University of Crete

Department of Philology - Division of Linguistics

Maria Pontiki

Fine-grained Sentiment Analysis

PhD Thesis

Rethymnon, July 2019

[Pick the date]

University of Crete

Department of Philology - Division of Linguistics

Maria Pontiki

Fine-grained Sentiment Analysis

PhD Thesis

Assessment Committee:

Prof. Elena Anagnostopoulou, *Professor of Linguistics in the Department of Philology- Sector of Linguistics at the University of Crete, Greece*.

Dr. Haris Papageorgiou, *Research Director at the Institute for Language and Speech Processing of the ATHENA Research and Innovation Center, Greece.*

Prof. Sophia Ananiadou, *Professor in the School of Computer Science at the University of Manchester, UK.*

Prof. Vasiliki Georgiadou, *Associate Professor of Political Science at the Department of Political Science and History, Panteion University, Greece*.

Dr. Stella Markantonatou, *Research Director at the Institute for Language and Speech Processing of the ATHENA Research and Innovation Center, Greece.*

Prof. Maria Liakata, *Assistant Professor in the Department of Computer Science at the University of Warwick, UK*.

Prof. Alexis Kalokerinos, *Professor of General Linguistics in the Department of Philology-Faculty of Letters at the University of Crete, Greece*.

Rethymnon, July 2019

«*Καί οἱ λέξεις φλέβες εἶναι,*
μέσα τους αἷμα τρέχει»

(Γιάννης Ρίτσος)

ACKNOWLEDGEMENTS

I am deeply grateful to the members of the advisory committee, Professor Elena Anagnostopoulou, Dr. Haris Papageorgiou, and Professor Sophia Ananiadou, and especially to Dr. Haris Papageorgiou for his continual supervision, guidance and encouragement throughout the research activity and the compilation of the work presented in this thesis.

I am equally grateful to the Institute of Language and Speech Processing (ILSP)/ Athena Research and Innovation Center, where the research work presented in this thesis was carried out, and especially to my colleagues and friends Dr. Dimitris Galanis and Konstantina Papanikolaou for their valuable help throughout the various stages of the research work of this thesis.

I would also like to sincerely thank my co-organizers at the ABSA shared task of the International Workshop on Semantic Evaluation for our exceptional cooperation, the research teams that applied part of the work proposed by this thesis to other languages, as well as the task participants for their valuable feedback.

Similarly, I am deeply grateful to the political and social scientists of the XENO@GR project, Professor Vasiliki Georgiadou and her team, for our fruitful collaboration and their valuable contribution in the interpretation of the verbal aggressiveness analysis results presented in this thesis.

Last but not least, I would like to thank my parents and my brother for being continuous supporters of my academic and professional route, and my life in general, my unconventional “Fokidos family” for their valuable support and reinforcement, and all my tango friends for their precious hugs when most needed.

Maria Pontiki

ABSTRACT

Sentiment Analysis constitutes a key data analytics tool in many contexts and domains, since it helps to automatically detect and analyze public opinions, emotions, attitudes, and needs in massive amounts of unstructured data using Natural Language Processing and Text Mining methods. The research activity of this PhD thesis focused on two types of opinionated user-generated content; evaluations expressed by customers about products and services and their aspects in particular domains of interest (restaurant and laptop reviews), and verbal attacks against predefined target groups of interest (e.g. refugees, immigrants) in the context of Computational Social Sciences, covering an industrial and a humanitarian use case of Sentiment Analysis, respectively. In this setting, this thesis presents: a) a principled unified knowledge representation framework and English benchmark datasets for Aspect Based Sentiment Analysis, and b) a linguistically inspired and data-driven framework for examining Verbal Aggression as an indicator of xenophobic attitudes in Greek Social Media.

ΠΕΡΙΛΗΨΗ

Η επιστημονική περιοχή της Ανάλυσης Συναισθήματος/Άποψης αποτελεί ένα βασικό εργαλείο ανάλυσης δεδομένων σε πολλούς κλάδους, καθώς βοηθάει στον αυτόματο εντοπισμό και την ανάλυση απόψεων, συναισθημάτων, συμπεριφορών και αναγκών που εκφράζονται από χρήστες του διαδικτύου σε τεράστιες ποσότητες μη δομημένων δεδομένων χρησιμοποιώντας μεθόδους Επεξεργασίας Φυσικής Γλώσσας και Εξόρυξης Κειμένου. Η ερευνητική δραστηριότητα αυτής της διδακτορικής διατριβής επικεντρώθηκε σε δύο τύπους περιεχομένου που παράγεται από χρήστες του διαδικτύου: αξιολογήσεις πελατών σχετικά με προϊόντα, υπηρεσίες και τα επιμέρους χαρακτηριστικά τους σε συγκεκριμένους τομείς επιχειρηματικής δραστηριότητας (κριτικές εστιατορίων και φορητών υπολογιστών), και λεκτικές επιθέσεις εναντίον προκαθορισμένων ομάδων στόχων (π.χ. πρόσφυγες, μετανάστες) στον χώρο των ανθρωπιστικών, και συγκεκριμένα, των Υπολογιστικών Κοινωνικών Επιστημών. Στο πλαίσιο αυτό, η παρούσα διατριβή παρουσιάζει: α) ένα ενιαίο πλαίσιο αναπαράστασης γνώσης και τρία σύνολα επισημειωμένων δεδομένων στην Αγγλική γλώσσα για την ανάλυση συναισθήματος/άποψης βασισμένης σε χαρακτηριστικά οντοτήτων, και β) ένα γλωσσικά εμπνευσμένο και καθοδηγούμενο από δεδομένα πλαίσιο για την εξέταση της λεκτικής επιθετικότητας ως δείκτη ξενοφοβικών συμπεριφορών στα ελληνικά μέσα κοινωνικής δικτύωσης.

LIST OF ABBREVIATIONS

ABSA	Aspect Based Sentiment Analysis
AC	Aspect Category
ACE	Aspect Category Extraction
ACP	Aspect Category Polarity Detection
AT	Aspect Term
ATE	Aspect Term Extraction
ATP	Aspect Term Polarity Detection
CSS	Computational Social Sciences
KD	Knowledge Database
LPT	Laptops (domain)
NLP	Natural Language Processing
OM	Opinion Mining
OTE	Opinion Target Expression
RQ	Research Question
RST	Restaurants (domain)
SA	Sentiment Analysis
SLA	Sentence-level Annotations
TG	Target Group
TLA	Text-level Annotations
VA	Verbal Aggression
VAM	Verbal Aggressive Message

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

1. INTRODUCTION

Grecian agora in ancient times was the heart of a city and of public life; citizens gathered to socialize, to buy and sell goods, politics were discussed, rhetoric was exercised as a way to persuade audiences to follow a proposal for action, and ideas were passed among great minds like Socrates, Aristotle and Plato. Nowadays, with the development of computer mediated communication, digital communities (e.g. online fora, blogs, social networks like Twitter and Facebook) have turn to become "contemporary agoras" (Rodriguez and Rojas-Galeano, 2018); users act and interact online, express and share their opinions, beliefs, and emotions on a variety of topics, entities, events, etc., and debate them openly and freely on a daily basis.

Opinions are key influencers of human behavior; our beliefs and perceptions of reality are conditioned on how others see the world, and whenever we need to make a decision we often seek out other's opinions (Liu, 2010). For instance, opinions are of great interest for companies who want to monitor their reputation and get timely feedback about their products and actions, while public opinion polls are widely used as a prediction instrument in various domains (e.g. electoral forecasting). In other words, opinions constitute valuable information for prediction and decision making. The availability of online user-generated opinionated data is beneficial not only for business, political or other purposes, but also for individual users. For example, nearly 95%[1] of shoppers read online reviews before making a purchase. According to a BrightLocal[2] consumer survey, 84% of people trust online reviews as much as they trust a recommendation from someone they know, and 74% of people say positive reviews dramatically improve trust in a business.

But there is also the other side of the coin. Online interaction differs from face-to-face communication, especially because anonymity and pseudonymity enable a more disinhibited self (Bandura, 2004); the fact that individuals can mask their identity or operate anonymously seems to influence online disinhibition, namely the tendency to say things in cyberspace that would not be said in person (Vandebosch and Cleemput, 2009). This permissiveness has led to online aggression incidents, where aggressive/abusive language is used in order to voice public criticism, personal indignation, or to simply let off steam; the forms of aggression are manifold and vary from expressions of disgust and contempt, to threats, slander, insults, and hatred (Rösner and Krämer, 2016). Hence, on the one hand there is a need to detect and

[1] https://learn.g2crowd.com/customer-reviews-statistics

[2] https://www.brightlocal.com/learn/local-consumer-review-survey/

moderate such content. On the other hand, the analysis of such content can help to examine individual negative and antisocial on-line verbal behaviors as well as complex social phenomena. For example, in April 2016, the Guardian[3] newspaper published an analysis of 1.4 million comments that it has blocked on its site since 2006, revealing regular incidents of xenophobia, racism, sexism and homophobia in users comment behavior calling it "the dark side of the Guardian comments".

Hence, user-generated content constitutes a valuable source of information for various stakeholders. The development of Natural Language Processing (NLP) and Text Mining methods enables effectively harnessing this wealth of data, which otherwise is impossible to manage; for example, Tripadvisor[4] hosts over 570 million user reviews covering a huge selection of travel listings worldwide (7.3 million accommodations, airlines, attractions and restaurants), while every second, on average, around 6,000 tweets are tweeted on Twitter[5], which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.

Sentiment Analysis (SA), defined as the computational study of subjective states (i.e. attitudes, stances, opinions, emotions, etc.) expressed in text (Liu, 2012), is considered a key data analytics tool in many contexts and domains ranging from business analytics (e.g. marketing, customer service), to social sciences (e.g. examining complex phenomena like xenophobia and racism), to urban sensing and citizen's behavior analysis (e.g. smart cities), and to clinical medicine, since it helps to automatically detect and analyze public opinions, emotions, needs and concerns in massive amounts of unstructured data using NLP and Text Mining methods.

The research activity of this PhD thesis is directed towards two aspects and sub-areas of SA with respect to the "two sides of the coin" presented above and proposes two linguistically inspired and data-driven frameworks for fine-grained SA focusing on: a) online customer reviews about specific target entities of interest and their aspects, and b) online verbal aggression in the context of Computational Social Sciences (CSS), respectively. In particular, this thesis presents:

- A principled unified knowledge representation framework and English benchmark datasets for Aspect Based Sentiment Analysis.
- A conceptual and computational framework for examining Verbal Aggression as an indicator of xenophobic attitudes in Greek Social Media.

[3] https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments

[4] https://expandedramblings.com/index.php/tripadvisor-statistics/

[5] http://www.internetlivestats.com/twitter-statistics/

The motivation, the research goals as well as the contribution in each case are described below in dedicated sections.

1.1 BACKGROUND AND MOTIVATION

1.1.1 ASPECT-BASED SENTIMENT ANALYSIS

Early work in SA focused mainly on the overall positive or negative classification of a given text or text span (Pang, Lee, and Vaithyanathan, 2002; Turney, 2002). While detecting the overall sentiment of a given text or snippet has a wide range of real-world applications, analyzing unstructured text only in terms of positivity and negativity is not sufficient enough to provide meaningful insights and is therefore of limited use; it is important to associate positive or negative polarity expressed in a text with the entities mentioned in context and their aspects. For example, customer reviews about laptops not only express the overall sentiment about a specific model (e.g. "*This is a great laptop*"), but also sentiments relating to its specific aspects, such as the hardware or the price. Subsequently, a review may convey contradictory sentiments (e.g. "*Its performance is ideal, I wish I could say the same about the price*") or objective information (e.g. "*This one still has the CD slot*") for different aspects of an entity. In this context, research has moved towards fine-grained approaches like aspect-based (or feature-based) sentiment analysis (ABSA) that involves identifying sentiment on different aspects of entities and entities themselves (Zhang and Liu, 2014).

ABSA extends the typical SA setting with a more realistic assumption that negative or positive polarity is associated with specific aspects (or product features) rather than the whole text unit, and allows a model to produce a fine-grained understanding of people's opinion towards a particular product (Ma, Peng and Cambria, 2018). Traditionally, such insights are obtained through closed-form customer satisfaction questionnaires the development and execution of which are expensive or may not be available. Hence, the availability of online customer comments offers a valuable source of information for business intelligence. Some review sites (e.g. TripAdvisor, Amazon) contain such information in the form of multi-aspect user ratings (e.g. stars). However, user ratings are not always available (e.g. Twitter), and in addition this type of information is not always sufficient; taking into account the textual component of user reviews provides also evidence to understand the reason behind the rating (Titov and McDonald, 2008a; McAuley, Leskovec and Jurafsky, 2012), and results in better

general or personalized review score predictions than those derived from the numerical star ratings given by the users (Ganu, Elhadad and Marian, 2009). An ABSA method can analyze large amounts of unstructured texts and extract information not included in the user ratings that are available in some review sites. Therefore, it is critical in mining and summarizing opinions from on-line reviews (Titov and McDonald, 2008a) and is considered a key data analytics tool in the business analytics domain, since it can provide valuable advantages like correlations with Key Performance Indicators (KPIs), reputation management, improving products and customer service, indicating potential customers, personalizing contents or displaying targeted commercials.

Although many computational methods and systems have been proposed - ranging from frequency (Zhuang et al., 2006) and syntactic relation (Qiu et al., 2011) based approaches to deep learning models (Lakkaraju et al., 2014) and neural networks (Alghunaim et al., 2015)- there is no established task decomposition in terms of a conceptual framework. Depending on the approach, "aspect" can be a synonym for both fine- and coarse grained types of information; coarse predefined categories similar to rateable aspects that do not necessarily occur as terms in a text, explicit (e.g. price) or implicit (e.g. expensive) terms denoting an aspect, opinion targets, etc. As a result, publicly available benchmark datasets adopt different annotation schemes within different tasks.

This diversity in the decomposition of ABSA as an information extraction task is translated to different computational approaches –generating different types of output even for the same domains- which are not directly comparable. In addition, the available datasets have been constructed to feed specific (types of) algorithms in each case; the annotations are typically presented as numbers of training and testing instances. No qualitative information is provided (e.g. main annotation problems, if and how they have been resolved), since no annotation guidelines of how to build a benchmark ABSA dataset are available. In other words, the computational framework conquers and somehow determines the conceptual framework. As in many NLP tasks, human annotated datasets are of critical importance not only for development and training purposes but also for evaluating the proposed methods and techniques in each case. However, the creation of high quality benchmark data is a labor intense task. In the case of ABSA, annotation is a very difficult task, since it involves labeling the targets (entities, aspects) of the expressed sentiments requiring definitions of what constitutes a target, whether targets are linked to opinion expressions, and how the boundaries of target spans should be defined (Farra, McKeown and Habash, 2015; Kim and Hovy, 2004; Somasundaran, Wiebe and Ruppenhofer, 2008).

In this setting, the research activity of this PhD thesis focused on the review of the scientific literature and the existing datasets in the field of ABSA as well as on collecting user-generated data about particular target entities of interest. The aim of the research was to decompose ABSA as an information extraction task focusing on the intended meaning of the text and how it can be formalized into a conceptual knowledge representation framework, as well as to perform a systematic annotation study examining the different ways in which aspects are linguistically instantiated. The ultimate goal was to compile a set of detailed annotation guidelines and to construct gold-standard annotations fostering ABSA research towards a more structured and meaningful output as well as to provide a common test bed for computational methods and techniques.

1.1.2 VERBAL AGGRESSION AND XENOPHOBIA

While Verbal Aggression (VA) predated the Internet, the extent and the nature of online communication tools amplifies incidents of aggression affecting billions of people; VA can manifold in a multitude of ways (e.g. flaming, cyberbullying, hate speech) in different contexts with somewhat different intentions and various effects on individuals, communities and social cohesion. For example, flames posted on online discussion groups can be defamatory with serious consequences to an organization's products, services, and good-will (Alonzo and Aiken, 2004), cyberbullying can have devastating consequences for the victims ranging from withdrawal from school activities, school absence, and school failure, to eating disorders, substance abuse, depression, and even suicide (Chibbaro, 2007; Klomek, Brunstein and Gould, 2011), while hate speech poses threat to the dignity of individuals, to personal liberties, to the social fabric of democracies (Waltman and Haas, 2010) as well as to the security of societies. Online VA may also escalate to physical violence; for instance, online hateful language resulted in massive violence in Kenya before and after the elections in 2007 and 2008 (Benesch, 2018). Similarly, a relationship between the online hateful debate on refugees and attacks on home for asylum seekers in Germany has been reported[6] (Köffer et al., 2018).

The recent refugee/immigrant crisis in Europe gave burst to xenophobic sentiments, attitudes and practices ranging from individual (re)actions to official state policies. VA constitutes an important component in the study of xenophobia, since verbal

[6] Wie aus Netzhass Gewalt wird und was dagegen hilft, http://www.spiegel.de/netzwelt/netzpolitik/netzhass-und-gewalt-was-man-dagegen-tun-kann-lobo-kolumne-a-1048799.html

attacks targeting foreigners can be indicative of xenophobic sentiments, attitudes and perceptions. For example, in an attempt to map xenophobia on the Estonian Internet by describing the use of VA directed against some more common groups in Estonia, Laineste (2012) describes the main objects of online flaming and the social and contextual background of the target choice. The close relation of online VA with xenophobia is also demonstrated by the hate speech literature and especially by approaches that focus on xenophobia-related types of hate speech like racist (Kwok and Wang, 2013; Waseem and Hovy, 2016) and hate speech directed to immigrants (Sanguinetti et al., 2018) or specific ethnic groups (Warner and Hirschberg, 2012), even though they do not make an explicit reference to xenophobia. Traditionally, xenophobia is examined using empirical and statistical methods; xenophobic attitudes are being measured using data coming from focus groups, interviews, and public sentiment polls using standard questions in order to capture opinions, emotions, perceptions and beliefs (e.g. Eurobarometer). The user-generated content available online constitutes a valuable source of information not only in terms of quantity (massive amounts of data), but also in terms the content itself, since the online disinhibition also allows aggressive forms of expression that cannot be captured by traditional methods that use face to face communications. Despite the numerous research efforts in automatically detecting and analyzing online VA, the user-generated content has been scarcely explored from the xenophobia standpoint at a large scale.

The various classification methods and algorithms that have been proposed for the detection of aggressive content employ different definitions and address somewhat different aspects of online VA; flames (Razavi et al., 2010), profanity-related offensive content (Sood, Antin and Churchill, 2012), cyberbullying (Dinakar et al., 2012), hate speech (Warner and Hirschberg, 2012), or abusive language in general (Chen et al., 2012) in different social media and online communities. Focusing on hate speech that is more close to the work presented in this thesis, some studies adopt a binary classification schema aiming to distinguish hateful from non-hateful content (Djuric et al., 2015), other studies attempt to differentiate hate speech and offensive language (Nobata et al., 2016), while another line of research focuses on specific types/categories of hate speech e.g. anti-Semitic hate speech (Warner and Hirschberg, 2012), racist and sexist hate speech (Waseem and Hovy, 2016). However, as Davidson et al. (2017) point out, some approaches conflate hate speech with offensive language making it difficult to ascertain the extent to which they are really identifying hate speech. In addition, identifying hate speech consistently is difficult and often yields the paradox that each person seems to have their own intuition about what hate speech is, but rarely are two people's understandings the same (Saleem et al., 2017).

There is a general diversity and lack of consensus in terminology of online VA that often results in overlap between several subtasks with the need for clear and operational definitions being stressed by several researchers (e.g. Waseem et al., 2017).

Furthermore, the detection of online aggressive content is not a trivial task; verbal attacks are shaped differently depending on individuals' intentions and strategic choices in language use ranging from expressions of negative emotions, name calling, swearing, threatening, and insulting to the use of humor, sarcasm and irony or even paralanguage (Tereszkiewicz, 2012). In addition, detecting and classifying an aggressive message is not enough; for example, an effect of hate speech depends on the originator, the content and the targeted one (Chetty and Alathur, 2018). However, only few studies incorporate these elements in their accounts.

In this setting, this thesis focuses on verbal attacks expressed in Twitter against specific predefined Target Groups (TGs) of interest and proposes a data-driven and linguistically-inspired SA framework for examining VA as an indicator of online xenophobic attitudes in the context of the XENO@GR[7] project, an interdisciplinary project aiming to examine the evolution of the phenomenon of xenophobia in the contemporary Greek society from the 1990s. This notion of VA is closely related to hate speech, however, given the lack of a universally agreed definition as well as the legal implications of the term hate speech, the general term VA is used instead for explicitly stated verbal attacks targeting specific groups of foreigners in Greece. The ultimate goal was to build a Knowledge Database (KD) that would help to formulate adequate responses to specific Research Questions (RQs) concerning the nature and the evolution of the phenomenon of xenophobia as a violent practice in the Greek society.

1.2 RESEARCH GOALS AND CONTRIBUTION

1.2.1 TOWARDS A PRINCIPLED UNIFIED ABSA FRAMEWORK

The aim of the research activity of this PhD thesis in the field of ABSA was to decompose it as an information extraction task focusing on the intended meaning of the texts and fostering research towards more end-user and application oriented ABSA outputs. In particular, the thesis focuses on the following the research goals:

[7] http://xenophobia.ilsp.gr/?lang=el

- Data-driven annotation frameworks and respective codebooks focusing on specific target entities of interest (e.g. restaurants, laptops) from the customer reviews domain; the ultimate goal was to formalize the ABSA problem into a more meaningful and structured knowledge representation framework that directly reflects the intended meaning of the texts as opposed to isolated pieces of information that are usually extracted from surface features (e.g. aspect terms lists).
- Gold-standard human-authored annotations (benchmark datasets) following the respective codebooks for the specific target entities of interest that could be used as a common (training and evaluation) framework for ABSA methods.

To achieve these goals, the research activity of this thesis was broken down into the following three phases:

- *Building an annotation framework for ABSA*. The first step was to perform a systematic annotation study and examine how existing definitions are applied to datasets for both fine- and coarse-grained ABSA.
- *Redefining ABSA*. Based on the key lessons learned during the first phase of this study, a new principled unified ABSA framework was designed in order to address representation issues having to do with completeness and meaningfulness.
- *Extending ABSA*. The third and last part of this thesis work extends the new ABSA framework towards text-level annotations, as well as to new domains and to languages other than English.

Detailed annotation guidelines and respective benchmark datasets were generated in all the three phases. The contribution of this thesis in the field of ABSA can be summarized as follows:

- A new definition of "aspect" that makes more explicit the difference between the entities and the attributes that are being evaluated. The new definition yields a new representation framework in which all the identified constituents of the expressed sentiments meet a set of guidelines/specifications and are linked to each other within sentence-level tuples that can be also be aggregated at the text level enabling the generation of meaningful opinion summaries.
- Three sets of detailed ABSA annotation codebooks (in correspondence to the three phases of the research activity). To the best of our knowledge, no annotation guidelines have been made available in the ABSA literature so far.
- Three sets of respective human-authored benchmark datasets for two domains (restaurants and laptops customer reviews) for the English language.

The proposed annotation frameworks and datasets were used to support the ABSA shared task that was introduced for the first time in the research community in the context of the International Workshop on Semantic Evaluation (SemEval). The task was organized and ran in parallel with this thesis research activity (the author was one of the task organizers) for three years providing training datasets and a common evaluation framework for ABSA methods. It started in 2014 (SE-ABSA14[8]) with four datasets following the state-of-the art frameworks in ABSA; in 2015 (SE-ABSA15[9]) the new ABSA framework and datasets were introduced, which in 2016 (SE-ABSA16[10]) were extended to new domains and seven languages (i.e. Arabic, Chinese, Dutch, French, Russian, Spanish, and Turkish). The task attracted a significant number of participants who contributed a large number of submissions and system description papers (consult Table 1).

	Participating teams	System submissions	System description papers
SE-ABSA14	32	163	28
SE-ABSA15	16	93	12
SE-ABSA16	29	245	20

Table 1: SemEval ABSA shared task participation statistics

The proposed annotation guidelines have been adopted for the creation of benchmark datasets in the same or new domains in languages other than English also outside the SemEval challenge; restaurant reviews in Czech (Steinberger, Brychcín and Konkol, 2014) and in Bangla (Rahman and Kumar Dey, 2018), restaurant and hotel reviews in Vietnamese (Thin et al., 2018), book (Al Smadi et al., 2015) and laptop (Al-Ayyoub et al., 2018) reviews in Arabic, product reviews in Hindi (Akhtar, Ekbal, and Bhattacharyya, 2016).

After the SemEval challenge, the generated datasets are still used for training and testing purposes by numerous researchers and constitute the standard benchmarks for ABSA (e.g. Gunes, 2016; Hasib and Rahin, 2017; Kushwaha and Chaundhary, 2017; Li and Lam., 2017; Akhtar et al., 2018; de Kok et al, 2018; Dilawar et al., 2018; Dong and de Melo, 2018; Li, Liu and Zhou, 2018; Moore and Rayson, 2018; Nguyen, 2018; Ouyang and Su, 2018; Piryani, Gupta, and Singh, 2018; Wang et al., 2018; Xiang, He and Zheng, 2018; Zhu and Qian, 2018; Zhu et al., 2018). In addition, the proposed datasets were recently enriched with a new annotation layer (sentiment expressions) by Kaljahi and Foster (2018). Overall, the SA research activity of this PhD thesis in

[8] http://alt.qcri.org/semeval2014/task4/
[9] http://alt.qcri.org/semeval2015/task12/
[10] http://alt.qcri.org/semeval2016/task5/

the field of ABSA produced three publications[11,12,13]. The outcome of the research activity of this thesis was presented as an invited talk[14] at the Xerox Research Center Europe.

1.2.2 VA AS AN INDICATOR OF ONLINE XENOPHOBIC ATTITUDES

The second part of the SA work presented in this thesis constitutes the fine-grained SA framework that was designed and implemented in the context of the XENO@GR[15] project, an interdisciplinary project aiming to examine the evolution of the phenomenon of xenophobia in the contemporary Greek society from the 1990s. The main research puzzle of the project was whether (or not) the phenomenon of xenophobia is an outcome of the recent financial crisis or it comprises a long-lasting social perception deeply rooted in the Greek society, and it was further decomposed into specific RQs. Looking beyond traditional empirical approaches of social science research, the project aimed at analyzing and providing an in-depth understanding of the evolution of the phenomenon of xenophobia as a violent practice in the Greek society based on social computational methods and big data analytics. More specifically, two principal data analytics workflows were employed: a) Event Analysis using news data aiming to capture physical attacks (e.g. violent attacks, sexual attacks, attacks against properties) against the predefined TGs of interest, and b) SA using Twitter data aiming to detect verbal attacks targeting the predefined TGs of interest.

In this context, the research activity of this thesis was directed towards a data-driven and linguistically-inspired conceptual and computational framework for the analysis of different types of online verbal attacks against the predefined TGs of interest (e.g. Immigrants, Pakistani, Albanians, etc.) aiming to address the following three RQs focusing on the amount, the type and the content of the verbal attacks, respectively:

[11] **Pontiki**, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014). Semeval-2014 task 4: Aspect Based Sentiment Analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 27–35.

[12] **Pontiki**, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 486–495.

[13] **Pontiki**, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. and Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp.19–30.

[14] Talk Title: "Redefining Aspect-based Sentiment Analysis: the ABSA challenge experience", Xerox Research Center Europe, Grenoble, France, 3/11/2016.

[15] http://xenophobia.ilsp.gr/?lang=el

- RQ1: Who are the main targets of Twitter verbal attacks?
- RQ2: Which are the main types of Twitter verbal attacks?
- RQ3: Are there stereotypes and prejudices against foreigners rooted deeply in the Greek society?

To address the above RQs a five-step methodology was followed; the first step was to gather data related to the predefined TGs of interest. In a second phase, samples of the collected data were explored in order to identify different aspects of VA related to the predefined TGs. Then, based on data observations and literature review findings, a linguistically-driven VA framework was designed according to which the VA messages (VAMs) are classified into distinct categories based on specific criteria. The next step was the design and the development of the resources and the models needed for the computational treatment of the VA framework (VA analyzer). Subsequently, the data collections were automatically processed. The output of the VA analyzer was then recorded in a KD and visualized in various ways in order to obtain a better understanding of the data and the results of the analysis. The KD that was built helped to formulate adequate responses the aforementioned RQs, which in turn contributed to the RQs of the XENO@GR project either standalone or in combination with other research findings (i.e. results of the Event Analysis workflow, findings from empirical studies).

The contribution of this thesis is three-fold:

- A comprehensive overview of the key concepts, types, causes and effects of (offline and online) VA, and the related state-of-the-art computational methods.
- A data-driven and linguistically-inspired conceptual and computational framework for fine-grained SA in terms of VA towards specific TGs of interest. The proposed VAM taxonomy illuminates different aspects of VA. The computational method enables applying the taxonomy at a large-scale in the Greek language providing valuable insights that facilitate the study of the formulation of VA in relation to specific TGs, and to measure and to monitor different aspects of VA as an important component of the manifestations of xenophobia as a violent practice in Greece.
- An interdisciplinary end-to-end fine-grained SA approach; typically, SA approaches focus only on the information extraction process, and conclude with the evaluation of the performance of the proposed methods. This thesis takes a step further by linking the analysis results to specific RQs and including the critical step of their interpretation; in collaboration with political and social scientists the VA analysis results are further analyzed both quantitatively and

qualitatively providing valuable insights for the research problem under investigation as well as a tangible example of how a carefully designed fine-grained SA approach can serve as a complementary research instrument in the context of CSS.

Overall, the SA research activity of this PhD thesis in the context of CSS produced two publications[16,17]. The VA framework and the analysis results were also presented as part of the XENO@GR project at national[18] and international[19,20] events.

1.3 THESIS OUTLINE

The remainder of this thesis is structured as follows. Chapter 2 presents the first part of this thesis work that is a fine-grained SA framework focusing on online customer reviews about specific target entities of interest and their aspects. Section 2.1 presents an overview of the landscape in the field of ABSA including the key definitions and representation models (2.1.1), available benchmark datasets (2.1.2), and the state-of-the-art methods and techniques (2.1.3). Then, sections 2.2 – 2.4 describe the knowledge representation framework and the benchmark datasets proposed by this thesis in each one of the three phases of the research activity; building an annotation framework for ABSA (2.2), redefining ABSA (2.3), and extending ABSA (2.4). In particular, during the first phase a data-driven ABSA annotation codebook was compiled (2.2.2) -following the state-of-the-art definitions of aspect and based on a systematic annotation study on existing datasets (2.2.1)-, and was then applied to the aforementioned datasets that were extended with new unseen sentences (2.2.3). The proposed benchmark datasets (2.2.4), along with the annotation framework and the guidelines were adopted from the SE-ABSA14 shared task (2.2.5). Then, based on the key lessons learned during the first part study (2.2.6), a new principled unified ABSA framework was designed (2.3.1) and was then applied to new datasets (2.3.3) following a new data-driven ABSA annotation codebook was compiled for three

[16] Galariotis, G., Papanikolaou, K., Georgiadou, V., Kafe, A., Lialiouti, Z., Papageorgiou, H., **Pontiki**, M. and Pappas, D. (2016). Xenophobia in Greece: A Computational Social Science Approach. *Poster presented at the 3rd Computational Social Science Winter Symposium 2016*, Cologne, Germany.

[17] **Pontiki**, M., Papanikolaou, K. and Papageorgiou, H. (2018). Exploring the Predominant Targets of Xenophobia-motivated behavior: A longitudinal study for Greece. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Natural Language Meets Journalism Workshop III*, Miyazaki, Japan, 7-12 May, 2018, pp.11 –15. [**best paper award**]

[18] Conference "Xenophobia in Greece: Evolution and Causes", 10 January 2017, Panteion University, Athens.

[19] Computational Social Science (CSS): New Frontiers of Collaboration? European University Institute (EUI) Workshop on CSS Approaches, 25 May 2016, Villa la Fonte, Florence, Italy.

[20] International workshop on Critical Reflections on Asylum, Migration and Xenophobia in Europe, 24 January 2017, EUI, Florence, Italy.

domains (2.3.2). The proposed datasets (2.3.4), along with the annotation framework and the guidelines were again adopted from the SE-ABSA15 shared task (2.3.5) as a follow up of the SE-ABSA14. Finally, section 2.4 presents the third part of this thesis work in the context of ABSA that extended the new ABSA framework towards two directions; text-level annotations that can be used for the generation of opinion summaries (2.4.1), and application of the proposed annotation framework to other languages and domains in the context of the SE-ABSA16 shared task as a follow up of the SE-ABSA15 (2.4.2).

Chapter 3 presents the second part of this thesis that is the fine-grained SA framework designed and implemented in the context of the XENO@GR project for examining VA as an indicator of xenophobic attitudes in Greek Social Media. Starting with the background and the research strand of this work (3.1), section 3.1.1 provides a comprehensive overview of the key concepts, types, causes and effects of offline and online VA (3.1.1.1), and the related state-of-the-art computational methods (3.1.1.2). The relation of xenophobia to VA along with an overview of the historical context of this phenomenon in Greece, and the research goals of this thesis are presented in sections 3.1.2.1 and 3.1.2.1, respectively. Section 3.2 presents the five-step methodology designed for building the VA analysis framework following the respective steps; data collection (3.2.1), explorative analysis (3.2.2), taxonomy of aggressive messages (3.2.3), computational analysis (3.2.4), and data visualization (3.2.5). Subsequently, section 3.3 discusses the analysis results with regard to the specific RQs that this thesis aims to address focusing on the amount (RQ1: main targets of attacks, 3.3.1), the type (RQ2, 3.3.2) and the content (RQ3: stereotypes and prejudices, 3.3.3) of the aggressive messages, respectively. Section 3.4 concludes with a discussion of the most interesting findings of the quantitative and the qualitative analysis and presents also some further insights about the nature of online xenophobic behavior in Greece (3.4.1). The limitations and the possible future research directions of the presented work are discussed in 3.4.2. Finally, chapter 4 provides a summary of the work presented in this thesis and its contribution in the context of ABSA (4.1) and online VA (4.2), respectively.

2. TOWARDS A PRINCIPLED UNIFIED ABSA FRAMEWORK

This chapter presents the first part of this thesis work; the fine-grained SA framework that focuses on online customer reviews about specific target entities of interest and their aspects. Section 2.1 presents an overview of the landscape in the field of ABSA including the key definitions and representations (2.1.1), the available benchmark datasets (2.1.2), and the state-of-the-art methods and techniques (2.1.3). Then, sections 2.2 – 2.4 describe the knowledge representation framework and the benchmark datasets proposed by this thesis in each one of the three phases of the research activity; building an annotation framework for ABSA (2.2), redefining ABSA (2.3), and extending ABSA (2.4), respectively.

2.1 BACKGROUND AND RESEARCH STRAND

2.1.1 KEY CONCEPTS AND DEFINITIONS

SA or Opinion Mining (OM) is defined as the computational study of opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, etc., expressed in text (Liu, 2012). The two terms are often used interchangeably to denote the same field of study, however SA actually focuses on emotions and different types of feelings and attitudes, while OM on evaluations and polarity detection, respectively; given that sentiment identification usually involves polarity detection, the two tasks are often combined or used as synonyms.

For example, the following review text about an iPhone (Liu, 2010) expresses several positive (sentences 2, 3, 4) and negative (sentences 5, 6) sentiments about different targets (entities or aspects):

"(1)*I bought an iPhone a few days ago.* (2)*It was such a nice phone.* (3)*The touch screen was really cool.*(4)*The voice quality was clear too.* (5) *However, my mother was mad with me as I did not tell her before I bought it.* (6) *She also thought the phone was* ***too*** *expensive, and wanted me to return it to the shop.*"

Sentences (2-4) express opinions/evaluations about a product (iPhone) and specific aspects/features of it ("*touch screen", "voice quality*"), while sentence (5) an emotion about the reviewer/iPhone owner (*"mad with me"*). Sentence (6) conveys an opinion about the price of the iPhone ("*expensive*"), but also a desire/suggestion for the

iPhone owner ("*wanted me to return it to the shop*"), which implies a negative sentiment towards the iPhone. Focusing on the sentiment/opinion targets, some of them are expressed explicitly using specific terms naming them ("*touch screen", "voice quality*"), other through pronouns (*it* for iPhone, *me* for the reviewer /iPhone owner), while other can be implicitly inferred (e.g. the aspect "price" can be implicitly inferred through "expensive"). Finally, the opinion holder in sentences (2-4) is the reviewer/iPhone owner, while in (5-6) the reviewer's mother ("*my mother", "she*").

As it is indicated by the above example, SA is a multifaceted problem that touches every aspect of NLP (e.g. named-entity recognition, co-reference resolution, negation handling, anaphora resolution, word-sense disambiguation) and as Cambria et al. (2013) mention "*it requires a deep understanding of the explicit and implicit, regular and irregular, and syntactic and semantic language rules*". Focusing on opinions/evaluations, a comprehensive definition of what constitutes an opinion is provided by Liu (2012); an opinion is a quintuple (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on the aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative or neutral, and can be expressed with different strength/intensity levels. Following this definition, SA consists of five respective subtasks; given a text or a text snippet:

- extract all the entities mentioned
- extract all the (explicit or implicit) aspects of these entities
- assign a respective sentiment polarity label on each aspect
- extract the opinion holder (i.e. a person or organization that holds the opinion)
- extract the time

The opinion holder and the time may or not occur as information in a given text or text snippet; sometimes these types of information are available at the metadata accompanying a text (e.g. a customer review usually contains information about who and when wrote it). Focusing on the entities, Liu (2012) uses the general term object (O) which can be a product, a topic, a person, an event, or an organization, and it is associated with a pair, O: (T, A), where T is a hierarchy or taxonomy of components (or parts) and sub-components of O (e.g. "*touch screen*", "*voice*"), and A is a set of attributes of O (e.g. "*price*"). Each component has its own set of sub-components and attributes (e.g. "*voice quality*").

A typical SA method would assign polarity labels or scores at the sentence or at the text level. An ABSA or Feature-based SA method involves identifying sentiment on

different aspects of entities and entities themselves (Zhang and Liu, 2014). Although many ABSA computational methods and systems have been proposed (see section 2.1.3), there is no established task decomposition in terms of a conceptual framework. Depending on the approach, aspect can be a synonym for both fine- and coarse grained types of information. The basic aspect definitions are summarized below:

- Coarse predefined categories (i.e. concept names) similar to rateable aspects (e.g. Ganu, Elhadad and Marian, 2009; McAuley, Leskovec and Jurafsky, 2012) that do not necessarily occur as terms in a text or text snippet.
- Aspects are opinion targets i.e. all the targets towards which opinion can be expressed (e.g. Qiu et al., 2011).
- Aspects or features (Hu and Liu, {2004a, 2004b}) or facets (Mei et al., 2007) denote components/ parts, subcomponents of the target entity, and attributes of the target entity or its components (Liu, 2006; Zhang and Liu, 2014).

For example, given the sentence "*The pizza was delicious but do not come here on an empty stomach.*" from a customer review about a particular restaurant, the output of an ABSA method would be as follows for each of the above representations respectively:

- [FOOD: positive & FOOD: negative] or [FOOD: conflict]
- "pizza": positive
- pizza [+5], size [-3] [u][21]

Hence, depending on the representation adopted, we may have coarse categories that do not occur as terms in the sentence, or we may have explicit mentions of the aspects like "*pizza*" as well as implicit aspects like the "*pizza size*". As for the sentiment classification schema, it can vary from labels like "*positive*" and "*negative*" or "*conflict*" to numerical scores. So, given the different representations, in many cases the ABSA methods are not directly comparable, since they adopt different classification schemes and focus on different domains using different datasets.

2.1.2 ABSA DOMAINS AND DATASETS

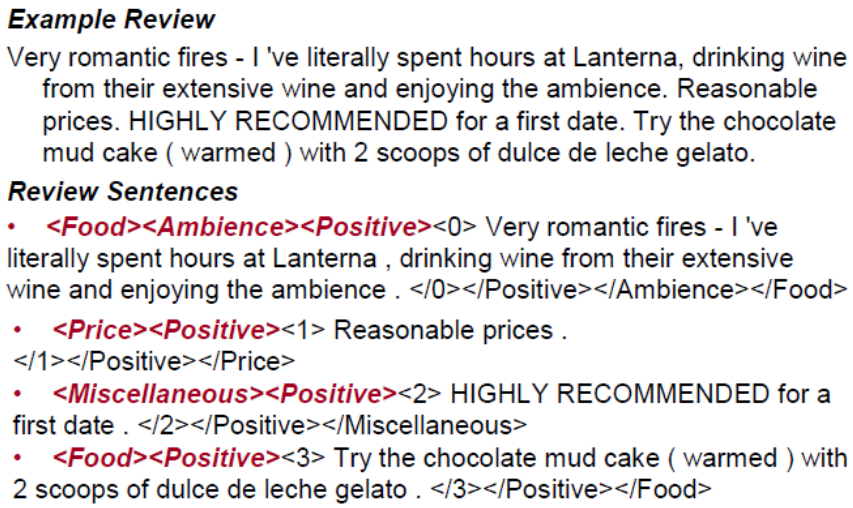
ABSA is mainly applied on product or service customer reviews from websites and e-commerce platforms (e.g. Amazon, Tripadvisor, Yelp, etc.). Publicly available ABSA datasets adopt different annotation schemes for different subtasks; coarse aspect categories and respective ratings (Ganu, Elhadad and Marian; McAuley, Leskovec

[21] [u] denotes feature/aspect not appeared in the sentence (Hu and Liu, {2004a, 2004b}).

and Jurafsky, 2012) or fine-grained aspect (terms) annotations (Hu and Liu, {2004a, 2004b}) at the sentence (Ganu, Elhadad and Marian; Hu and Liu, {2004a, 2004b}) or at the text (review) level (McAuley, Leskovec and Jurafsky, 2012).

The coarse aspects are predefined labels for each domain (and dataset) and not necessarily terms occurring in a sentence or text. For example, the restaurant reviews dataset of Ganu, Elhadad and Marian (2009) includes annotations for coarse aspect categories and overall sentence polarities (Fig. 1); each sentence is tagged following a six-way classification schema for aspects (FOOD, SERVICE, PRICE, AMBIANCE, ANECDOTES (i.e. sentences describing the reviewer's personal experience or context, but that do not usually provide information on the restaurant quality), and MISCELLANEOUS (i.e. sentences that do not belong to the other five categories including sentences that are general recommendations), and a four-way classification schema for sentiment polarity (positive, negative, neutral, conflict).

The datasets of McAuley, Leskovec and Jurafsky (2012) provide aspects and respective ratings at the review level (i.e., aspects and numeric ratings associated with entire reviews, not particular sentences)[22] about Beers, Pubs, Toys and Games, and Audiobooks. The reviews are obtained from sites that allow users to evaluate a product not only in terms of its overall quality, but also focusing on specific predefined aspects (e.g. SMELL and TASTE for Beers, FUN and EDUCATIONAL VALUE for Toys and Games). An example of an annotated review is provided in Fig. 2.


Example Review

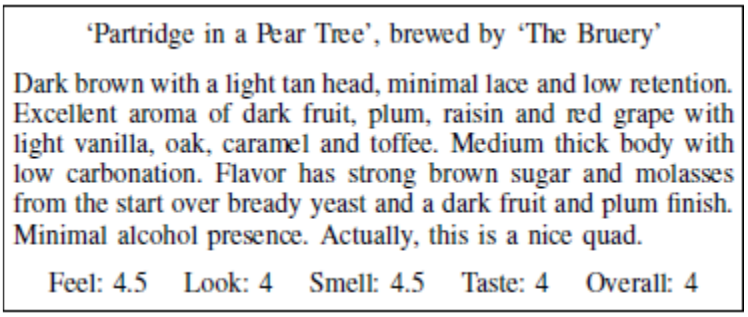
Very romantic fires - I 've literally spent hours at Lanterna, drinking wine from their extensive wine and enjoying the ambience. Reasonable prices. HIGHLY RECOMMENDED for a first date. Try the chocolate mud cake (warmed) with 2 scoops of dulce de leche gelato.

Review Sentences

- ***<Food><Ambience><Positive>***<0> Very romantic fires - I 've literally spent hours at Lanterna , drinking wine from their extensive wine and enjoying the ambience . </0></Positive></Ambience></Food>
- ***<Price><Positive>***<1> Reasonable prices . </1></Positive></Price>
- ***<Miscellaneous><Positive>***<2> HIGHLY RECOMMENDED for a first date . </2></Positive></Miscellaneous>
- ***<Food><Positive>***<3> Try the chocolate mud cake (warmed) with 2 scoops of dulce de leche gelato . </3></Positive></Food>

Figure 1: Example of an annotated review from citysearch.com (Ganu et al., 2009)

[22] A subset of the datasets has been annotated with aspects at the sentence level.


'Partridge in a Pear Tree', brewed by 'The Bruery'

Dark brown with a light tan head, minimal lace and low retention. Excellent aroma of dark fruit, plum, raisin and red grape with light vanilla, oak, caramel and toffee. Medium thick body with low carbonation. Flavor has strong brown sugar and molasses from the start over bready yeast and a dark fruit and plum finish. Minimal alcohol presence. Actually, this is a nice quad.

Feel: 4.5 Look: 4 Smell: 4.5 Taste: 4 Overall: 4

Figure 2: Example of an annotated review from BeerAdvocate (McAuley et al., 2012)

The SentiHood dataset (Saeidi et al., 2016) includes aspect annotations for more than one entity that are locations or neighborhoods of the city of London. The data was taken from question answering platform of Yahoo! Answers and annotated using a set of predefined aspects (e.g. SAFETY, PRICE, QUIET, DINING, NIGHTLIFE, TRANSIT-LOCATION, TOURISTY, SHOPPING, GREEN-CULTURE AND MULTICULTURAL) and a binary (positive/negative) classification schema (Fig. 3).

Sentence	Labels
The cheap parts of London are **Edmonton** and **Tottenham** and they are all poor, crime ridden and crowded with immigrants	(Edmonton,price,Positive) (Tottenham,price,Positive) (Edmonton,safety,Negative) (Tottenham,safety,Negative)
Hampstead area, more expensive but a better quality of living than in **Tufnell Park**	(Hampstead,price,Negative) (Hampstead,live,Positive)

Figure 3: Examples of annotated sentences in Sentihood (Saeidi et al., 2016)

Other datasets contain more fine-grained annotations. For example, the very popular dataset of Hu and Liu ({2004a, 2004b}), contains reviews of five particular electronic products (e.g., Nikon Coolpix 4300) from amazon.com. Each sentence is annotated with aspect terms (i.e. terms naming particular aspects of the reviewed products) and respective numeric sentiment polarity scores. The annotation schema allows also tagging implicit features (indicated with the attribute [u]), which however are not resolved/linked to any actual product aspect categories (Fig. 4).

camera[+2]##This is my first digital camera and what a toy it is...
size[+2][u]##it is small enough to fit easily in a coat pocket or purse.

Figure 4: Example of annotated sentences from amazon.com consumer electronic reviews (Hu and Liu, {2004a, 2004b})

The datasets of Pavlopoulos (2014) contain also aspect terms annotations in three domains (customer reviews for restaurants, laptops and hotels). Contrary to Hu and Liu ({2004a, 2004b}), the datasets include also sentences expressing conflicting opinions (e.g. "*The screen is clear but small*") and neutral sentences (e.g. "*It has a 4.8-inch screen*"). The restaurants dataset is a subset of the dataset of Ganu, Elhadad and Marian (2009) that is further annotated with aspect terms and their polarity without however providing any information linking the annotated terms to the existing annotated coarse aspect categories. On the other hand, as it is illustrated below in Fig.5, in the movie reviews dataset of Zhuang et al. (2006) the movie features (Fword) are attributed to one of 20 predefined categories (Ftype). The opinion words (Oword) and their semantic orientations (Otype) are also tagged.

⟨Sentence⟩ I have never encountered a movie whose supporting cast was so perfectly realized.⟨FO Fword="supporting cast" Ftype="PAC" Oword="perfect" Otype="PRO"/⟩ ⟨/Sentence⟩

Figure 5: Example of an annotated sentence from Internet Movie Database (Zhuang et al., 2006)

Other approaches go further to more annotation layers and more fine-grained annotated datasets. For example, the JDPA corpus (Kessler et al., 2010) that consists of blog posts about cars and cameras is annotated with a complex set of entities and relations, including aspects, subjective phrases, polarities, part-of relations, feature-of relations, opinion holders and others, while Toprak, Jakob and Gurevych (2010) introduce a corpus of consumer reviews from the rateitall and the eopinions websites annotated according to a two-level annotation schema; sentence-level annotations for given topics (e.g. the name of the university or the service being reviewed) and topic-specific evaluations, and expression-level annotations that provide further information about the properties (semantic orientation, intensity), and the functional components of the evaluations (opinion terms, targets and holders).

(17) I have **no**$_{[modifier1]}$
complaints$_{[opinionexpression1]}$ *about the entire* **PhD**
journey$_{[target1]}$ *and* **highly**$_{[modifier2]}$
recommend$_{[opinionexpression2]}$ **this school**$_{[target2]}$.

Figure 6: Example of expression-level annotations in the dataset of Toprak et al. (2010)

Except from individual research efforts, benchmark datasets have been also released in the context of shared tasks that provide training datasets and the opportunity for direct comparison of different approaches on common test sets. The IGGSA Shared Tasks on German SA (Ruppenhofer et al., 2014) provided human annotated datasets of political speeches (STEPS task) and reviews about products (StAR task) like coffee machines and washers. The StAR task focused on the extraction of evaluative phrases (e.g., "*bad*") and aspect expressions (e.g., "*washer*"). The STEPS dataset includes annotations for evaluative phrases, opinion targets, and the corresponding sources (opinion holders). The 'Concept-Level Sentiment Analysis Challenge' of ESWC 2014 included in its tasks the extraction of aspects of each sentence and a sentiment score (positive or negative) per aspect using the dataset of Blitzer, Dredze, and Pereira (2007), which contains customer reviews of DVDs, books, kitchen appliances, and electronic products, with an overall sentiment score for each review; the aspects were intended to be concepts from ontologies, not simply aspect terms.

All the above datasets are monolingual and focus mainly on the English language. The USAGE corpus (Klinger and Cimiano, 2014) consists of annotations of Amazon reviews in German and English for eight product categories ("washing machine", "coffee machine", "trash can", "microwave", "vacuum cleaner", "dish washer", "toaster", and "cutlery") and is annotated with aspects, subjective evaluating phrases, polarities and their relation. Multilingual datasets provide additional benefits enabling the development and testing of cross-lingual methods (Lambert, 2015).

In general, as in many NLP tasks, human annotated datasets are of critical importance in ABSA too, not only for development and training purposes, but also for evaluating the proposed methods and techniques in each case. However, the creation of high quality benchmark data is a labor intense task. The common practice is to construct annotated corpora to feed specific (types of) algorithms that aim to address specific types of research problems in each case (e.g. aspect term extraction or aspect category detection in different domains of interest), instead of performing a systematic annotation study and report, for example, the main annotation problems in each case,

and if and how they have been resolved. In other words, the computational framework conquers and somehow determines the conceptual framework.

2.1.3 METHODS AND TECHNIQUES

Several ABSA methods have been proposed for various domains including consumer electronics (Hu and Liu {2004a, 2004b}), movies (Thet, Na and Khoo, 2010), restaurants (Ganu, Elhadad and Marian, 2009), and Beers, Pubs, Toys, Games and Audiobooks (McAuley, Leskovec and Jurafsky, 2012), among others. Some methods treat aspect extraction and sentiment classification separately (Mei et al., 2007; Brody and Elhadad, 2010), while other approaches model the two problems jointly (Jo and Oh, 2011; Lakkaraju et al., 2014). Some methods adopt domain-independent solutions (Lin and He, 2009), while other make use of domain-specific knowledge to improve their results (Thet, Na and Khoo, 2010). Each approach exploits a variety of features to address aspect detection and sentiment classification. The basic types of features used for ABSA are summarized below:

- Lexical features e.g. n-grams, Token shape
- Morpho-Syntactic features e.g. Lemma, Part-Of-Speech (POS), Dependency trees
- Semantic features e.g. Word clusters, Semantic Dependencies
- Lexicon based features e.g. Sentiment Lexica, WordNet
- Word Vector Representations e.g. Word2Vec (Mikolov et al., 2013), GloVe (Pennington, Socher and Manning, 2014)

Early ABSA approaches were based on *word frequencies* (nouns and noun phrases) in the text, with the assumption that aspect words were more likely to be repeated (Hu and Liu, {2004a, 2004b}; Scaffidi et al., 2007). For example, Hu and Liu ({2004a, 2004b}) capture high frequency feature words by using association rules and generate a summary by using high frequency feature words and ignoring infrequent features. Ding, Liu and Yu (2008) further improved this method by manually adding some rules to handle different kinds of sentence structures. However, a limitation of this approach is that the capability of recognizing phrase features is limited by the accuracy of recognizing noun-group boundaries (Jin and Ho, 2009). In addition, this approach may work well if the text contains high-frequency terms, but may fail if terms are infrequent.

Another line of research exploits *syntactic relations* (Zhang et al., 2010; Qiu et al., 2011; Poria, Cambria and Gelbukh, 2016) focusing on rule-based linguistic patterns

that first identify sentiment words, and then use grammatical relations to build the syntactic structure of sentences and to detect the aspects. For example, Qiu et al. (2011) use a double-propagation method for the bidirectional transfer of sentiment values onto targets and back to unknown sentiment terms based on dependency relations. The lexical relation between sentiment words and aspects is the key element in this method, which is able to identify low-frequency aspects (Schouten and Frasincar, 2016). A key advantage of the syntactic methods is that they only need a small seed set to work properly and do not require human labeled data as compared to supervised approaches. A drawback is reliance on grammatical accuracy of the sentence and the requirement for manipulation (Poria, Cambria and Gelbukh, 2016).

Another type of unsupervised approach is based on *topic modelling* (Lin and He, 2009; Moghaddam and Ester, 2011; Titov and McDonald, {2008a, 2008b}; Brody and Elhadad, 2010; Jo and Oh, 2011) usually using Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan, 2003) to learn distributions of words used to describe each aspect; the topics categories usually comprise of a set of words, and a topic distribution indicates the proportion of a document that discusses each topic. Hence, each topic can be considered an aspect category represented by a set of descriptive words. In this way both explicit and implicit aspects can be detected, however, given that the generated topics are unlabeled, there is no direct correspondence between them and specific aspects.

Supervised learning approaches (Jin and Ho, 2009; Jakob and Gurevych, 2010; Choi and Cardie, 2010) usually treat aspect extraction as a sequential labelling task using Conditional Random Fields (CRF) (Lafferty, McCallum and Pereira., 2001) or Hidden Markov Models (HMMs) (Rabiner, 1989); sequences of words are labelled based on hidden state sequences using a variety of features that rely on labelled training data. Typical features include syntactic structures and lexical features (Jakob and Gurevych, 2010; Toh and Su, 2016; Hamdan, Bellot and Bechet, 2015), cross domain knowledge based features (Jakob and Gurevych, 2010; Mitchell et al., 2013), and more recently features learned by deep learning models (Yin et al., 2016; Li and Lam, 2017).

Deep learning approaches have become very popular over the past several years due to their effectiveness in many Artificial Intelligence tasks. Several researchers (Dong et al. 2014; Lakkaraju et al., 2014; Nguyen and Shirai, 2015; Wang et al. 2016) utilize deep neural networks to generate dense vector representations (embeddings) of sentences and then feed them to classifiers as low dimensional feature vectors. Some approaches (Wang et al. 2016) enhance the representation using for example the

attention mechanism, usually a multi-layer neural network that takes as input the word sequence and aspects, and quantifies for each word of the sentence its sentiment salience as well as the relevance to the given aspect.

A detailed overview on ABSA state-of-the-art methods is available by (Schouten and Frasincar, 2016; Rana and Cheah, 2016), while Do et al. (2018) provide a comparative review of ABSA deep learning approaches. As already mentioned earlier in the introduction, the datasets that were generated by this thesis and adopted to support the first shared task on ABSA organized in the context of SemEval for three years (2014-2016) provided a common test bed for ABSA methods attracting numerous system submissions from a significant number of participants. The best performing approaches in each case are presented below in the respective sections (SE-ABSA14: 2.2.5, SE-ABSA15: 2.3.5, SE-ABSA16: 2.4.2).

2.2 BUILDING AN ANNOTATION FRAMEWORK FOR ABSA

This section presents the first part of this thesis work in the context of ABSA that was to examine how existing definitions are applied to datasets for both fine- and coarse-grained ABSA in order to build a respective annotation framework. The starting point was the restaurants and laptop reviews datasets of Pavlopoulos (2014) that combined annotations for coarse aspect categories and aspect terms along with their sentiment polarity (2.2.1). Following the definitions of aspect (term or category) as a part/component of an entity, an attribute of an entity, or an attribute of a part/component of an entity (Liu, 2006; Zhang and Liu, 2014) and based on a systematic annotation study, a data-driven ABSA annotation codebook was compiled for the first time (2.2.2), and was then applied to the aforementioned datasets that were extended with new unseen sentences (2.2.3). The annotations -as well as the annotation guidelines- were finalized after several iterations in order to ensure consensus among the annotators and consistent annotations. The proposed benchmark datasets (2.2.4), along with the annotation framework and the guidelines were adopted from the ABSA shared task that was organized in the context of the SemEval 2014 workshop and provided for the first time a common evaluation framework for (both coarse- and fine-grained) ABSA (2.2.5). This section concludes with a discussion of the key lessons learned during the systematic annotation study indicating the need for new definitions and a principled unified ABSA framework (2.2.6).

2.2.1 DATASETS AND ANNOTATION SCHEMA

The restaurant reviews dataset of Pavlopoulos (2014) combined annotations for aspect terms and aspect categories, along with their sentiment polarity. In particular, it was a subset (3,710 English sentences) of the reviews dataset of Ganu, Elhadad and Marian (2009) that included annotations for coarse aspect categories and overall sentence polarities; the dataset was modified to include annotations for aspect terms occurring in the sentences, aspect term polarities, and aspect category-specific polarities. For example, sentence (1) in addition to the existing annotation {category: "FOOD"} was enriched with the annotations: {category= "FOOD": "*positive*", aspect term= "*dessert*": "*positive*"}.

1. *The dessert was divine*.
2. *The restaurant was expensive, but the menu was great.*

Similarly, sentence (2) was tagged as follows: {category= "PRICE": "*negative*", category= "FOOD": "*positive*", aspect term= "*menu*": "*positive*"}. Even though the sentence refers also to the prices, and a possibility would be to add also an annotation of an implicit aspect term (e.g. "price" or "expensive"), since it is not mentioned in an explicit manner as in the case of the terms "*dessert*" and "*menu*" that are explicit mentions of food, no aspect term annotation was provided.

The laptops dataset consisted of 3,085 English sentences of 394 online customer reviews and contained annotations only for aspect terms and their sentiment polarity e.g. "*The screen is bright and the keyboard is nice".* → {aspect term: "*screen*": "*positive*", aspect term: "*keyboard*": "*positive*"}.

The author of this thesis was asked to inspect the existing annotations, identify possible inconsistencies, proceed to the appropriate modifications/corrections and compile a respective set of annotation guidelines in order to make these two datasets benchmarks that could support a shared task on ABSA in the context of the 8th International Workshop on Semantic Evaluation (SemEval 2014) following the ABSA task decomposition (annotation schema) proposed by Pavlopoulos (2014). In particular, given a sentence from a customer review about a target entity of interest - a restaurant or laptop-, the task of an annotator (system or human) was to identify the following types of information:

- **Aspect Terms**: Single or multiword terms explicitly naming particular aspects of the target entity (i.e. a restaurant or a laptop). For example, in "*I liked the service*

and the staff, but not the food", the aspect terms are "*service*", "*staff*" and "*food*"; in "*The hard disk is very noisy*" the only aspect term is "*hard disk*".

- **Aspect term polarity**: Each aspect term is assigned one of the following polarities based on the sentiment that is expressed in the sentence about it: Positive, negative, conflict (both positive and negative sentiment), or neutral (neither positive nor negative sentiment). For example, in "*I hated their fajitas, but their salads were great*", the aspect term "fajitas" has negative polarity and "salads" has positive polarity; in "*The fajitas were their starters*", "fajitas" has neutral polarity; and in "*The fajitas were great to taste, but not to see*", "fajitas" has conflict polarity.

For the restaurants domain, two further types of information were included:

- **Aspect category**: Each sentence is assigned one or more aspect category labels based on the six-way schema of Ganu, Elhadad and Marian (2009): FOOD, SERVICE, PRICE, AMBIENCE, ANECDOTES, and MISCELLANEOUS. The first four categories are typical parameters of restaurant reviews (e.g. Zagat ratings). ANECDOTES is used for sentences describing the reviewer's personal experience or context, but that do not usually provide information on the restaurant quality (e.g. "*I knew upon visiting NYC that I wanted to try an original deli*"), while MISCELLANEOUS for sentences that do not belong to the other five categories including sentences that are general recommendations (e.g. "*Your friends will thank you for introducing them to this gem!*")

- **Aspect category polarity**: Similarly to aspect terms, each aspect discussed by a particular sentence had to be assigned one of the following polarities based on the sentiment that is expressed in the sentence about it: positive, negative, conflict, neutral. For example, in "*The restaurant was expensive, but the menu was great*", the aspect category PRICE has negative polarity, whereas FOOD has positive polarity.

2.2.2 ANNOTATION STUDY AND GUIDELINES

The inspection of all the annotations revealed issues and inconsistencies having to do mainly with what is annotated as aspect term in each case, with the boundaries of multi-word aspect terms as well as with sentiment polarity ambiguity cases. The author made a selection of sentences that were representative of the problematic cases detected during the annotation study and asked for feedback from another

computational linguist as well as from two senior computer scientists. Based on literature findings, the findings during the systematic annotation study and the feedback received, a data-driven codebook was compiled in order to resolve problematic cases and to achieve consistency. To this end, definitions of each information unit that should be annotated in each case were provided along with examples and exceptions as presented below in the following sections (2.2.2.1-2.2.2.3).

2.2.2.1 ASPECT TERMS ANNOTATION

For a given target entity -a restaurant or laptop-, the task of the annotator was to identify nouns, nominal phrases, verbs or verbals (words formed from a verb, but functioning as a different part of speech e.g. gerunds and participles) explicitly naming particular aspects of the given target entity, as indicated in bold in the following examples:

1. *The **screen** is bright and the **keyboard** is nice.*
2. *Of course, I also have several great **software packages** that came for free including **iWork**, **GarageBand**, and **iMovie**.*
3. *Fresh, delicious, and reasonably **priced**.*
4. *It is pretty sweet when you want **gaming** on the laptop.*

The identified aspect terms should be annotated as they appear, even if misspelled e.g

5. *Still under **warrenty** so called Toshiba, no help at all.*

The identified aspect terms should be annotated even if they appear in quotation marks or brackets. Notice that “okra (bindi)” is a single aspect term below:

6. *I recommend the **garlic shrimp**, **okra (bindi)**, and anything with **lamb**.*

If an aspect term appears in a sentence more than once, then all of its occurrences in the sentence should be annotated e.g.

7. *The only disappointment was the **coat check girls** who didn't seem to know what a customer is on a relatively non-busy night (for the **coat check girls**).*

In order to facilitate the annotation task and avoid inconsistencies, the guidelines include also information of the types of information that are not considered aspect terms in the context of the specific annotation framework. In particular, no aspect term annotations should be provided for:

- References to the target entity (the restaurant or laptop the review is about) as a whole (e.g. "*product", "restaurant"*) and mentions of other entities (e.g. "*New York City"* below):

8. *Great product.*
9. *This is my favorite Italian restaurant in all of New York City.*

- The name, the type or the model of the laptop, including the name of the manufacturer (e.g. "*Notebook*", "*Toshiba Qosmio*", "*Toshiba*") or the name of the restaurant (e.g. *Rao's*):

10. *The Notebook PC, Toshiba Qosmio is the best gift my father could have ever gotten me.*
11. *I was at Rao's last Wed.*

- Pronouns (e.g. "it", "they", "this") even if they refer to an aspect. For example, "*it*" should not be annotated below.

12. I love the ***screen***, it is amazing.

- Implicit aspect terms, i.e., aspect terms that are not explicitly mentioned, but can be inferred from subjectivity indicators (i.e. words/phrases expressing opinion, evaluation etc.) or other expressions. For example, sentence (13) refers to an implicit aspect term "price", because of the adjective "*inexpensive*". Only explicitly mentioned aspect terms should be annotated, like "*prices*" in sentence (14):

13. *I picked it out because it was inexpensive ($400).*
14. ***Prices*** are in line.

Similarly, subjectivity indicators (e.g. "*malfunction"*) are not considered parts of aspect terms.

15. *It had a* ***cooling system*** *malfunction after 10 minutes of general use, and would not move past this error.*

However, some terms (e.g. "*fresh*" in the following examples) can be used both as parts of aspect terms (e.g. in 16) as well as subjectivity indicators (e.g. in 17):

16. *Both the **fresh mozzerella slices** and the **Plain Cheese slice** are phenomenal.*
17. *The **food** is fresh, delicious, and reasonably priced.*

Finally, terms that are often used as aspect terms, for example "*screen*" in (1), may not always be aspect terms; for example, "*blue screen crash*" in (18) is an operating system malfunction. Similarly, the aspect term in (19) is the "*backlit keyboard*", while in (20) only the "*keyboard*" which is not backlit. In sentence (21), "*place*" refers to the restaurant as a whole and, hence, is not an aspect term; by contrast, in (22) "*place*" is an aspect term referring to the space or room of the restaurant.

18. *It gave me a blue screen crash twice.*
19. *There is a **backlit keyboard** which is perfect for typing in the dark.*
20. *No backlit **keyboard**, but not an issue for me.*
21. *Would recommend - perfect for those looking for a place close to grand central.*
22. *The **staff** was accomodating, the **food** was absolutely delicious and the **place** is lovely.*

Focusing on the boundaries of the multiword aspect terms, the decision was to annotate the maximal phrase as illustrated in the following examples:

23. *The **cover for the DVD drive** soon came off, too--a mark of poor **construction quality**.*
24. *I ordered the **smoked salmon and roe appetizer** and it was off flavor.*
25. *The **noodle and rices dishes** taste great.*

Notice that in (25) the entire conjunction "*noodle and rices dishes*" has been annotated as a single term, while in (24), there is only one aspect term: the "*smoked salmon and roe appetizer*", since this is a single dish, rather than two separate aspect terms "*smoked salmon*" and "*roe appetizer*". As it is illustrated in the above examples, determiners (e.g. "a", "the", "some", "many", "all") are not be included in aspect terms, unless they are parts of embedded noun phrases e.g. "***cover for <u>the</u> DVD drive***" in (23).

2.2.2.2 ASPECT CATEGORY ANNOTATION

The inspection of the original aspect category annotations provided by Ganu, Elhadad and Marian (2009) did not reveal any major issues or inconsistencies. With the exception of some few missing annotations (e.g. the AMBIENCE category was missing in the sentence "*With the theater 2 blocks away we had a delicious meal in a beautiful room*"), the only problem was that the distinction between the categories ANECDOTES

and MISCELLANEOUS was not always clear. Hence, these two categories have been merged to one (ANECDOTES/MISCELLANEOUS) for the purposes of this annotation framework and was used for sentences not belonging in any of the previous four aspect categories (e.g. *Overall I would recommend it and go back again*. → ANECDOTES/MISCELLANEOUS)

In this setting, a sentence may be classified into one or more aspect categories (FOOD, PRICE, AMBIENCE, SERVICE, ANECDOTES/MISCELLANEOUS) based on its overall meaning. An important notice is that aspect categories may not necessarily occur as terms in the sentence; for example, the sentence "*Anybody who likes this place must be from a different planet, where greasy, dry, tasteless and complimentary*" discusses the aspect category FOOD, without mentioning particular aspect terms related to the food as compared for example to the sentence "*While the ambiance was great, the food and service could have been a lot better",* where the categories AMBIENCE, FOOD, and SERVICE are explicitly mentioned through the aspect terms "*ambiance*", "*food*", and "*service*", respectively.

2.2.2.3 ASPECT (TERM/CATEGORY) SENTIMENT POLARITY ANNOTATION

Each identified aspect term or category should be classified as "*positive*", "*negative*", or "*conflict*" if the sentiment that is expressed in the sentence about it is positive, negative or both positive and negative, respectively as illustrated in the following examples:

1. *Other than not being a fan of click pads (industry standard these days) and the lousy internal speakers, it's hard for me to find things about this notebook I don't like, especially considering the $350 price tag*. →
   {"*click pads*": negative, "*internal speakers*": negative, "*price tag*": positive}

2. *Small screen somewhat limiting but great for travel.* → {"*screen*": conflict}

3. *The sweet lassi was excellent as was the lamb chettinad and the garlic naan but the rasamalai was forgettable* →
   {"*sweet lassi*": positive, "*lamb chettinad*": positive, "*garlic naan*": positive, "*rasamalai*": negative}, {FOOD: conflict}

4. *My husband had the mesclun, salmon, and ice cream and he enjoyed all 3 courses.* →
   {"*mesclum*": positive, "*salmon*": positive, "*ice cream*": positive, "*courses*": positive}, {FOOD: positive}

Notice that in sentences (1-3), the opinion holder is the reviewer, whereas in sentence (4), the opinion holder is a third person.

As for the "neutral" polarity, it applies to the following cases:

- When factual information (no sentiment) about the aspect (term or category) is provided e.g. *Went there for an office lunch.* → {"*office lunch*": neutral}, {ANECDOTES/MISCELLANEOUS: neutral}

- When a neutral sentiment, wish, or desire toward the aspect term is expressed, e.g. *I would like at least a 4 hr. battery life* → {"*battery life*": neutral}

- When positive or negative polarity about the named aspect might be inferred, without being explicit e.g. *We were told that the wait was about twenty minutes and there would be no problem for our 8:00 pm curtain call.* → {"*wait*": neutral}

- When expression like "*moderate*", "*in line*", "*nothing out of the ordinary*", "*not an issue*" etc. are used e.g.
  No backlit keyboard, but not an issue for me. → {"*keyboard*": neutral}
  Prices are in line. → {"*prices*": neutral}

If a sentence conveys both neutral and negative/positive opinions about an aspect category, then the negative/positive polarities dominate over the neutral ones. There are also cases like in the following example, where a positive opinion is expressed about the menu, but there are no opinions for its particular items: *The menu was impressive with selections ranging from a burger, to steak, to escargot* → {"*burger*": neutral, "*steak*": neutral, "*escargot*": neutral}, {FOOD: positive}.

2.2.3 ANNOTATION PROCESS

The datasets of Pavlopoulos (2014) were extended with new unseen sentences and were modified or annotated form scratch in the case of the new sentences according to the guidelines. Each sentence was inspected/annotated by two annotators, the author this thesis (annotator A) and a graduate student in computational linguistics (annotator B). The annotators used BRAT (Stenetorp et al., 2012), a web-based annotation tool, which was configured appropriately for the needs of the ABSA task. Fig. 7 shows an annotated sentence in BRAT, as viewed by the annotators.

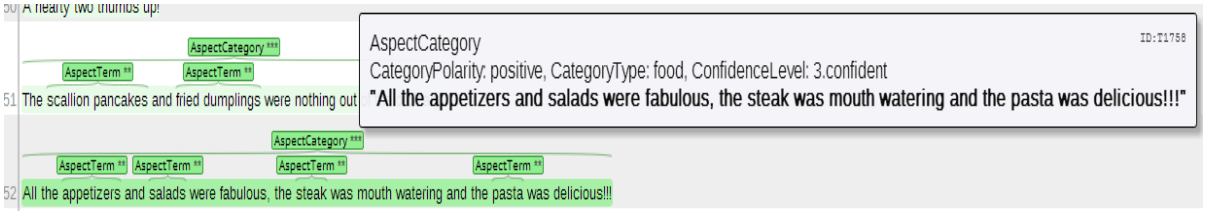


Figure 7: Example of an annotated sentence in the BRAT annotation tool

In order to facilitate the process, for each annotation, the annotators had to provide a confidence level rating according to the following three-level scale: 1. Not Confident (i.e. the annotator is not sure about an aspect term) 2. Average Confidence (i.e. the annotator is confident about the aspect term but not about its polarity) 3. Confident. The two annotators worked collaboratively to resolve non-confident cases. When A and B disagreed, a decision was made collaboratively by them and a third annotator (a computer scientist).

The disagreements between the two annotators were confined to borderline cases. Most uncertainties and disagreements were due to the lack of context, since the datasets consist of isolated sentences taken from customer reviews. In several sentences, it was unclear if the reviewer expressed positive or negative opinion, or no opinion at all (just reporting a fact), due to lack of context. For example, in "*12.44 seconds boot time*" it is unclear if the reviewer expresses a positive, negative, or no opinion about the aspect term '*boot time*'. Similarly, in some cases, it was unclear if a noun or noun phrase was used as the aspect term or if it referred to the entity being reviewed as whole. In "*This place is awesome*", for example, '*place*' most probably refers to the restaurant as a whole (hence, it should not be tagged as an aspect term), but in "*Cozy place and good pizza*" it probably refers to the ambience of the restaurant. A broader context would again help in some of these cases.

Other disagreements concerned the extent of the aspect terms when adjectives that may or may not have a subjective meaning were also present. For example, if 'large' in "*large whole shrimp*" is part of the dish name, then the guidelines require the adjective to be included in the aspect term; otherwise (e.g. in "*large portions*") 'large' is a subjectivity indicator not to be included in the aspect term. Despite the guidelines, in some cases it was difficult to isolate and tag the exact aspect term, because of intervening words, punctuation, or long-term dependencies.

Overall, the laptops domain proved to be harder for the annotators than the restaurants one, since it involves more entities (e.g. hardware and software components) and complex concepts (e.g. usability, portability) that are often discussed implicitly in the

text. Determining the aspect categories of the sentences and their polarities was an easier task compared to detecting aspect terms and their polarities. The annotators needed less time and it was easier to reach agreement. Exceptions were some sentences where it was difficult to decide if the categories AMBIENCE or ANECDOTES/MISCELLANEOUS applied (e.g. "*One of my Fav spots in the city*"). The decision was to classify those sentences as ANECDOTES/MISCELLANEOUS only if they conveyed general views about a restaurant, without explicitly referring to its atmosphere or environment.

The annotations -as well as the annotation guidelines- were finalized after several iterations in order to ensure consensus between the annotators and consistent annotations. Despite the detailed guidelines, still some annotation decisions were considered borderline cases mainly due to the lack of context. When the annotations were finalized, the datasets were further refined by removing some sentences (duplicates or very similar sentences, problematic cases that were left for future research e.g. comparative opinions). The annotation process resulted in two datasets that consist of a total of 7686 sentences from customer reviews about laptops and restaurants, containing a total of 7839 manually annotated aspect terms along with their sentiment polarity (see below Table 2).

2.2.4 THE ABSA-2014 BENCHMARK DATASETS

The laptops (LPT) dataset consists of 3845 sentences annotated with a total of 3012 aspect terms, and the restaurants (RST) dataset consists of 3845 sentences annotated with a total of 4827 aspect terms.

Dataset	Sentences	Aspect Terms				
		Positive	Negative	Conflict	Neutral	Total
LPT	3845	1328	994	61	629	3012
RST	3841	2892	1001	105	829	4827
Total	7686	4220	1995	166	1458	7839

Table 2: Sizes of the datasets and number of aspect terms annotations and their polarities per domain

The majority of the aspect terms are single-words in both datasets (2148 in laptops, 4827 in restaurants, out of 3012 and 4827 total aspect terms, respectively). As it is illustrated above in Table 2, restaurants reviews contain many more aspect terms than laptop reviews. This is because, as it was observed during the annotation process, laptop reviews often evaluate each laptop as a whole, rather than expressing opinions

about particular aspects. Furthermore, when they express opinions about particular aspects, they often do so by using adjectives that refer implicitly to aspects (e.g. 'expensive', 'heavy'), rather than using explicit aspect terms (e.g. 'cost', 'weight'); the annotators were instructed to tag only explicit aspect terms, not adjectives implicitly referring to aspects. On the other hand, restaurants reviews contain many references to specific dishes that are evaluated. Another difference between the two datasets is that the neutral class is much more frequent in (the aspect terms of) the laptops dataset, since laptop reviews often mention features without expressing any (clear) sentiment (e.g. "*the latest version does not have a disc drive*"). Nevertheless, the positive class is the majority in both datasets, but it is much more frequent in the restaurants dataset.

The restaurants dataset contains two additional annotation layers: aspect categories and the respective sentiment polarities. As illustrated below in Table 3, FOOD is the most frequent category which is not a surprise, since customers evaluate mainly their eating experience in a restaurant. Again, the positive class is the majority one.

Category	Positive	Negative	Conflict	Neutral	Total
FOOD	1169	278	82	121	1650
PRICE	230	143	20	11	404
SERVICE	425	281	40	23	769
AMBIENCE	339	119	60	31	549
ANEC./MISC.	673	240	45	408	1366
Total	2836	1061	247	594	4738

Table 3: Aspect categories distribution per sentiment class

The proposed datasets, along with the annotation framework and the guidelines were adopted from SE-ABSA14 shared task that was organized in the context of the SemEval 2014 workshop and provided for the first time a common evaluation framework for (both coarse- and fine-grained) ABSA (section 2.2.5). Based on the experience of the annotation study and process, the expectations were that systems would perform better in aspect category detection than in aspect term extraction, and that the restaurants domain would be easier.

2.2.5 THE ABSA-2014 SEMEVAL CHALLENGE

The International Workshop for Semantic Evaluation (SemEval) is an ongoing series of evaluations of computational semantic analysis systems. The evaluations are intended to explore the nature of meaning in language and to provide frameworks for the development of robust systems for a variety of semantic analysis tasks, with SA being one of them. The first SemEval shared task on ABSA was introduced in 2014. The rationale behind organizing the SE-ABSA14 task[23] was to provide a common evaluation framework for the two main state-of-the-art ABSA representations that is coarse- and fine-grained ABSA focusing on restaurants and laptops customer reviews, and in particular on the annotation framework and the datasets described above in the previous sections. Accordingly, the task consisted of four subtasks:

- Subtask 1 (SB1): Aspect Term Extraction (ATE).
- Subtask 2 (SB2): Aspect Term Polarity Detection (ATP).
- Subtask 3 (SB3): Aspect Category Extraction (ACE).
- Subtask 3 (SB4): Aspect Category Polarity Detection (ACP).

Participants were free to choose the subtasks and domains they wished to participate in. The task provided training and testing data on both domains (restaurants and laptops) for the first two subtasks (SB1, SB2), and only for the restaurants domain for the last two subtasks (SB3, SB4). In particular, the datasets described in the previous section were split for training (TR) and testing (TE) purposes as illustrated below in Table 4 and Fig. 8. In addition, the task provided baselines for each subtask and domain.

Dataset	Sentences	Aspect terms				
		Positive	Negative	Conflict	Neutral	Total
LPT-TR	3045	987	866	45	460	2358
LPT-TE	800	341	128	16	169	654
RST-TR	3041	2164	805	91	633	3693
RST-TE	800	728	196	14	196	1134

Table 4: Sizes and aspect term annotations of the TR and TE datasets per domain

[23] http://alt.qcri.org/semeval2014/task4/

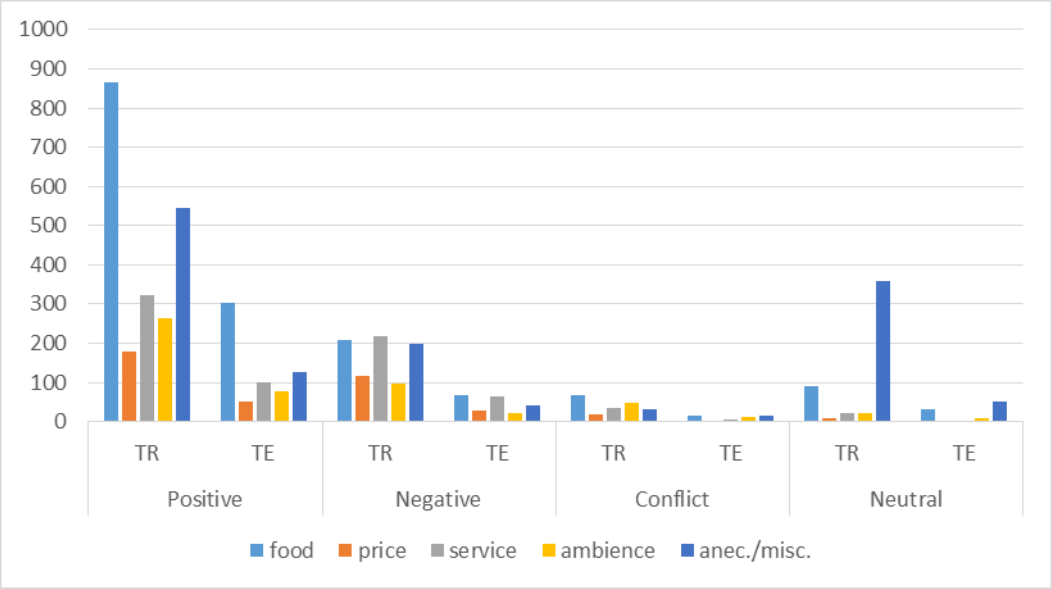


Figure 8: Aspect categories distribution per sentiment class in the TR and TE datasets

The task attracted 165 submissions from 32 teams that experimented with a variety of features (e.g. based on n-grams, parse trees, named entities, word clusters), techniques (e.g. rule-based, supervised and unsupervised learning), and resources (e.g. sentiment lexica, Wikipedia, WordNet). The evaluation of the submitted systems ran in two phases. In Phase A, participants were asked to return the aspect terms (SB1) and the aspect categories (SB3) for the provided test datasets. Subsequently, in Phase B, the participants were given the gold aspect annotations for the sentences of Phase A and they were asked to return the polarities of the aspect terms (SB2) and the polarities of the aspect categories of each sentence (SB4). ATE (SB1) and ACE (SB3) were evaluated using the F1 measure defined as usually. For ATP (SB2) and ACP (SB4) the accuracy of each system was calculated, defined as the number of correctly predicted aspect term or aspect category polarity labels, respectively, divided by the total number of aspect term or aspect category annotations.

The expectations that systems would perform better in ACE than in ATE were confirmed; the best score in ACE was 88.57%, whilst the best scores for ATE were 84.01% and 74.55% for the restaurants and laptops domain, respectively. Furthermore, as it is indicated by these scores and was also expected based on the annotation experience, the systems achieved significantly higher scores (+10%) in the restaurants domain, as compared to laptops. The best scores in SB1 were achieved by methods that modelled ATE as a sequential labeling task using CRFs, along with POS and dependency tree based features, and word clusters created from additional reviews from YELP and Amazon (Toh and Wang, 2014; Chernyshevich, 2014). In SB3 the best scores were achieved by SVM methods. For example, the NRC team

(Kiritchenko et al., 2014) relied on five binary (one-vs-all) SVMs, one for each aspect category; the SVMs used features based on various types of n-grams (e.g., stemmed) and information from a lexicon learnt from YELP data, which associates aspect terms with aspect categories. Another effective approach was the hybrid approach of the XRCE team (Brun, Popa and Roux, 2014) that used information identified by its syntactic parser as well as BoW features to train a logistic regression model that assigns to the sentence probabilities of belonging to each aspect category. Finally, in the sentiment polarity classification subtasks (SBs 2 and 4) the best scores were achieved by SVMs with features mainly based on n-grams, parse trees, and several out-of-domain, publicly available sentiment lexica (e.g. MPQA, SentiWordnet and Bing Liu's Opinion Lexicon) which had a significant impact on systems' performance (Kiritchenko et al., 2014; Brun, Popa and Roux, 2014; Wagner et al., 2014). More details about the submitted systems and the evaluation results are available at the task overview paper (Pontiki et al., 2014).

2.2.6 DISCUSSION AND LESSONS LEARNT

This section presented the first part of this thesis work in the context of ABSA that was to examine how existing definitions and ABSA representations are applied to datasets for both fine- and coarse-grained ABSA. Based on a systematic annotation study, a data-driven annotation codebook was compiled and applied to existing customer reviews datasets on the restaurants and the laptops domain. The proposed benchmark datasets, along with the annotation framework and the guidelines were adopted from the SE-ABSA14 shared task. The task attracted 165 submissions from 32 teams that experimented with a variety of features, techniques, and resources achieving high scores in both fine-grained (ATE and ATP) and coarse-grained ABSA (ACE and ACP).

However, the observations made during the annotation process indicated that there were still issues that needed to be addressed in the ABSA representation having to do with completeness and meaningfulness. On the one hand, many opinionated sentences are not annotated with this type of ABSA representation, since it does not include implicit aspects and mentions of the target entities. This lack of completeness is very noticeable especially in the laptops domain, which, in addition, lacks annotations for coarse aspect categories. Hence, there is a need for additional information types (annotation layers) like aspect categories for the laptops domain, implicit aspects and mentions of the target entities.

On the other hand, the existing annotations do always reflect the intended meaning of the texts and aspect terms cannot always be linked to aspect categories in order to construct a more structured and meaningful output. For example, in sentence (1) we can we link "*pizza*" to FOOD and PRICE and resolve the conflict polarity label, but in sentence (2) the term "*appetizers*" cannot be linked to the category SERVICE:

1. *The pizza was pricey but delicious*.

   → AT: {*pizza*: conflict}

   → AC: {PRICE: negative, FOOD: positive}

2. *Appetizers took over an hour*.

   → AT: {*Appetizers*: neutral}

   → AC: {SERVICE: negative}

Another important issue is the "*aspect of the aspect*" problem resulting from the fact that "aspect" indicates both entities and attributes. For example, sentences (3) and (4) discuss the same aspect(s), the quality of the food. However, what is annotated as aspect depends on its lexicalization every time ("*food quality*" as unique aspect term or "*food*" and "*quality*" as separate terms). Similarly, sentences (5) and (6) discuss the same aspects (the quality and the price of the food), but this is not reflected in the resulting annotations.

3. ***Food quality*** *is excellent.*

4. *The* ***food*** *was poor* ***quality****.*

5. ***Quality of food*** *is excellent and* ***price*** *is cheap.*

6. *The* ***food*** *is cheap and the* ***quality*** *excellent.*

However, sentences that discuss the same aspects should be assigned common labels irrespectively of the linguistic evidence used. Hence, there is a need for an ABSA representation framework that distinguishes:

a) what is meant from what is written (the intended meaning vs the linguistic evidence), and

b) entities from attributes taking into account the "*dyadic relation*" between them, namely that an entity or a part of an entity is evaluated with regard to an attribute (or as a whole).

Furthermore, even though extracting aspect terms in general including aspect terms for which no sentiment is expressed (neutral polarity) is a useful task (e.g. useful for constructing an ontology of aspect terms and to identify frequently discussed aspects), a framework that includes aspects for which no sentiment is expressed may result to misleading outputs (see for example sentence 2). A final, yet important, issue that emerged during the annotation process is the need for context; in order to be able to assign the correct labels in many cases annotators (human or systems) have to consider a broader context (the entire review text) and not each sentence in isolation.

In order to address these issues, the second step of the work presented in this thesis was to design a new ABSA framework that is more structured, meaningful and representative of the intended meaning of the texts (see below section 2.3).

2.3 REDEFINING ABSA

This section presents the second part of this thesis work in the context of ABSA; it aimed at designing and providing a new principled unified ABSA framework (2.3.1) based on the key lessons learned during the first part of this study. Following this new framework, a data-driven ABSA annotation codebook was compiled for three domains (2.3.2), and was then applied to new datasets (2.3.3). The annotations -as well as the annotation guidelines- were finalized after several iterations in order to ensure consensus between the annotators and consistent annotations. The proposed benchmark datasets (2.3.4), along with the annotation framework and the guidelines were adopted from the ABSA shared task that was organized in the context of the SemEval 2015 (2.3.5) as a follow up of the SE-ABSA14.

2.3.1 A PRINCIPLED UNIFIED FRAMEWORK FOR ABSA

As discussed in the previous section, one of the main problems during the annotation process following the state-of-the-art definitions for "aspect" was that the term is used to denote both entities and attributes of these entities; in particular, an aspect (term or category) indicates (Liu, 2006; Zhang and Liu (2014): (a) a part/component of *e* (e.g., battery), (b) an attribute of *e* (e.g., price), or (c) an attribute of a part/component of *e* (e.g., battery life). This results in the "*aspect of the aspect*" problem that has been

discussed above. In addition, these definitions and the up-to-date ABSA settings did not consider the target entity (e.g. laptop, restaurant) as part of the analysis.

In this context, and taking into account that usually an entity is evaluated either in general terms or by focusing on specific attributes, this thesis proposes a new definition of aspect that includes also the target entity and aims to distinguish entities and attributes:

- An aspect is defined as a combination of an entity type E and an attribute type A.
- E can be the reviewed entity e itself (e.g., laptop), a part/component of it (e.g., battery or customer support), or another relevant entity (e.g., the manufacturer of e), while A is a particular attribute (e.g., durability, quality) of E.
- E and A are concept names (classes) from a given domain ontology and do not necessarily occur as terms in a sentence.

In this setting, aspect corresponds more to the notion of the aspect categories of the coarse-grained way of ABSA representation that was based on predefined inventories of classes/ labels like the ones used in Ganu, Elhadad and Marian (2009) or the TripAdvisor. However, the difference is that according to this new definition, an aspect is a dyadic relation formed by the entity or the part of an entity that is evaluated and the attribute with regard to which it is evaluated, rather than a list of single labels like food, price, service, etc. As for the "aspect terms", within this new framework they correspond to the linguistic evidence of the entities E (e.g. service, pizza) or attributes A (e.g., price, quality) under evaluation in the text if any. In this setting, an ABSA tuple is structured as follows:

{E#A, OTE$_E$, OTE$_A$, polarity, Pev}

where:

- E and A are concept names of entity types and attribute labels that are defined for each domain and do not necessarily occur as terms in a text.
- OTE$_E$ and OTE$_A$ correspond to the explicit mentions (linguistic evidence) -if any- of the entities and attributes under evaluation, respectively.
- Polarity is the sentiment class corresponding to the semantic orientation of the expressed opinion about the attribute of an entity (e.g. positive, negative).
- Pev is the linguistic evidence -if any- for the expressed opinion.

Below are some examples of sentences annotated according to this new framework. The examples include also annotations from the previous ABSA framework discussed

in section 2.2 making more clear the differences between them. As it is demonstrated by the examples, this new framework focuses on what is really meant (at the concept level) and not on the way it is lexicalized. As for the linguistic evidence of the basic elements (entities, attributes, polarity) they can be explicitly lexicalized (e.g. *pizza, keyboard, pricey, delicious, backlit, drawback*) or they may not appear in a text (e.g. NULL values in 2 and 4) and can be only implicitly inferred. In some cases, some terms may indicate both attributes and sentiment polarity (e.g. *pricey* and *delicious* in 2).

1. *The pizza was pricey but delicious.*

   Previous framework:

   → AT: {*pizza*: conflict}

   → AC: {PRICE: negative, FOOD: positive}

   New framework:

   → {FOOD#PRICES, OTEE: *pizza*, OTEA: *pricey*, negative, Pev: *pricey*},
   {FOOD#QUALITY, OTEE: *pizza*, OTEA: *delicious*, positive, Pev: *delicious*}

2. *Appetizers took over an hour.*

   Previous framework:

   → AT: {*Appetizers*: neutral}

   → AC: {SERVICE: negative}

   New framework:

   → {SERVICE#GENERAL, OTEE: *null*, OTEA: *null*, negative, Pev: *null*}

3. *One drawback, I wish the keys were backlit.*

   Previous framework:

→ AT: {*keys*: negative}

New framework:

→ {KEYBOARD#DESIGN&FEATURES, OTEE:*keys*, OTEA:*backlit*, negative, Pev: *drawback*}

4. *They sent it back with a huge crack in it and it still didn't work; and that was the fourth time I've sent it to them to get fixed.*

   Previous framework:

   → no annotations

   New framework:

   → {SUPPORT#QUALITY, OTEE: *null*, OTEA: *null*, negative, Pev: *null*}

In this setting, the ABSA problem has been formalized into a unified framework in which all the identified information (opinion target expressions, aspect categories and polarities) are linked to each other and organized in opinion tuples that directly reflect the intended meaning of a sentence.

2.3.2 ANNOTATION SCHEMA AND CODEBOOK

The proposed new ABSA framework is applied to new datasets for the restaurants and the laptops domain that consist of whole review texts instead of isolated sentences. In addition, it is applied at a small scale in a new domain that is hotel reviews. The annotation framework that has been designed in the context of this thesis focuses on aspect categories (defined as E#A pairs for each domain), sentiment polarity and OTEEs. Hence, the ABSA tuple is structured as follows: {E#A, OTE, polarity}, where OTE corresponds to explicit mentions of the entities E and is used for simplicity instead of OTEE. The E#A pair defines an aspect (category). In particular, given a customer review about a particular entity (a restaurant, a hotel, or a laptop), the task (of system/human annotator) is to identify the following types of information:

- **Aspect Category.** Identify the entity E and the attribute A pairs E#A towards which opinions are expressed. The identified entities can be assigned one or more attribute labels based on the context of the sentence they appear in. E and A should be chosen from the inventories of entity types and attribute labels per domain. In particular, the entity E can be assigned 22 possible labels for the laptops domain (e.g. LAPTOP, SOFTWARE, SUPPORT), 6 labels for the restaurants domain (e.g. RESTAURANT, FOOD), and 7 labels for the hotels domain (e.g. HOTEL, ROOMS). The attribute A can be assigned 9 possible labels for the laptops domain (e.g. USABILITY, OPERATION_PERFORMANCE), 5 labels for the restaurants domain (e.g. QUALITY, PRICE), and 8 labels for the hotels domain (e.g. COMFORT, PRICE). The aspect category inventories are described in detail in the following sections (2.3.2.1-2.3.2.3).

- **Opinion Polarity**. Each identified E#A pair of a sentence is assigned a polarity, from a set P = {**positive**, **negative, neutral}**. Contrary to the previous schema, here the ***neutral*** label applies for ***mildly positive or negative sentiment*** (e.g. 3, 4), thus it does not indicate objectivity. For example, sentence (5) below has not been assigned any label, since it conveys only objective information:

  1. *A mac is very easy to use and it simply makes sense.* → {**LAPTOP#USABILITY** , positive}
  2. *And the room did not even have a shower curtain!* → {**ROOMS#DESIGN_FEATURES** , negative}
  3. *Food was okay, nothing great.* →{**FOOD#QUALITY**, neutral}
  4. *Prices are in line.* →{**RESTAURANT#PRICES**, neutral}
  5. *I went to this restaurant with a woman that I met recently.* → {}

  Another difference with the previous annotation framework is that here the "*conflict*" label is not used, since –due to the adopted fine-grained aspect classification schema– it is very rare to encounter (in a sentence) both a positive and a negative opinion about the same attribute A of an entity E. In the few cases that this may happen, the dominant sentiment is chosen (e.g. "*The OS takes some getting used to but the learning curve is so worth it*!" → {OS[24]#USABILITY, positive}).

  The {E#A, polarity} annotations are assigned at the sentence level taking into account the context of the whole review. For example, sentence 6_A is assigned a

[24] OS entity label in the laptops domain is for the operating system.

negative opinion about the quality of the customer support and not about the operation of the laptop, as it is implied by 6_B. Similarly, in 7_A, even though the reviewer starts by saying how happy he/she was with the laptop, he/she is expressing a negative opinion towards the laptop as it can be inferred from 7_B.

6. A. *Horrible customer support-they lost my laptop for a month-got it back 3 months later*. → {**SUPPORT#QUALITY** , negative}
   B. *Laptop still did not work, blue screen within a week...* → {**SUPPORT#QUALITY** , negative}

7. A. *I was so happy with my new Mac*. → {**LAPTOP#GENERAL** , negative}
   B. *For two months…*

- **OTE**. An opinion target expression (OTE) is an explicit reference (mention) to the reviewed entity E of the E#A pair. This reference can be a named entity (e.g. 8), a common noun (e.g. 10) or a multi-word term (e.g. 9). The identified OTEs are annotated as they appear, even if misspelled. When an evaluated entity E is only implicitly inferred or referred to (e.g. through pronouns), the OTE slot is assigned the value "NULL" (e.g. 11-13). This annotation layer is applied only in the restaurants and the hotels domain. Below are some examples:

8. *Leon is an East Village gem*. → {**RESTAURANT#GENERAL**, "***Leon***", positive}
9. *The lobster sandwich is good and the spaghetti with Scallops and Shrimp is great*. → {**FOOD#QUALITY**, "***lobster sandwich***", *positive*}, {**FOOD#QUALITY**, "***spaghetti with Scallops and Shrimp***", positive}
10. *The towels were thin and worn*. → {**ROOM_AMENITIES#QUALITY**, "***towels***", negative}
11. *Everything was wonderful*" → {**RESTAURANT#GENERAL**, NULL, positive}
12. *They never brought us complimentary noodles, ignored repeated requests for sugar, and threw our dishes on the table*. → {**SERVICE#GENERAL**, NULL, negative}
13. *Pleasantly surprised at $69 night."* → {**HOTEL#PRICES**, NULL, positive}

The laptops domain does not include OTE annotations, since most entities are instantiated through a limited set of expressions (e.g. MEMORY: "memory", "ram", CPU: "processing power", "processor", "cpu") as opposed to the restaurants domain, where for example, the entity "FOOD" is instantiated through a variety of food types and dishes (e.g. "pizza", "Lobster Cobb Salad"). Furthermore, LAPTOP which is the majority category label in laptops (see below section 3.3.5) is instantiated mostly

through pronominal mentions, while the explicit mentions are limited to nouns like laptop, computer, product, etc.

Quite often reviews contain opinions towards entities that are not directly related (e.g. part-of, manufacturer) with the reviewed entity, for example, restaurants/hotels that the reviewer has visited in the past, other laptops or products (and their components) of the same or a competitive brand. Such entities as well as comparative opinions are considered to be out of the scope of the proposed annotation framework (examples of such cases per domain are presented in the following sections).

2.3.2.1 ASPECT CATEGORIES FOR THE LAPTOPS DOMAIN

The entity E of an {E#A} pair can be assigned one of the following 22 labels:

- **LAPTOP**: This label is assigned when the reviewed entity is a specific laptop and is being evaluated with regard to particular attributes or in general as a whole (e.g. 1, 7, 9, 15, 21-24, 29, 30, 33, 35-38, 40-50).

- 14 labels that refer to single hardware components, parts or a set of components: **DISPLAY** (*monitor, screen*) (e.g. 16), **CPU** (*processor*) (e.g. 32), **MOTHERBOARD** (e.g. 3), **HARD DISC** (e.g. 31, 43), **MEMORY** (e.g. 43), **BATTERY** (e.g. 10, 11), **POWER SUPPLY** (*charger, charger unit, power supply cord, (power) adapter*) (e.g. 3, 27), **KEYBOARD** (*keys, numpad*), (e.g. 25, 34), **MOUSE** (*mouse pad and the buttons on it*) (e.g. 12, 17, 33), **FANS&COOLING** (*fan, cooling system, heat sink*) (e.g. 26), **OPTICAL DRIVES** (*CD, DVD or Blue-ray players, DVD drive, disc drive, DVD burner*) (e.g. 2), **PORTS** (*USB, HDMI, VGA, card reader, Firewire, SD, DVI, Thunderbolt*) (e.g. 33), **GRAPHICS** (*graphics card, video card, graphics chip*) (e.g. 32, 43), and **MULTIMEDIA DEVICES** (*sound, audio, microphone, (built-in) camera, webcam, speakers, headphone*) (e.g. 28).

- A general entity label **HARDWARE** is used for sentences that refer to the hardware in general or to hardware related entities that do not fall into one of the existing entity types (e.g. 4).

- **OS**: This label applies to sentences discussing the operating system and its features (e.g. *start menu, safe mode, boot manager, drag and drop feature*) (e.g. 13, 18, 32).

- **SOFTWARE**: This label applies for the rest of the software applications (e.g. *Office Suite (Office, iWork, Mac version of Microsoft Office, word processor, Microsoft Word, PowerPoint), browsers, Skype, iPhoto, iLife, photo detection software, Pages, Keynotes, antivirus programs, firewall, games*), as well as for sentences/reviews that refer to the software in general (e.g. 13, 14, 19, 20).

Furthermore, we have the following 4 entity types that refer to the manufacturing company as a brand and to the services/products it provides:

- **WARRANTY** that is provided by the manufacturer (e.g. 39).

- **SHIPPING** for the delivery service when the laptop is bought or during repairs (e.g. 5).

- **SUPPORT** for pre- and after-sales customer support, customer service, repair service, product support, replacement policy and the staff (e.g. 6, 8).

- **COMPANY** for sentences that refer in general to the manufacturing company. For example, in (a-c) the reviewers are expressing opinions not only about the reviewed laptop but also about the manufacturing company by mentioning the brand name (e.g. 41, 42).

The attribute A of an {E#A} pair can be assigned one of the following 9 labels:

- **QUALITY** for opinions referring to the following attributes of an entity: construction/build quality, materials quality, enduring/long-lasting quality (=durability, longevity), broken components, noise, overheating problems, general feel, security (virus-resistant), screen quality (picture quality, screen colors, resolution and clearness), quality of service/shipping e.g.

  1. *I dropped this once from the table when my baby girl grabbed me one day and it is still working with NO issues!* → {**LAPTOP#QUALITY**, positive}
  2. *The DVD burner broke after burning 3 DVD'd during that time!* → {**OPTICAL_DRIVES#QUALITY**, positive}
  3. *The board has a bad connector with the power supply and shortly after warrenty expires the power supply will start having issues*. → {**MOTHERBOARD#QUALITY**, negative}, {**POWER_SUPPLY#QUALITY**, negative}

4. *This is likely due to poor grounding and isolation between the components, and I'm hoping that it can be fixed with a ground loop isolator*. → {**HARDWARE#QUALITY**, negative}
5. *The computer is currently in West Virginia doe to the method of shipping chosen by Toshiba.* → {**SHIPPING#QUALITY**, negative}
6. *Then HP sends it back to me with the hardware screwed up, not able to connect.* →{**SUPPORT#QUALITY**, negative}

Sentence (6) has been assigned the entity label SUPPORT and not HARDWARE, since the reviewer is expressing an opinion towards the quality of the technical support and not about the HARDWARE as in (4).

- **PRICE** for opinions focusing on the price (cheap or expensive) of the laptop and the services provided by the manufacturer (support, shipping and warranty) e.g.

7. *Luckily, for all of us contemplating the decision, the Mac Mini is priced just right*. → {**LAPTOP#PRICE**, positive}
8. *I took it to the shop and they said it would cost too much to repair it*. → {**SUPPORT#PRICE**, negative}

- **OPERATION_PERFORMANCE** for opinions that focus on the operation, the speed, the power, the stability and the responsiveness of an entity, opinions referring to freezing, crashing issues, as well as for opinions evaluating the battery life e.g.

9. *It works exactly like it did the day I took it out of the box.* → {**LAPTOP#QUALITY**, positive}, {**LAPTOP#OPERATION_PERFORMANCE**, positive}
10. *After replacing the hard drive the battery stopped working (3 months of use) which was frustrating.* → {**BATTERY#QUALITY**, negative}, {**BATTERY #OPERATION_PERFORMANCE**, negative}
11. *The battery life seems to be very good.*→ {**BATTERY# OPERATION_PERFORMANCE**, positive}
12. *Sometimes you will be moving your finger and the pointer will not even move.* → {**MOUSE#OPERATION_PERFORMANCE**, negative}
13. *Love the stability of the Mac software and operating system.* → {**SOFTWARE# OPERATION_PERFORMANCE**, positive}, {**OS#OPERATION_PERFORMANCE**, positive}
14. *The Internet Explorer was very slow from the very beginning*. → {**SOFTWARE#OPERATION_PERFORMANCE**, positive}

15. *I got the blue screen of death the first month I got it.* → {**LAPTOP#OPERATION_PERFORMANCE**, negative}
16. *Sometimes the screen even goes black on this computer.* → {**DISPLAY#OPERATION_PERFORMANCE**, negative}

Sentence (10) has been assigned the entity label BATTERY and not HARD DISC, since the reviewer is expressing an opinion about the BATTERY. In (9) and (10) the opinions do no refer only to the OPERATION of the LAPTOP and the BATTERY, but to their QUALITY (durability) too (*like it did the day I took it out of the box, 3 months of use*). Sentence (15) has been assigned the entity label LAPTOP and not DISPLAY, since blue screen issues are related to the operation of the laptop. On the other hand, a black screen (16) or other types of screen issues may be related to the graphics, to the operation of the laptop or the screen itself. Such cases are assigned the entity label DISPLAY.

- **USABILITY** for opinions focusing on the easiness or convenience to use/ learn/ (un)install/ handle/ operate/ set up/ work with/ navigate/ update/ configure/ etc., as well as for opinions evaluating properties like the upgradeability, the compatibility, and ergonomics[25]* e.g.

17. *The mouse jumps around all the time and it clicks stuff I don't want it to!* → {**MOUSE#OPERATION_PERFORMANCE**, negative}, {**MOUSE#USABILITY**, negative}
18. *The OS takes some getting used to especially after being a Windows user for so long!* → {**OS#USABILITY,** neutral}
19. *The applications are also very easy to find and maneuver.* → {**SOFTWARE#USABILITY,** positive}
20. *The only downfall is a lot of the software I have won't work with Mac.* → {**SOFTWARE#USABILITY**, negative}
21. *I had a USB connect but, I can't use it because it is not compatible.* → {**LAPTOP#USABILITY**, positive}
22. *What's really great about this product is you may have a family member who is computer illiterate and you can pretty much just let them loose on this computer without any real supervision.* → {**LAPTOP#USABILITY**, positive}
23. *Memory is upgradable*. → {**LAPTOP#USABILITY**, positive}

[25] Ergonomics* is an attribute that is related both to DESIGN&FEATURES and USABILITY in that a bad/good design of an entity may affect its usability. Therefore, in sentences (xxviii) and (xxix) both attribute labels should be assigned.

24. *Upgrading from Windows 7 Starter, thru Windows 7 Home Premium, to Windows 7 Professional was a snap;* → {**LAPTOP#USABILITY**, positive}

Note that in (20) the opinion expressed refers to the compatibility and not the operation of the software. In (23) and (24) the opinions are expressed about the LAPTOP and not the MEMORY or the OS respectively, since the capability of upgrading them is related to the laptops usability.

- **DESIGN&FEATURES** for opinions focusing on the design, the appearance (shape, color, look), the size, the weight and ergonomics[1]* of an entity, the placement of components, the software design, opinions referring to (extra/missing) features/components, as well as for opinions focusing on the duration and the terms/conditions of the warranty.

25. *The backlit keys are wonderful when you are working in the dark.* → {**KEYBOARD#DESIGN&FEATURES**, positive}
26. *Fan vents to the side, so no cooling pad needed, great feature!* → {**FANS&COOLING#DESIGN&FEATURES**, positive}
27. *The magnetic plug-in power charging power cord is great (I even put it to the test by accident)-excellent innovation*! → {**POWER_SUPPLY# DESIGN&FEATURES**, positive}
28. *I dislike the quality and the placement of the speakers* → {**MULTIMEDIA DEVICES# DESIGN&FEATURES,** negative}, {**MULTIMEDIA DEVICES# QUALITY**, negative}
29. *It also does not have Bluetooth.* → {**LAPTOP# DESIGN&FEATURES**, negative}
30. *The unibody design is edgy and durable.* → {**LAPTOP#DESIGN&FEATURES**, positive}, {**LAPTOP#QUALITY**, positive}
31. *Not to mention it has shit gigs.* → {**HARD_DISC # DESIGN&FEATURES**, negative}
32. *The processor screams, and because of the unique way that Apple OSX 16 functions, most of the graphics are routed through the hardware rather than the software.* → {**CPU #OPERATION_PERFORMANCE**, positive}, {**OS # OPERATION_PERFORMANCE**, positive}, {**GRAPHICS # DESIGN&FEATURES**, positive}
33. *The headphone and mic jack are in front of touch-pad making the touch-pad hard to use when using headphones/mic, not to mention the laptop was designed for right handed person.* →{**PORTS # DESIGN&FEATURES**, negative}, {**MOUSE# USABILITY**, negative}, {**LAPTOP#DESIGN&FEATURES**, negative}, {**LAPTOP# USABILITY**, negative}

34. *I do transcription work on the side, and the flatline keyboard makes typing quick and easy as well.* → {**KEYBOARD#DESIGN&FEATURES**, positive}, {**KEYBOARD#USABILITY**, positive}

- **PORTABILITY** for opinions focusing on the easiness to transfer the laptop and/or use it in limited space e.g.

35. *Very convenient when you travel...* → {**LAPTOP#PORTABILITY**, positive}
36. *This laptop is very large and barely fits in any carrying cases*. → {**LAPTOP#DESIGN&FEATURES**, positive}, {**LAPTOP#PORTABILITY**, positive}

- **CONNECTIVITY** for opinions referring to the ability or the easiness to connect via ports, VGA, HDMI, USB, Bluetooth to peripherals etc., as well as for opinions focusing on wireless and internet connections e.g.

37. *The internet capabilities are also very strong and picks up signals very easily*→{**LAPTOP#CONNECTIVITY**, positive}
38. *I can barely use any usb devices because they will not stay connected properly.* →{**LAPTOP#CONNECTIVITY**, negative}

- **GENERAL** for general opinions expressed about an entity as a whole (e.g. laptop, hardware, software, company) not focusing on any specific attribute.

39. *Also, the extended warranty was a problem.* → {**WARRANTY#GENERAL**, negative}
40. *Do not buy it!* →{**LAPTOP#GENERAL**, negative}
41. *Apple continues to shine and provide a much more enjoyable computer experience!* → {**LAPTOP#GENERAL**, positive}, {**COMPANY#GENERAL**, positive}
42. *I can guarantee this will be the last Dell I will ever purchase!* → {**LAPTOP#GENERAL**, negative}, {**COMPANY#GENERAL**, negative}
43. *It has plenty of memory, lots of hard drive, and great graphics*. → {**MEMORY#DESIGN&FEATURES**, positive}, {**HARD_DISC# DESIGN&FEATURES**, positive}, {**GRAPHICS#GENERAL**, positive}

In sentences (42, 43), the reviewers' negative/positive opinions about the laptop under review are generalized for the manufacturing company (and its products).

- **MISCELLANEOUS** for attributes that do not fall into any of the above cases. Such cases may be:

- Opinions focusing on specific types of a laptop's usage (e.g. *personal use* or *recommendations* for specific purposes like gaming, programming, daily/school/business use etc.). For example, sentence (44) has been assigned the label MISCELLANEOUS**,** since it conveys a negative opinion towards the laptop as a gaming or media machine, while in (45) the reviewer expresses a positive opinion about the quality of the laptop (solid machine) recommending it at the same time for college students.

44. *This is not a serious gaming laptop or a serious media machine*. → {**LAPTOP#MISCELLANEOUS**, negative}
45. *I highly recommend this computer for students looking for a solid machine to get them through college*. → {**LAPTOP#MISCELLANEOUS**, positive}, **LAPTOP#QUALITY**, positive}

- Opinions referring to other types of advantages/disadvantages related to the target entities (e.g. the free printer in (46) or the absence of a hardcopy manual in (47), and to miscellaneous problems (e.g. 48), attributes (e.g. 49) and opinions in general (e.g. 50):

46. *And the best part is that it even comes with a free printer*→ {**LAPTOP#MISCELLANEOUS**, positive}
47. *The one thing I wish it had was a detailed hardcopy manual*. → {**LAPTOP#MISCELLANEOUS**, negative}
48. *MY ONLY PROBLEM IS I CAN NOT REG. THE PRODUCT KEY.* → {**LAPTOP#MISCELLANEOUS**, negative}
49. *I will NEVER buy (Refurbished) again, I don't care how cheap it is*. → {**LAPTOP#MISCELLANEOUS**, negative}
50. *Oh and if thats not bad enough it doesn't come with a recovery cd so you can make one if you know how to or buy one if you buy it the cost is $25 for two cds.*→ {**LAPTOP#MISCELLANEOUS**, negative}

Overall the 22 entities and 9 attribute labels give rise to more than 80 E#A pairs (combinations of entity and attribute labels) (see below section 2.3.4).

Opinions expressed towards entities not described above are considered to be **out of the scope** of the current annotation framework and the corresponding sentences should be tagged accordingly. Such entities are laptops or products of the same or a competitive brand, theirs components, or other companies (e.g. competitive brands or retailers like Amazon, Best Buy, MacConnection etc.). For example, in sentences (51,

52) the reviewers are expressing opinions about other products. Even though these opinions are related somehow to the reviewed entity, they are also considered to be out of scope. Furthermore, comparative opinions are out of the scope of this annotation task. These opinions could be expressed towards either specific entities that are explicitly (e.g. 54) or vaguely mentioned (e.g. 55) or classes of entities (e.g. 56) such as NetBooks, PCs, etc.

51. *I previously owned a Toshiba and it only lasted about 2 years.* → {**OutOfScope**} (Previous sentence: *The apple MacBook is the best investment that I have ever made)*
52. *I love my Samsung TV and Galaxy S smartphone, but this Netbook was a very poor computer*. → {**OutOfScope**}
53. *I would recommend anyone to buy from pcconnection express.* → {**OutOfScope**}
54. *Mac software is just so much simpler than Microsoft software.* → {**OutOfScope**}
55. *The Toshiba laptop I am using is easier to use than most I have tried.* → {**OutOfScope**}
56. *I wasn't a big fan of the Netbooks but this one was very well designed.* → {**OutOfScope**}

2.3.2.2 ASPECT CATEGORIES FOR THE RESTAURANTS DOMAIN

The entity E of an {E#A} pair can be assigned one of the following 6 labels:

- **FOOD**: for opinions focusing on the food in general or in terms of specific dishes, dining options etc. (e.g. 1, 2, 4, 6, 7, 8).

- **DRINKS**: for opinions focusing on the drinks in general or in terms of specific drinks, drinking options etc. (e.g. 3, 4, 5).

- **SERVICE**: for opinions focusing on the (customer/kitchen/counter) service, on the promptness and quality of the restaurant's service in general, the food preparation, the staff's attitude and professionalism, the wait time, the options offered (e.g. *takeout*), etc. (e.g. 7, 8).

- **AMBIENCE**: for opinions focusing on the atmosphere or the environment of the restaurant's interior or exterior space (e.g. terrace, yard, garden), the décor, entertainment options, etc. (e.g. 6, 7).

- **LOCATION**: for opinions focusing on the location of the reviewed restaurant in terms of its position, the surroundings, the view, etc. (e.g. 9).

- **RESTAURANT**: for opinions evaluating the restaurant as a whole and not focusing on any of the above five entity types (e.g. 9-13).

The attribute A of an {E#A} pair can be assigned one of the following 5 labels:

- **QUALITY**: for opinions focusing on the taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality of the FOOD and the DRINKS served in the restaurant e.g.

    1. *The spicy tuna roll was unusually good and the rock shrimp tempura was awesome, great appetizer to share!* →{**FOOD#QUALITY**, positive}
    2. *Food was okay, nothing great.* →{**FOOD#QUALITY**, neutral}
    3. *Always ask the bartender for the SEASONAL beer!!!* → {**DRINKS#QUALITY**, positive}

- **STYLE&OPTIONS**: for opinions referring to the presentation, the serving style, the portions size, the food/menu options or variety (e.g. innovative dishes/drinks, vegetarian options) of the FOOD and of the DRINKS served in the restaurant e.g.

    4. *The portions are small but being that the food was so good makes up for that*. → {**FOOD#STYLE&OPTIONS**, negative}, {**FOOD#QUALITY**, positive}

- **PRICES**: for opinions that refer to the prices of the FOOD, the DRINKS or the RESTAURANT in general.

    5. *The wine list is interesting and has many good values*. → {**DRINKS#STYLE&OPTIONS**, positive}, {**DRINKS#PRICES**, positive}

- **GENERAL**: this attribute label is assigned to sentences that express general positive or negative sentiment about the RESTAURANT as well as to the sentences that express opinions about the SERVICE, the AMBIENCE and the LOCATION, since no fine-grained attributes have been defined in the current annotation schema for these entity types e.g.

6. *The food was very good, a great deal, and the place its self was great.* → {**FOOD#QUALITY**, positive}, {**FOOD#PRICES**, positive}, {**AMBIENCE#GENERAL**, positive}
7. *Excellent atmosphere, delicious dishes good and friendly service.* → {**AMBIENCE#GENERAL**, positive}, {**FOOD#QUALITY**, positive}, {**SERVICE#GENERAL**, positive}
8. *Bagels are ok, but be sure not to make any special requests!* →{**FOOD#QUALITY**, neutral}, {**SERVICE#GENERAL**, negative}
9. *Its location is good and the fact that Hutner College is near and their prices are very reasonable, makes students go back to Suan again and again.* → {**LOCATION#GENERAL**, positive}, {**RESTAURANT#PRICES**, positive}
10. *Rao is a good restaurant, but it's nothing special.* →{**RESTAURANT#GENERAL**, neutral}
11. *Go there once and oh yes...you will go back...you will...* →{**RESTAURANT#GENERAL**, positive}

- **MISCELLANEOUS**: for attributes that do not fall into any of the aforementioned cases e.g.

12. *Not a great place for family or general dining.* → {**RESTAURANT#MISCELLANEOUS**, negative}
13. *Good luck getting a table.* → {**RESTAURANT#MISCELLANEOUS**, negative}

Overall, the 6 entity types and 5 attribute classes described above can result in 12 possible combinations (E#A pairs) as illustrated in the following table:

	GENERAL	PRICES	QUALITY	STYLE&OPTIONS	MISCELLANEOUS
RESTAURANT	✓	✓	✗	✗	✓
FOOD	✗	✓	✓	✓	✗
DRINKS	✗	✓	✓	✓	✗
AMBIENCE	✓	✗	✗	✗	✗
SERVICE	✓	✗	✗	✗	✗
LOCATION	✓	✗	✗	✗	✗

Table 5: Possible E#A pairs in the Restaurants domain

Opinions expressed towards entities that are not described above (e.g. other restaurants that the reviewer has visited) as well as comparative opinions are considered to be out of the scope of the current annotation framework, and the corresponding sentences should be tagged accordingly e.g.

14. *I was in love with Pongsri on 48th, but compared to Suan it is slow in service and overpriced".* → {**OutOfScope**}
15. *The service was attentive, yet unimposing, the food was far better than many notorious restaurants in Midtown and the wine list is extensive and well priced.* → {**OutOfScope**}

2.3.2.3 ASPECT CATEGORIES FOR THE HOTELS DOMAIN

The entity E of an {E#A} pair can be assigned one of the following 7 labels:

- **HOTEL** for opinions evaluating the hotel as whole or in terms of the lack or presence of extra features/facilities (e.g. 1, 5, 9-11, 14, 15).

- **ROOMS** for opinions evaluating the rooms in terms of their size, general condition, view, furniture, bathroom, sleep quality and the lack or presence of extra features/amenities (e.g. 3, 4).

- **ROOM_AMENITIES** for opinions evaluating the rooms in terms of the amenities they include (e.g. *air condition, refrigerator, microwave, mini bar, hair dryer, TV, toiletries, safe, balcony, coffee maker, linen*) (e.g. 6).

- **FACILITIES** for opinions focusing on the hotel facilities in terms of specific installations/areas (e.g. *swimming pool, spa&sauna, beauty salon, restaurants, café, night club, casino, business center, gymnasium, access facility for the differently-abled, parking, etc*.) or guest services offered by a hotel (e.g. *shuttle, laundry, baby sitting or wake up services, sports activities, 24-hour concierge &front desk, information desk, in-room dining, internet access, availability of touristic material etc*.) (e.g. 2).

- **SERVICE** for opinions focusing on the staff's attitude and promptness, easiness to problem solving, execution of service in time, or the rooms/ check-in/ check-out/ reception etc. service, etc. (e.g. 13).

- **LOCATION** for opinions focusing on the location of the reviewed hotel in terms of its position, the surroundings, the view, etc. (e.g. 12).

- **FOOD&DRINKS** for opinions focusing on the breakfast, the food and the drinks in general or in terms of specific dishes and drinks, dining/drinking options etc. (e.g. 7, 8).

The attribute A of an {E#A} pair can be assigned one of the following 8 labels. In the examples below the respective polarity label is also provided.

- **PRICES** for opinions that refer to the prices of the rooms, the food & drinks, the facilities/services offered by the hotel or the hotel in general e.g.

1. *Pleasantly surprised at $69 night.* → {**HOTEL#PRICES**, positive}
2. *The only downside was a per minute charge to use the business center computers.* → {**FACILITIES#PRICES**, negative}

- **DESIGN&FEATURES** for opinions that refer to the design, the appearance/decor, the size of an entity (hotel, rooms, facilities), to extra or missing features (e.g. amenities/facilities), etc. e.g.

3. *The bathroom was small and all white, and lacked a soap dish in the shower and no grab bars with a rather tricky exit required out of the shower/tub.* → {**ROOMS#DESIGN_FEATURES**, negative}

- **CLEANLINESS** for opinions that refer to the neatness or hygiene of the rooms, common areas and the hotel in general e.g.

4. *The room was spacious and clean.* → {**ROOMS#DESIGN_FEATURES**, positive}, {**ROOMS#CLEANLINESS**, positive}

- **COMFORT** for opinions evaluating an entity in terms of its comfortableness or convenience for the guests, (e.g. stay and sleep quality, accessibility). e.g.

5. *The building appears to be on permanent lock-down, as the only way in is through the front door, away from the main parking area.* → {**HOTEL#COMFORT**, negative}

- **QUALITY** for opinions focusing on the quality of the FOOD&DRINKS (e.g. taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality of the food and the drinks served in the hotel) or the quality of the hotel facilities and room amenities e.g.

6. *The towels were thin and worn.* → {**ROOM_AMENITIES#QUALITY**, negative}
7. *The breakfast is excellent*! → {**FOOD&DRINKS#QUALITY**, positive}

- **STYLE&OPTIONS** for opinions referring to the food/drinks presentation, the serving style, the portions size, the food/menu options or variety (e.g. innovative dishes/drinks, vegetarian options) of the food and of the drinks served in the restaurant e.g.

8. *It is rare that hotels in this class serve hot meals, yet they do!* → {**FOOD&DRINKS#STYLE&OPTIONS**, positive}

- **GENERAL**. This attribute label is assigned to sentences that express general positive or negative sentiment about an entity type (hotel, room amenities, rooms, facilities, location, service) e.g.

9. *Not bad for one night.*→ {**HOTEL#GENERAL**, neutral}
10. *It's not a recipe for another stay."* → {**HOTEL#GENERAL**, negative}
11. *An elevator was broken during our last stay and it was most annoying, but did not greatly impact the overall experience*.→ {**FACILITIES#QUALITY**, negative}, {**HOTEL#GENERAL**, positive}
12. *Close to the airport and restaurants*. → {**LOCATION#GENERAL**, positive}
13. *Front desk staff were very friendly and helpful; made us feel very welcome to their property*." → {**SERVICE#GENERAL**, positive}

- **MISCELLANEOUS** for attributes that do not fall into any of the aforementioned cases (e.g. recommendations for specific purposes). e.g.

14. *If you plan to do any hiking, this is a perfect place to stay*." → {**HOTEL#MISCELLANEOUS**, positive}
15. *There was construction being done on the street in front of the hotel; which made it very difficult driving around.* → {**HOTEL#MISCELLANEOUS**, negative}

Overall, the 7 entity types and 8 attribute classes described above can result in 34 possible combinations (E#A pairs) as illustrated in the following table:

	GENERAL	PRICES	DESIGN& FEATURES	CLEANLINESS	COMFORT	QUALITY	STYLE&OPTIONS	MISCELLANEOUS
HOTEL	✓	✓	✓	✓	✓	✓	x	✓
ROOMS	✓	✓	✓	✓	✓	✓	x	✓
ROOM_AMENITIES	✓	✓	✓	✓	✓	✓	x	✓
FACILITIES	✓	✓	✓	✓	✓	✓	x	✓
SERVICE	✓	x	x	x	x	x	x	x
LOCATION	✓	x	x	x	x	x	x	x
FOOD&DRINKS	x	✓	x	x	x	✓	✓	✓

Table 6: Possible E#A pairs in the Hotels domain

2.3.3 ANNOTATION PROCESS

A collection of 900 customer reviews texts (400 reviews for restaurants and 500 for laptops that is more complicated domain) were collected and annotated from scratch following the annotation framework proposed in the previous section. In addition, a small sample of 50 hotel review texts was collected and annotated as a pilot study for new domain modelling according to the new annotation framework. The annotation process was similar with the one described in Section 2.2.3); two annotators, the author of this thesis (annotator A) and a graduate student in computational linguistics (annotator B) worked collaboratively using BRAT (Stenetorp et al., 2012) as an annotation tool. When A and B disagreed, a decision was made collaboratively by them and a third annotator (a computer scientist). However, this time the input for the annotators was full review texts; the annotation was still at the sentence level, but it took into account the broader context.

The main difficulties and disagreements encountered at the annotation process are summarized below. In the laptops domain the main difficulty was that in some negative evaluations the annotators were unsure about the actual problem/target. For example, in "*Sometimes the screen even goes black on this computer*", the black screen may be related to the graphics, the laptop operation (e.g., motherboard issue) or the screen itself. The decision for such cases was to assign the E#A pair that reflected what the reviewer said and not the possible interpretations that a technician would give. So, if someone reports screen issues without providing further details,

then the opinion is considered to be about the screen[26]. Another issue occured when an attribute could be inferred from an explicitly evaluated attribute. For example, DESIGN affects USABILITY (e.g. *"With the switch being at the top you need to memorize the key combination rather than just flicking a switch"*). In such cases annotators assigned both attribute labels. The annotation in the restaurants domain was easier, due to the less fine-grained schema. A common problem was that (as in SE-ABSA14) the distinction between the GENERAL and MISCELLANEOUS and between the RESTAURANT and AMBIENCE labels was not always clear. In the case of the OTEs, the annotators found it easier to identify explicit references to the target entities as opposed to the more general aspect terms of the previous annotation framework (SE-ABSA14). However, the problem of distinguishing aspect terms when they appear in conjunctions or disjunctions remains. In this case the maximal phrase (e.g. the entire conjunction or disjunction) was annotated (e.g. *"Greek or Cypriot dishes"* instead of *"Greek dishes*", "*Cypriot dishes*"). As for sentiment polarity classification, the majority of cases for which the annotators could not easily decide about the correct polarity label can be classified in the following categories:

- *Change of sentiment over time*. Some reviewers start their review by saying how excited they were at first (e.g. with the laptop) and continue by reporting problems or negative evaluations.
- *Negative fact vs. positive opinion*. Some reviewers do mention particular deficiencies of a laptop or a restaurant saying, however, at the same time that they do not bother (e.g. *"Overheats but put a pillow and problem solved!"*).
- *Mildly positive and negative sentiments are both denoted by the "neutral" label*. In some cases the annotators reported that it would be helpful to have a more fine-grained schema (e.g. "negative", "somewhat negative", "neutral", "somewhat positive", "positive").

Finally, in some cases it is difficult to decide a polarity label without knowing the reviewer's intention (e.g. *"50% of the food was very good"*).

Again, the annotations -as well as the annotation guidelines- were finalized after several iterations in order to ensure consensus between the annotators and consistent annotations. When the annotations were finalized, the datasets were further refined by removing some reviews (very similar texts, problematic cases that were left for future research). The annotation process resulted in 5761 opinion tuples in total that

[26] "Blue screen" is an exception since it is well-known that it refers to the laptop operation.

correspond to more than 15000 label assignments (E, A, OTE, polarity); consult Table 2 for more information.

2.3.4 THE ABSA-2015 BENCHMARK DATASETS

The numbers of the review texts, sentences and ABSA tuple annotations per domain are provided below in Table 7:

Dataset	Texts	Sentences	ABSA tuples
LPT	450	2500	2923
RST	350	2000	2499
HTL	30	266	339
Total	830	4766	5761

Table 7: Sizes of the datasets and number of ABSA tuples per domain

As already mentioned in previous sections, the restaurants and hotels datasets contain {E#A, OTE, polarity} annotations, while the laptops only {E, A, polarity} tuples, since LAPTOP is the most frequent entity label; as illustrated below (Fig. 9), 1922 out of the 2923 ABSA tuples in this dataset refer to attributes of the entity LAPTOP, while the second most frequent entity class is the customer SUPPORT. As illustrated in Fig. 10, general evaluations about the laptop are the majority category (33%) in the specific customer reviews dataset, while the most frequent attributes that are evaluated are the OPERATION_PERFORMANCE, the DESIGN&FEATURES, and the QUALITY of a LAPTOP.

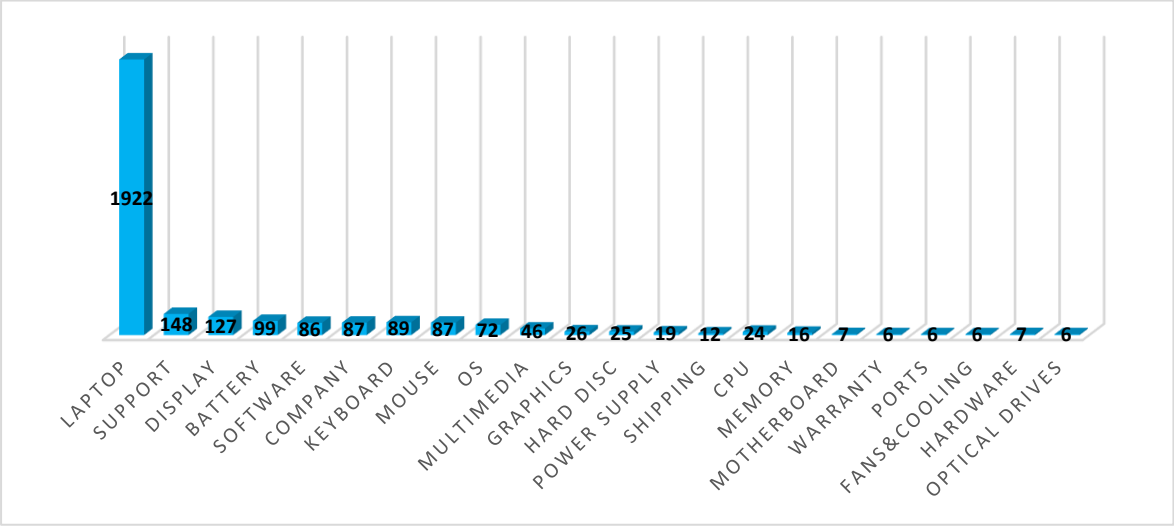


Figure 9: Number of annotations per entity in the laptops domain

Overall, the annotation process in the specific dataset resulted in 83 possible combinations (aspect categories) between entity and attribute types that reflect the complexity of this domain.

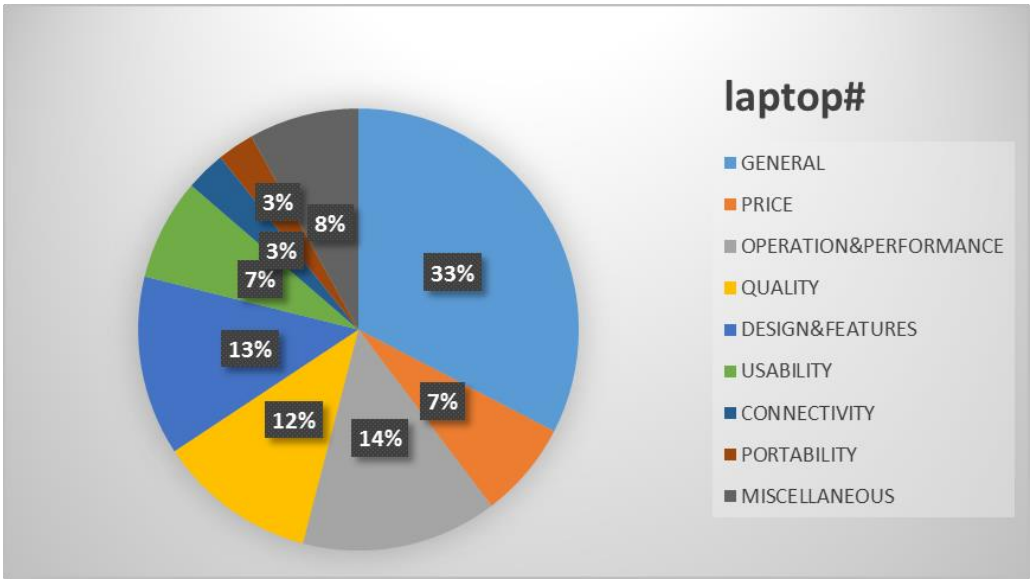


Figure 10: Distribution of aspect category annotations for the entity laptop

The most frequent entities in the restaurants domain is the FOOD (1071 annotations), the RESTAURANT (599) as a whole and the SERVICE (443). As expected, GENERAL (1148) is the majority class, since some entities (SERVICE, AMBIENCE, LOCATION) can only be assigned this label; the second most frequent attribute label is QUALITY (898) that applies for FOOD and DRINKS. The overall distribution of the annotated E#A pairs is illustrated below in Fig. 11:

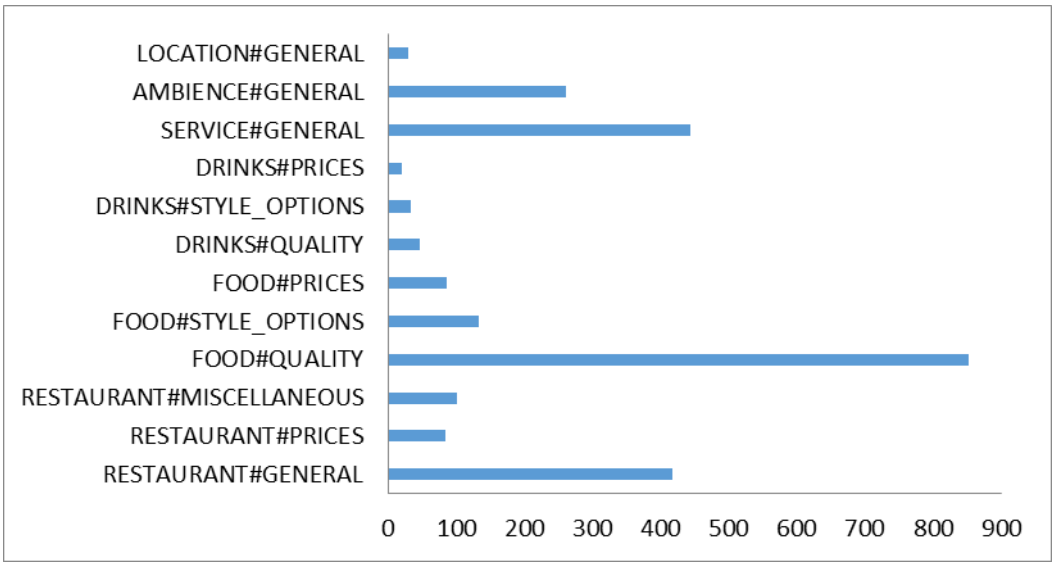


Figure 11: Distribution of aspect categories (E#A) annotations in the restaurants domain

The annotation process in the hotels domain -even though the data sample is rather limited (30 review texts)- resulted in 28 combinations of entity and attribute labels (E#A pairs); this may indicate that hotel customers discuss a variety of aspects in their reviews. Fig. 12 presents the 15 most frequent E#A pairs.

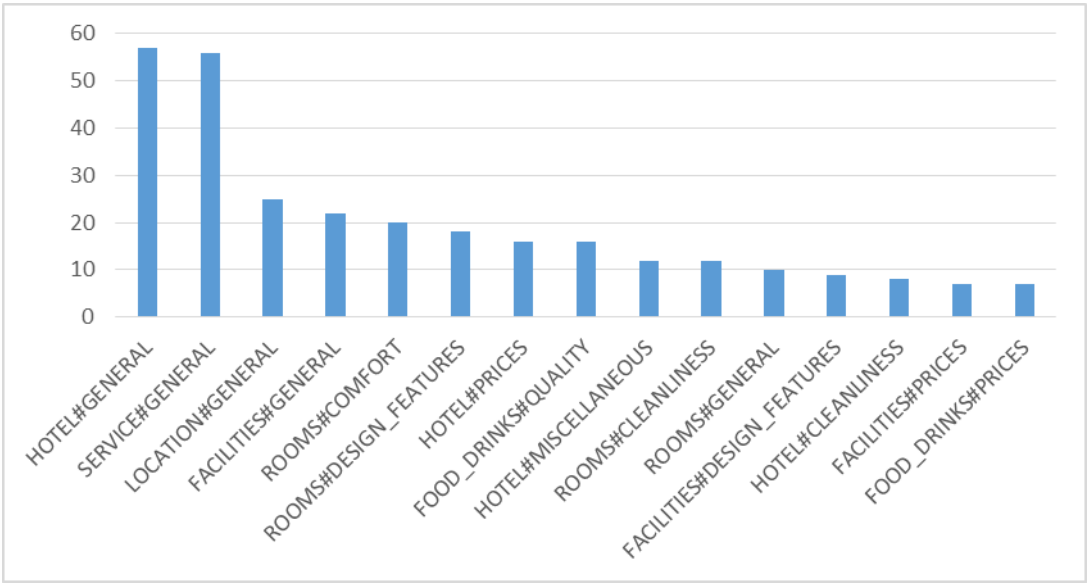


Figure 12: Distribution of aspect categories (E#A) annotations in the hotels domain

As for the sentiment polarity annotations, as illustrated below in Fig. 13, "positive" is the majority class in all domains. In addition, "positive" is significantly most frequent in the restaurants and the hotels domain as compared to the laptops one.

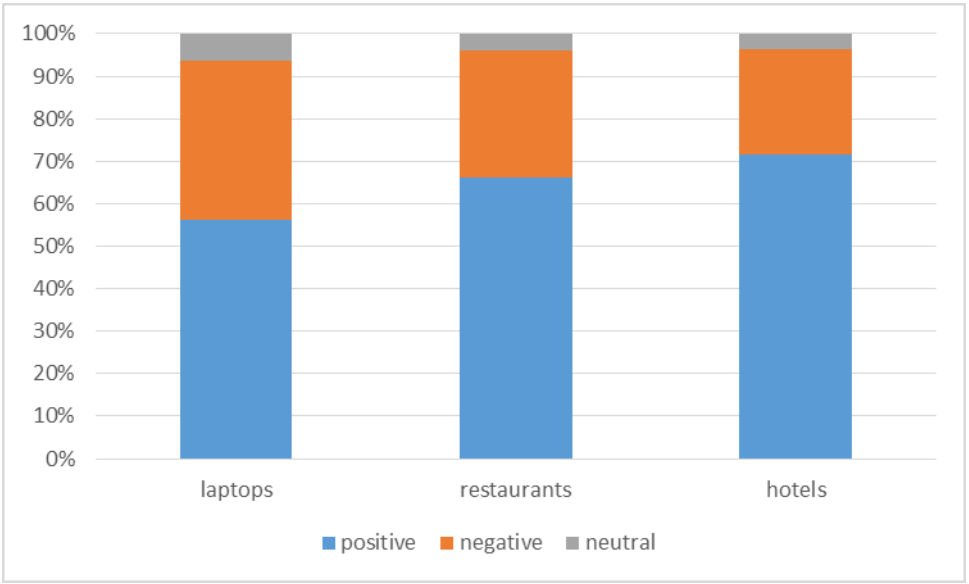


Figure 13: Distribution of polarity annotations per domain

The proposed datasets, along with the annotation framework and the guidelines were adopted from the SE-ABSA15 shared task that was organized in the context of the SemEval 2015 workshop as a follow up of the SE-ABSA14 task by the same organizing committee (section 2.3.5). Based on the experience of the annotation process and the ABSA2014 experience, the expectations were that systems would perform better in the restaurants domain, since in this domain the classification schema is less fine-grained (it contains 6 entity types and 5 attribute classes that result in 12 possible combinations, as opposed to the laptops domain where the 22 entities and 9 attribute labels give rise to more than 80 combinations), and that OTE would be an easier task as compared to aspect category detection.

2.3.5 THE ABSA-2015 SEMEVAL CHALLENGE

The SE-ABSA15 task[27] built upon the SE-ABSA14 task and - following the new ABSA framework proposed in this thesis (section 2.3.1)- consolidated its subtasks (aspect category extraction, aspect term extraction, polarity classification) into a principled unified framework in which all the identified constituents of the expressed opinions (i.e. aspects, opinion target expressions and sentiment polarities) meet a set of guidelines and are linked to each other within sentence-level tuples. In addition, SE-ABSA15 included an aspect level polarity classification subtask for the hotels domain in which no training data were provided (out-of-domain ABSA). In particular, the task consisted of the following subtasks and slots:

- **Subtask 1: In-domain ABSA.** Given a review text about a laptop or restaurant, identify all the opinion tuples with the following types (tuple slots) of information:

  - **Slot 1**: **Aspect Category Detection** (E#A pairs).
  - **Slot 2: Opinion Target Expression** (**OTE**)[28].
    - **Slots 1&2**: Link the extracted OTEs to the respective Aspect Categories.
  - **Slot 3: Sentiment Polarity Classification**.

- **Subtask 2: Out-of-domain ABSA** (Hotels Reviews). The gold annotations for Slots 1 and 2 were provided and the teams had to return the sentiment polarity values (Slot 3).

Participants were free to choose the subtasks, slots and domains they wished to participate in. The task provided training and testing data on both domains

[27] http://alt.qcri.org/semeval2015/task12/
[28] This slot was required only in the restaurants domain.

(restaurants and laptops) for the first subtask (SB1, SB2), and only testing data on the hotels domain for Subtask 2. In particular, the datasets described in the previous section were split for training and testing purposes as illustrated below in Fig. 14 and 15. The task provided also baselines for each slot and domain. The distribution of the category annotations in the restaurants domain (Fig. 14) is similar across the train and test set.

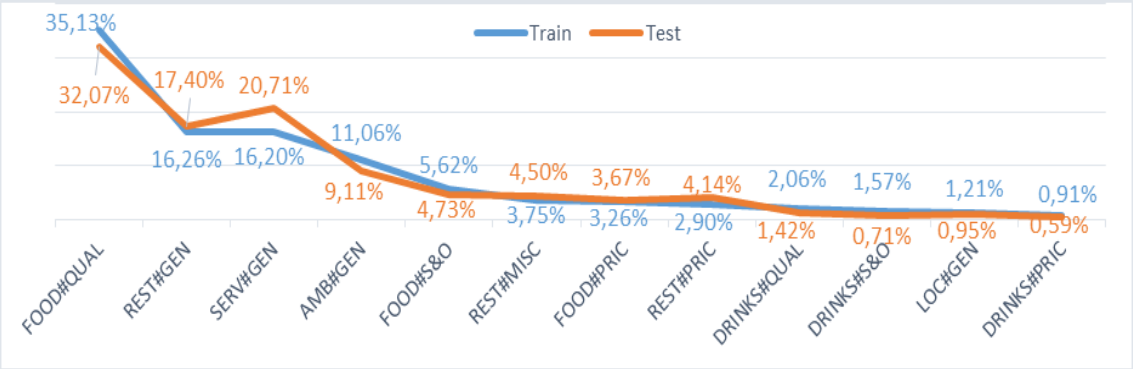


Figure 14: Aspect category (E#A) distribution in the restaurants domain. REST = restaurant, SERV = service, AMB = ambience, LOC = location, GEN=general, PRIC = price, S&O = style&options, MISC= miscellaneous

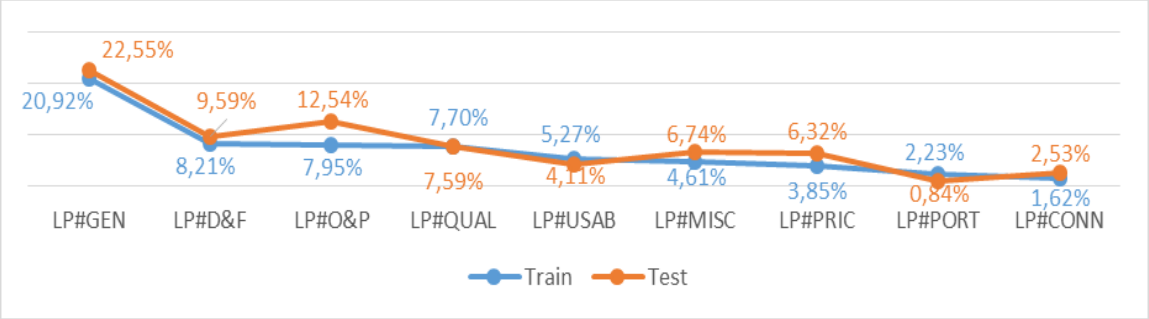


Figure 15: LAPTOP#ATTRIBUTE categories distribution in the laptops domain. LP= laptop, O&P= operation&performance, QUAL= quality, D&F= design &features, USAB=usability, CONN=connectivity, PORT=portability

Fig. 15 presents the distribution for all the attributes of the LAPTOP entity in the train and test sets. Again, the category distributions are similar. The training set contains 81 E,A combinations (different pairs), while the test set 58. LAPTOP is the majority entity class in both sets; 62.36% in train, 72.81% in test data. The remaining 37.64% of the annotations in the laptops train data correspond to 72 categories with frequencies ranging from 6.53% to 0.05%. In the test set the remaining 27.19% of the annotations correspond to 49 categories with frequencies ranging from 2.32 % to 0.11%.

Regarding the polarity, "positive" is the majority class in all domains (Fig. 16). The polarity distribution is balanced in the laptops domain, while in the restaurants domain there is a significant imbalance between the "positive" and "negative" classes across the training and the test sets.

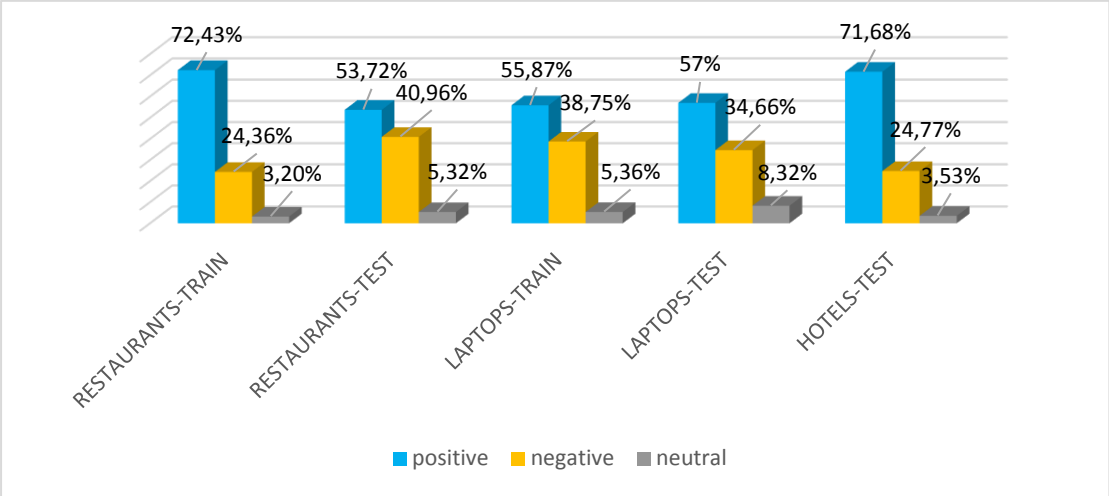


Figure 16: Polarity distribution per domain (RS-restaurants, LP-laptops, HT-hotels). TR and TE indicate the training and test sets.

The task attracted 92 submissions from 16 teams. Similarly to SE-ABSA14, the evaluation ran in two phases. In Phase A, participants were asked to return the {category, OTE} tuples for the restaurants domain and only the category slot (Slot1) for the laptops domain. Subsequently, in Phase B, participants were given the gold annotations for the reviews of Phase A and they were asked to return the polarity (Slot3). Aspect category detection (Slot1) and OTE extraction (Slot2) were evaluated using the F1 measure, while sentiment polarity classification (Slot 3) was evaluated using accuracy. The participating teams experimented with a variety of features, techniques, and resources as summarized in Fig. 17 below.

As expected, the systems achieved significantly higher scores (+12%) in the restaurants domain, since in this domain the classification schema is less fine-grained and complex. The best F-1 scores in both domains, 50.86% for laptops and 62.68% for restaurants, were achieved by NLANGP team (Toh and Su, 2015), which modeled aspect category detection as a multiclass classification problem with features based on tokenization, parsing as well as on word clusters learnt from Amazon and Yelp data, for laptops and restaurants respectively. In the OTE slot, the best F1 score (70.05%) was achieved by EliXa team (Vicente, Saralegi and Agerri, 2015) that addressed the problem using an averaged perceptron with a BIO tagging scheme. The features they used included n-grams, token classes, token prefixes and suffixes, and word clusters

learnt from additional data; Yelp for Brown and Clark clusters and Wikipedia for word2vec clusters. Similarly, NLANGP (67.11%) was based on a CRF model with features based on word strings, head words, name lists, and Brown clusters. Also, candidate terms are extracted using the double propagation algorithm (Qiu et al., 2011). Finally, as expected, the scores are significantly lower when systems have also to link the extracted OTEs to the relevant aspect categories (Slot1&2). The best F-1 score (42.90%) was achieved by the NLANGP team that simply combined the output for each slot to construct the corresponding tuples.

In Slot 3, the best accuracy scores in both domains (79.34% for laptops and 78.69% for restaurants) were achieved by Sentiue (Saias, 2015) using features based on n-grams, POS tagging, lemmatization, and publicly available sentiment lexica (MPQA, Bing Liu's lexicon, AFINN). Most teams performed (slightly) better in the laptops domain. This is probably due to the fact that in the restaurants domain the positive polarity is significantly more frequent in the training than in the test data, which may have led to biased models. As for the hidden domain, the results of some systems suggest that it was easier, but other systems achieved significantly lower scores, compared to the in-domain ABSA scores. More details about the baselines, the submitted systems and the evaluation results are available at the task overview paper (Pontiki et al., 2015).

Figure 17: Summary of features, techniques, and resources used in SE-ABSA15. Font size indicates frequency.

2.3.6 CONCLUDING REMARKS AND NEXT STEPS

This section presented the new ABSA framework that is proposed in this thesis and aims to address problems encountered during the annotation process following state-of-the-art definitions and classification schemes. According to the new definition, an aspect is a dyadic relation constructed by the entity or the part of an entity that is evaluated and the attribute with regard to which it is evaluated. The new definition yielded a new representation framework in which all the identified constituents of the expressed opinions (i.e. opinion target expressions, aspects and sentiment polarities) meet a set of guidelines/specifications, and are linked to each other within sentence-level tuples that directly reflect the intended meaning of the texts. Furthermore, a set of aspect inventories for three domains and a detailed data-driven annotation codebook were compiled enabling us to apply this new framework to new datasets that consist of whole reviews, not isolated sentences. The proposed framework and the datasets were adopted for the set-up and the support of the SE-ABSA15 shared task that was organized as a follow up of the SE-ABSA14.

The performances of the systems submitted in the context the SemEval ABSA shared task in both years (2014 and 2015) following different ABSA representations indicate that what was easier and more straightforward for human annotators was significantly more difficult for systems; for example, as expected, there was a significant drop in the systems performance in the restaurant domain when they had to detect 12 dyadic classes (E#A pairs) as compared to 5 single classes (best scores in aspect category detection: 88.57% and 62.68% in 2014 and 2015, respectively). Similarly, identifying opinion targets rather than aspect terms in general was harder (best scores in aspect term extraction (2014) and opinion target expression detection (2015): 84.01% and 70.05%, respectively). However, this new framework is more structured, meaningful and representative of the intended meaning of the texts. The third step and final step of the work presented in this thesis was to extend this framework towards two directions; text-level annotations/opinion summaries, and other languages and domains (see below section 2.4).

2.4 EXTENDING ABSA

This section presents the third and last part of this thesis work in the context of ABSA, namely the extension of the new ABSA framework presented in the previous section towards two directions; create text-level annotations that can be used for the generation of opinion summaries (2.4.1), and apply the proposed annotation framework to other languages and domains in the context of the SE-ABSA16 task as a follow up of the SE-ABSA15 (2.4.2).

2.4.1 TEXT LEVEL ANNOTATIONS

As already discussed in the previous section, the sentence-level ABSA tuples directly reflect the intended meaning of the texts. These tuples are important since they indicate the part of the text within which a specific opinion is expressed. However, a user might also be interested in the overall rating of the text with regard to a particular aspect. Such ratings can be used to estimate the mean sentiment per aspect from multiple reviews (McAuley, Leskovec and Jurafsky, 2012). Hence, moving towards a more application and end-user oriented ABSA framework, this section presents how the sentence-level annotations (SLA) can be aggregated at the text level (Text Level Annotations-TLA) towards a more meaningful and structured output.

2.4.1.1 FROM SLA TO TLA

The inspection of the existing sentence-level annotations in the ABSA-2015 benchmark datasets (described above in 2.3.4) from an opinion summarization standpoint revealed that each review text may fall into one of the following cases:

1) Each aspect (i.e. E#A pair) of the target entity is discussed in the text only once; there is a unique annotated tuple for each aspect category at the sentence level (Fig. 18). In such cases the sum of all the existing SLA is considered to directly reflect and summarize the intended meaning of the text (Fig. 19).

```
<Review rid="252">
    <sentences>
        <sentence id="252:0">
            <text>So far, a great product.</text>
            <Opinions>
                <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
            </Opinions>
        </sentence>
        <sentence id="252:1">
            <text>High price tag, however.</text>
            <Opinions>
                <Opinion category="LAPTOP#PRICE" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="252:2">
            <text>Still trying to learn how to use it.</text>
           <Opinions>
                <Opinion category="LAPTOP#USABILITY" polarity="neutral"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 18: SLA for Review id "252" (laptops domain)

```
    <Opinions>
        <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
        <Opinion category="LAPTOP#PRICE" polarity="negative"/>
        <Opinion category="LAPTOP#USABILITY" polarity="neutral"/>
    </Opinions>
```

Figure 19: TLA for Review id "252" (laptops domain)

However, in some cases the sentiment polarity that has been assigned to a particular aspect (usually the target entity in general) at the sentence level may not be valid at the text level. For example, in the following review (Fig. 20) the positive label for the "LAPTOP#GENERAL" category has to change to "negative" for the TLA (Fig. 21).

```
<Review rid="139">
    <sentences>
        <sentence id="139:0">
            <text>HP Pavilion DV9000 Notebook PC      When I first got this computer, it really rocked.</text>
            <Opinions>
                <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
            </Opinions>
        </sentence>
        <sentence id="139:1">
            <text>But as time went on I found it almost impossible to keep the thing on-line through wi-fi.</text>
            <Opinions>
                <Opinion category="LAPTOP#CONNECTIVITY" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="139:2">
            <text>Eventually the screen went blank and the computer would not turn on.</text>
            <Opinions>
                <Opinion category="DISPLAY#OPERATION_PERFORMANCE" polarity="negative"/>
                <Opinion category="LAPTOP#OPERATION_PERFORMANCE" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="139:3">
            <text>HP said it was out of warranty.</text>
        </sentence>
        <sentence id="139:4">
            <text>Guess I'll stay away from HP.</text>
            <Opinions>
                <Opinion category="COMPANY#GENERAL" polarity="negative"/>
            </Opinions>
        </sentence>
    </sentences>
```

Figure 20: SLA for Review id "139" (laptops domain)

```
<Opinions>
    <Opinion category="LAPTOP#GENERAL" polarity="negative"/>
    <Opinion category="LAPTOP#CONNECTIVITY" polarity="negative"/>
    <Opinion category="DISPLAY#OPERATION_PERFORMANCE" polarity="negative"/>
    <Opinion category="LAPTOP#OPERATION_PERFORMANCE" polarity="negative"/>
    <Opinion category="COMPANY#GENERAL" polarity="negative"/>
</Opinions>
```

Figure 21: TLA for Review id "139" (laptops domain)

2) An aspect of the target entity may be discussed in several parts of the text; there are more than one annotated tuples for each or some aspect categories at the sentence level, but all the tuples of a particular aspect have the **same polarity** label. For example, in the following review (Fig. 22) the reviewer expresses negative opinion towards the laptop in general and its quality twice. Given that there is a unique sentiment polarity label per aspect category irrespectively of how many times each category is discussed in the text, again the sum of all the existing SLA is considered to directly reflect and summarize the intended meaning of the text if the duplicate tuples are removed (Fig. 23).

```
<Review rid="19">
    <sentences>
        <sentence id="19:0">
            <text>it is the worst computer dell ever made.</text>
            <Opinions>
                <Opinion category="LAPTOP#GENERAL" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="19:1">
            <text>it is hard to fix and makes it a hassle to own one.</text>
            <Opinions>
                <Opinion category="LAPTOP#QUALITY" polarity="negative"/>
                <Opinion category="LAPTOP#GENERAL" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="19:2">
            <text>breaks easily.</text>
            <Opinions>
                <Opinion category="LAPTOP#QUALITY" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="19:3">
            <text>only good thing is the graphics quality.</text>
            <Opinions>
                <Opinion category="GRAPHICS#QUALITY" polarity="positive"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 22: SLA for Review id "19" (laptops domain)

```
<Opinions>
    <Opinion category="LAPTOP#GENERAL" polarity="negative"/>
    <Opinion category="LAPTOP#QUALITY" polarity="negative"/>
    <Opinion category="GRAPHICS#QUALITY" polarity="positive"/>
</Opinions>
```

Figure 23: TLA for Review id "19" (laptops domain)

3) An aspect of the target entity may be discussed in several parts of the text with different sentiment polarity (i.e. conflicting opinions); there are more than one annotated tuples for some aspect categories at the sentence level, where one or more tuples for a particular aspect have **different polarity** labels. In this case the **dominant sentiment** for the particular aspect has to be chosen. For example, in the following text (Fig. 24) the dominant sentiment about the quality of the food is considered to be "negative" and not "neutral" (Fig. 25):

```
<Review rid="1016296">
    <sentences>
        <sentence id="1016296:0">
            <text>I was very disappointed with this restaurant.</text>
            <Opinions>
                <Opinion target="restaurant" category="RESTAURANT#GENERAL" polarity="negative" from="34" to="44"/>
            </Opinions>
        </sentence>
        <sentence id="1016296:1">
            <text>Ive asked a cart attendant for a lotus leaf wrapped rice and she replied back rice and just
            walked away.</text>
            <Opinions>
                <Opinion target="cart attendant" category="SERVICE#GENERAL" polarity="negative" from="12" to="26"/>
            </Opinions>
        </sentence>
        <sentence id="1016296:2">
            <text>I had to ask her three times before she finally came back with the dish Ive requested.</text>
            <Opinions>
                <Opinion target="NULL" category="SERVICE#GENERAL" polarity="negative" from="0" to="0"/>
            </Opinions>
        </sentence>
        <sentence id="1016296:3">
            <text>Food was okay, nothing great.</text>
            <Opinions>
                <Opinion target="Food" category="FOOD#QUALITY" polarity="neutral" from="0" to="4"/>
            </Opinions>
        </sentence>
        <sentence id="1016296:4">
            <text>Chow fun was dry; pork shu mai was more than usually greasy and had to share a table with loud
            and rude family. </text>
            <Opinions>
                <Opinion target="Chow fun" category="FOOD#QUALITY" polarity="negative" from="0" to="8"/>
                <Opinion target="pork shu mai" category="FOOD#QUALITY" polarity="negative" from="18" to="30"/>
                <Opinion target="NULL" category="AMBIENCE#GENERAL" polarity="negative" from="0" to="0"/>
            </Opinions>
        </sentence>
        <sentence id="1016296:5">
            <text>I/we will never go back to this place again.</text>
            <Opinions>
                <Opinion target="place" category="RESTAURANT#GENERAL" polarity="negative" from="32" to="37"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 24: SLA for Review id "1016296" (restaurants domain)

```
<Opinions>
    <Opinion category="RESTAURANT#GENERAL" polarity="negative"/>
    <Opinion category="SERVICE#GENERAL" polarity="negative"/>
    <Opinion category="FOOD#QUALITY" polarity="negative"/>
    <Opinion category="AMBIENCE#GENERAL" polarity="negative"/>
</Opinions>
```

Figure 25: TLA for Review id "1016296" (restaurants domain)

In some cases the dominant sentiment has to be decided between a positive and a negative opinion towards the same aspect. For example, in the following text (Fig. 26) the review starts with a positive opinion about the laptop in general and a negative opinion about the customer support quality and continues with negative opinions about the support to end up with a negative recommendation about the product itself

and the company. So, in this case negative is considered to be the dominant sentiment for the "LAPTOP#GENERAL" aspect.

```
<Review rid="134">
    <sentences>
        <sentence id="134:0">
            <text>I was happy with My purchase of a Toshiba Satellite L305D-S5934 laptop until it came time to have
            it repaired under the Toshiba Warranty.</text>
            <Opinions>
                <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
                <Opinion category="SUPPORT#GENERAL" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="134:1">
            <text>The computer was shipped to their repair depot  on june 24 and returned on July 2 seems like a
            short turn around time except the computer was not repaired when it was returned.</text>
            <Opinions>
                <Opinion category="SHIPPING#QUALITY" polarity="positive"/>
                <Opinion category="SUPPORT#QUALITY" polarity="negative"/>
            </Opinions>
        </sentence>
        <sentence id="134:2">
            <text>....</text>
        </sentence>
        <sentence id="...">
        </sentence>
        <sentence id="134:11">
            <text>I would not recommend the purchase of this modle of Toshiba Computer or any Toshiba product for
            that matter.</text>
            <Opinions>
                <Opinion category="LAPTOP#GENERAL" polarity="negative"/>
                <Opinion category="COMPANY#GENERAL" polarity="negative"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 26: SLA for Review id "134" (laptops domain)

In other cases a positive opinion may dominate over a negative one. For example, in the following review (Fig. 27) the positive opinion about the quality of the food is considered to dominate over the negative opinion about the "frites".

```
<Review rid="1730014">
    <sentences>
        <sentence id="1730014:0">
            <text>I went here with a friend on a whim, we went someplace else first and couldn't get a table.</text>
        </sentence>
        <sentence id="1730014:1">
            <text>Service here was great, food was fantastic.</text>
            <Opinions>
                <Opinion target="Service" category="SERVICE#GENERAL" polarity="positive" from="0" to="7"/>
                <Opinion target="food" category="FOOD#QUALITY" polarity="positive" from="24" to="28"/>
            </Opinions>
        </sentence>
        <sentence id="1730014:2">
            <text>Guacamole+shrimp appetizer was really great, we both had the filet, very good, didn't much like the
            frites that came with, but the filet was so good, neither of us cared.</text>
            <Opinions>
                <Opinion target="Guacamole+shrimp appetizer" category="FOOD#QUALITY" polarity="positive" from="0" to="26"/>
                <Opinion target="filet" category="FOOD#QUALITY" polarity="positive" from="61" to="66"/>
                <Opinion target="frites" category="FOOD#QUALITY" polarity="negative" from="100" to="106"/>
            </Opinions>
        </sentence>
        <sentence id="1730014:3">
            <text>Will absolutely visit again.</text>
            <Opinions>
                <Opinion target="NULL" category="RESTAURANT#GENERAL" polarity="positive" from="0" to="0"/>
            </Opinions>
        </sentence>
        <sentence id="1730014:4">
            <text>Maybe tomorrow ;-)</text>
        </sentence>
    </sentences>
</Review>
```

Figure 27: SLA for Review id "1730014" (restaurants domain)

In the case of conflicting opinions where the dominant sentiment is not clear the **conflict** label has to be assigned. For example, in the following review in Fig. 28, the reviewer expresses a positive opinion about one dish (pad seew chicken) and a negative opinion about another one (pad thai) without providing any further information. So, in this case the opinion about the quality of the food is considered to be conflict (Fig. 29).

```
<Review rid="505535">
    <sentences>
        <sentence id="505535:0">
            <text>this little place has a cute interior decor and affordable city prices.</text>
            <Opinions>
                <Opinion target="interior decor" category="AMBIENCE#GENERAL" polarity="positive" from="29" to="43"/:
                <Opinion target="place" category="RESTAURANT#PRICES" polarity="positive" from="12" to="17"/>
            </Opinions>
        </sentence>
        <sentence id="505535:1">
            <text>the pad se ew chicken was delicious, however the pad thai was far too oily.</text>
            <Opinions>
                <Opinion target="pad se ew chicken" category="FOOD#QUALITY" polarity="positive" from="4" to="21"/>
                <Opinion target="pad thai" category="FOOD#QUALITY" polarity="negative" from="49" to="57"/>
            </Opinions>
        </sentence>
        <sentence id="505535:2">
            <text>i would just ask for no oil next time.</text>
        </sentence>
    </sentences>
</Review>
```

Figure 28: SLA for Review id "505535" (restaurants domain)

```
<Opinions>
    <Opinion category="AMBIENCE#GENERAL" polarity="positive"/>
    <Opinion category="RESTAURANT#PRICES" polarity="positive"/>
    <Opinion category="FOOD#QUALITY" polarity="conflict"/>
    <Opinion category="RESTAURANT#GENERAL" polarity="positive"/>
</Opinions>
```

Figure 29: TLA for Review id "505535" (restaurants domain)

Similarly, in the following review (Fig. 30) in the laptops domain the DESIGN&FEATURES aspect is assigned the conflict label (Fig. 31) since the different opinions refer to different attributes that are both denoted by the particular label (DESIGN&FEATURES) (i.e. positive opinion about the "appearance", negative opinion about the "(extra/missing) features/components").

```
<Review rid="375">
    <sentences>
        <sentence id="375:0">
            <text>This laptop is a great price and has a sleek look.</text>
            <Opinions>
                <Opinion category="LAPTOP#PRICE" polarity="positive"/>
                <Opinion category="LAPTOP#DESIGN_FEATURES" polarity="positive"/>
            </Opinions>
        </sentence>
        <sentence id="375:1">
            <text>Runs smooth and quick.</text>
            <Opinions>
                <Opinion category="LAPTOP#OPERATION_PERFORMANCE" polarity="positive"/>
            </Opinions>
        </sentence>
        <sentence id="375:2">
            <text>I wish it had a webcam though, then it would be perfect!</text>
            <Opinions>
                <Opinion category="LAPTOP#DESIGN_FEATURES" polarity="negative"/>
                <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 30: SLA for Review id "375" (laptops domain)

```
<Opinions>
    <Opinion category="LAPTOP#PRICE" polarity="positive"/>
    <Opinion category="LAPTOP#DESIGN_FEATURES" polarity="conflict"/>
    <Opinion category="LAPTOP#OPERATION_PERFORMANCE" polarity="positive"/>
    <Opinion category="LAPTOP#GENERAL" polarity="positive"/>
</Opinions>
```

Figure 31: TLA for Review id "375" (laptops domain)

2.4.1.2 DATASETS AND ANNOTATION

Based on the above observations-guidelines, the ABSA-2015 benchmark datasets were annotated with TLA. In a first phase, the existing SLA were aggregated automatically for each review and the duplicates were removed. Then, the resulting tuples were inspected and, when needed, the appropriate modifications in the polarity labels were made. Each review text was assigned also an overall (positive, negative, neutral, or conflict) sentiment label about the target entity (LAPTOP#GENERAL, RESTAURANT#GENERAL) even if it was not explicitly stated. Furthermore, a set of 170 new review texts (90 for restaurants and 80 for laptops) was collected and annotated from scratch both at the sentence- and the text-level. Hence, the ABSA-2015 benchmark datasets were extended with more review texts annotated at the sentence level and also enriched with TLA as illustrated below in Table 8. The asterisk indicates that the TLA correspond to a part of the data collections (in particular 425 review texts and 475 texts contain TLA in the RST and the LPT domain, respectively), since some texts were removed after the annotation process (e.g. reviews consisting of less than 3 sentences).

Dataset	Texts	Sentences	SLA ABSA tuples	TLA ABSA tuples
LPT	530	3308	3710	2627*
RST	440	2676	3366	1857*
Total	970	5984	7076	4484

Table 8: The ABSA-2015 benchmarks extended

The proposed datasets, along with the text-level annotation framework were adopted by the ABSA shared task that was organized in the context of the SemEval 2016 workshop as a follow up of the SemEval 2015 ABSA (section 2.4.3).

2.4.2 THE ABSA-2016 SEMEVAL CHALLENGE

The SE-ABSA16 task[29] was a follow up of the SE-ABSA15 task. In addition to sentence-level annotations, SE-ABSA16 accommodated text-level ABSA annotations and provided the respective training and testing data. Furthermore, the SE-ABSA15 annotation framework was extended to new domains and applied to languages other

[29] http://alt.qcri.org/semeval2016/task5/

than English (Arabic, Chinese, Dutch, French, Russian, Spanish, and Turkish). Similarly to the SE-ABSA15 task, the input for the participating systems consisted of whole review texts. Participants were free to choose the subtasks/languages/domains they wished to participate in. In particular, the task consisted of the following subtasks and slots:

- **Subtask 1: Sentence-level ABSA**. Similarly to SE-ABSA15, given a review text about a target entity of interest laptop or restaurant, the goal was identify all the opinion tuples with the following types (tuple slots) of information:

  - **Slot 1**: **Aspect Category Detection** (E#A pairs).
  - **Slot 2: Opinion Target Expression** (OTE).
    - **Slots 1&2**: Link the extracted OTEs to the respective Aspect Categories.
  - **Slot 3: Sentiment Polarity Classification**.

- **Subtask 2: Text-level ABSA.** Given a customer review about a target entity, the goal was to identify a set of {*cat, pol*} tuples that summarize the opinions expressed in the review. Following the text-level annotation framework proposed in this thesis in the previous section, *cat* could be assigned the same values as in SB1 (E#A tuple), while *pol* could be set to "*positive*", "*negative*", "*neutral*", or "*conflict*".

- **Subtask 3: Out-of-domain ABSA**. In SB3 participants had the opportunity to test their systems in domains for which no training data was made available; the domains remained unknown until the start of the evaluation period. Test data for SB3 were provided only for the museums domain in French.

The task provided baselines, training and testing data for SB1 and SB2 in several languages and domains. In particular, the extended SE-ABSA15 datasets described in the previous section were provided for training and testing for the restaurants and laptops domain in SBs 1& 2 for English (**En**). The restaurants domain was supported also for Dutch[30] (**Du**), French[31] (**Fr**), Russian[32] (**Ru**), Spanish[33] (**Es**), and Turkish[34] (**Tu**) by respective research teams following the new framework and the annotation

[30] Research team: LT3, Ghent University, Ghent, Belgium

[31] Research team: LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Orsay, France

[32] Research team: Lomonosov Moscow State University, Moscow, Russian Federation and Vyatka State University, Kirov, Russian Federation

[33] Research team: Universitat Pompeu Fabra, Barcelona , Spain and SINAI, Universidad de Jaén, Spain

[34] Research team: Dept. of Computer Engineering, Istanbul Technical University, Turkey and Turkcell Global Bilgi, Turkey

guidelines proposed in this thesis. Below are some examples of sentences from restaurants customer reviews discussing the aspect categories FOOD#QUALITY, DRINKS#STYLE_OPTIONS, SERVICE#GENERAL, and AMBIENCE#GENERAL in six languages. The OTEs and the sentiment polarity in each case are provided in the brackets:

- **FOOD#QUALITY**

En: *Salads are a delicious way to begin the meal.* → {"***Salads****"*, positive}
Du: *Nergens in Hasselt zijn de pannekoeken zo lekker als hier!!!!*→ {"***pannekoeken****"*, positive}
Fr: *Ces pauvres poulpes auraient pu mourir d'un excès de cholestérol s'ils n'avaient pas fini sur la plancha.* → {"***poulpes****"*, negative}
Ru: Все блюда очень вкусные,приготовлены по-домашнему. → {"**блюда***"*, positive}
Es: *Para niños croquetas buenísimas y hamburguesas de buena calidad***.** → {"***croquetas",*** positive}, {" ***hamburguesas****"*, positive}
Tu: *Hamburgerini pek beğenmedim ama diğer yemekleri lezzetli olabilir belki.* → {"***yemekleri",*** positive}, {" ***Hamburgerini****"*, negative}

- **DRINKS#STYLE_OPTIONS**

En: *The sake menu should not be overlooked!* → {"***sake menu****"*, positive}
Du: *Het aan bod van bieren en andere is zeer beperkt.* → {"***bieren****"*, negative}
Fr: *Carte des vins inexistante.* → {"***carte des vins****"*, negative}
Ru: Так же большой выбор коктейлей, некоторые очень неплохие. → {"**коктейлей***"*, positive}
Es: *Debe mejorar muy mucho su carta de vinos.* → {"***carta de vinos****"*, negative}
Tu: *60'a yakın çeşitte çaydan birini mutlaka seviceksiniz.* → {"***çaydan****"*, positive}

- **SERVICE#GENERAL**

En: *Service was slow, but the people were friendly.* → {"***Service***", negative}, {"***people***", positive}
Du: *Snelle bediening en vriendelijke personeel moet ook gemeld worden!!* → {"***bediening***", positive}, {"***personeel***", positive}
Fr: *Le service est impeccable, personnel agréable.* → {"***service***", positive}, {"***personnel***", positive}

Ru: Про сервис ничего негативного не скажешь- быстро подходят, все улыбаются, подходят спрашивают, всё ли нравится. → {"**сервис***"*, neutral}
Es: *También la rapidez en el servicio*. → {"***servicio****"*, positive}
Tu: *Servisi hizli valesi var.* → {"***Servisi****"*, positive}

AMBIENCE#GENERAL

En: *LOVE the atmosphere - felt like I was in Paris*. → {"***atmosphere****"*, positive}
Du: *Bovendien houden wij van de gezellige familiale sfeer*. → {"***sfeer****"*, positive}
Fr: *La salle est très agréable et tranquille.* → {"***salle****"*, positive}
Ru: Все детали продуманы до мелочей и вместе создают замечательную атмосферу тепла и комфорта. → {"**атмосферу***"*, positive}
Es: *Y el lugar está muy bien decorado, sencillo, pero elegante.* → {"***lugar****"*, positive}
Tu: *Dekorasyonu renkleri cok sicak ve sevimli.* → {"***Dekorasyonu****"*, positive}, {"***renkleri****"*, positive}

As it can be seen in the examples, the unstructured text is transformed into structured information within tuples using common annotation guidelines and labels set across all languages for aspect categories (E#A pairs). In addition, through the OTE slot we also obtain the linguistic evidence for the entities that are evaluated in each case and in each language. Based on the sentence-level annotations (SB1), the Du, Ru, Sp, and Tu datasets were further annotated with text-level ABSA tuples (SB2) following the respective guidelines presented in the previous section. Similarly, the hotels annotation schema and guidelines proposed in this thesis were adopted for the creation of Arabic[35] datasets in the specific domain both at the sentence (SB2) and text-level (SB2). Here is an example, of an annotated hotel customer review in Arabic at the sentence-level:

[35] Research team: Computer Science Dept., Jordan University of Science and Technology Irbid, Jordan

```
<Review rid="1271">
    <sentences>
        <sentence id="1271:0">
            <text>منتجع مذهل وخصوصي  كانت تجربتي تتميز بتعرف حصري للنبيذ العالمي في هذا
            المنتجع، لقد اخترنا هذا الفندق بناءا على التعليقات الجيدة المختلفة التي تلقيناها
            </text>
            <Opinions>
                <Opinion target="منتجع" category="FACILITIES#QUALITY" polarity="positive"/>
                <Opinion target="للنبيذ" category="FOOD_DRINKS#QUALITY" polarity="positive"/>
            </Opinions>
        </sentence>
        <sentence id="1271:1">
            <text>البحر الكاريبي رائع حقاً لكن هذا المنتجع يصنع الفرق لأنه يجمع بين روعة
            المكان مع المستوى الممتاز من الاهتمام من قبل الموظفين. أقترح عليكم زيارة هذا
            المنتجع المذهل كلاوديو كينتانيلا بيونس آيرس، الأرجنتين</text>
            <Opinions>
                <Opinion target="المنتجع" category="LOCATION#GENERAL" polarity="positive"/>
                <Opinion target="الاهتمام" category="SERVICE#GENERAL" polarity="positive"/>
                <Opinion target="منتجع" category="FACILITIES#QUALITY" polarity="positive"/>
            </Opinions>
        </sentence>
    </sentences>
</Review>
```

Figure 32: Annotated hotel customer review in Arabic at the sentence-level

The laptops annotation schema was extended to two other domains of consumer electronics, digital cameras and mobile phones. The mobile phones domain was supported for Chinese[36] (**Ch**) and Dutch (**Du**), while the cameras domain for Chinese. Examples of annotated sentences in the laptops (1), phones (2,3) and cameras (4) domains are shown below:

1. *It is extremely portable and easily connects to WIFI at the library and elsewhere*. → {**LAPTOP#PORTABILITY**, positive} , {**LAPTOP#CONNECTIVITY**, positive}
2. *Apps starten snel op en werken vlot, internet gaat prima*. → {**SOFTWARE#OPERATION_PERFORMANCE**, positive}, {**PHONE#CONNECTIVITY**, positive}
3. wifi 不能自动连接。 → {**PHONE# CONNECTIVITY**, negative}
4. 更轻便的机身也便于携带。→ {**CAMERA# PORTABILITY**, positive}

In addition, the ABSA framework proposed in this thesis was extended to two new domains, telecommunications and museums, for which annotation guidelines were compiled with respect to the specific annotation framework. The telecommunications domain was supported by Turkcell Global Bilgi (Turkcell Global Bilgi, 2015)[37] for the Turkish language (Twitter data), while the museums domain for French (Apidianaki, Tannier, and Richart, 2016). Below are two examples:

[36] Research team: Harbin Institute of Technology, Harbin, Heilongjiang, P.R. China

[37] Turkcell Global Bilgi. Web. 7 Dec. 2015. <http://www.global-bilgi.com.tr/>.

5. *#Internet kopuyor sürekli :(@turkcell* → {**INTERNET#COVERAGE**, "*Internet*", positive}

6. *5€ pour les étudiants, ça vaut le coup.* → {**MUSEUM#PRICES**, NULL, positive}

Overall, a total of 39 datasets were provided in the context of the SE-ABSA16 task; 19 for training and 20 for testing. The texts were from 7 domains and 8 languages. The datasets for the domains of restaurants (REST), laptops (LAPT), mobile phones (PHNS), digital cameras (CAME), hotels (HOTE) and museums (MUSE) consist of customer reviews, whilst the telecommunication domain (TELC) data consists of tweets. A total of 70790 manually annotated ABSA tuples were provided for training and testing; 47654 sentence-level annotations (SB1) in 8 languages for 7 domains, and 23136 text-level annotations (SB2) in 6 languages for 3 domains. Table 1 provides more information on the distribution of texts, sentences and annotated tuples per dataset. The full inventories for each domain and information about the annotation process in each language are available at the at the task overview paper[38] (Pontiki et al., 2016).

Lang.	Domain	Subtask	Train			Test		
			#Texts	#Sent.	#Tuples	#Texts	#Sent.	#Tuples
EN	REST	SB1	350	2000	2507	90	676	859
EN	REST	SB2	335	1950	1435	90	676	404
EN	LAPT	SB1	450	2500	2909	80	808	801
EN	LAPT	SB2	395	2375	2082	80	808	545
AR	HOTE	SB1	1839	4802	10509	452	1227	2604
AR	HOTE	SB2	1839	4802	8757	452	1227	2158
CH	PHNS	SB1	140	6330	1333	60	3191	529
CH	CAME	SB1	140	5784	1259	60	2256	481
DU	REST	SB1	300	1711	1860	100	575	613
DU	REST	SB2	300	1711	1247	100	575	381
DU	PHNS	SB1	200	1389	1393	70	308	396
FR	REST	SB1	335	1733	2530	120	696	954
FR	MUSE	SB3	-	-	-	162	686	891
RU	REST	SB1	302	3490	4022	103	1209	1300
RU	REST	SB2	302	3490	1545	103	1209	500
ES	REST	SB1	627	2070	2720	286	881	1072
ES	REST	SB2	627	2070	2121	286	881	881
TU	REST	SB1	300	1104	1535	39	144	159
TU	REST	SB2	300	1104	972	39	144	108
TU	TELC	SB1	-	3000	4082	-	310	336

Table 9: Datasets provided for SE-ABSA16

The task attracted 245 submissions from 29 teams. The majority of the submissions (216 runs) were for SB1. The newly introduced SB2 attracted 29 submissions from 5

[38] http://www.aclweb.org/anthology/S16-1002

teams in 2 languages (English and Spanish). As expected, most of the submissions (168) were runs for the restaurants domain, since the restaurants classification schema is less fine-grained (complex) compared to the other domains (e.g. lapt). In addition, this domain was supported for 6 languages enabling also multilingual or language-agnostic approaches. Regarding the participation per language, the majority of the submissions (156/245) were for English. Most teams (20) submitted results only for one language. Of the remaining teams, 3 submitted results for 2 languages, 5 teams submitted results for 3-7 languages, while only one team participated in all languages.

The evaluation process was similar to the one followed in SE-ABSA15. In Phase A, the participants were asked to return separately the aspect categories (Slot1), the OTEs (Slot2), and the {Slot1, Slot2} tuples for SB1. For SB2 the respective text-level categories had to be identified. In the second phase (Phase B), the gold annotations for the test sets of Phase A were provided and participants had to return the respective sentiment polarity values (Slot3). Similarly to SE-ABSA15, F-1 scores were calculated for Slot1, Slot2 and {Slot1, Slot2} tuples, by comparing the annotations that a system returned to the gold annotations (using micro-averaging), and accuracy for Slot 3 (sentiment polarity classification).

Fig. 33-36 below present a comparison of the scores achieved for aspect category detection in the restaurants and the laptops domain in SE-ABSA15 and SE-ABSA16 respectively, since the results for the English language are directly comparable to the ones of the previous year. As illustrated below, in 2016 we had significantly more submissions in both domains. In the restaurants domain the best systems in 2016 performed almost 10% higher; this was probably due to the fact that they had more training data or the experience from the previous year. However, this was not the case in the laptops domain where we had slightly better results for the first team and the rest of the results were almost the same. This was probably due to the fine-grained classification schema. As for the sentiment polarity classification slot, again there were more submissions in 2016, better results in the restaurants domain and slightly better results in the laptops domain.

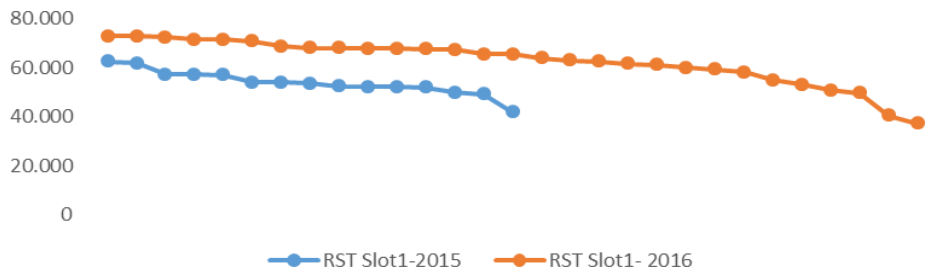


Figure 33: Restaurants Slot 1: ABSA2015 and ABSA2016 F-1 scores

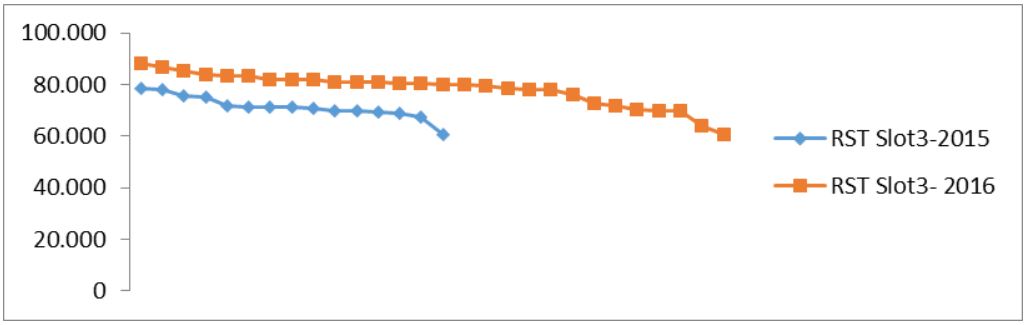


Figure 34: Restaurants Slot 3: ABSA2015 and ABSA2016 Accuracy scores

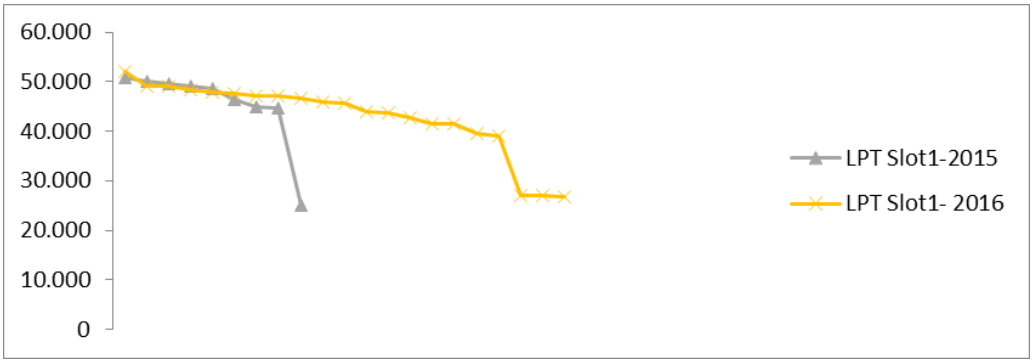


Figure 35: Laptops Slot 1: ABSA2015 and ABSA2016 F-1 scores

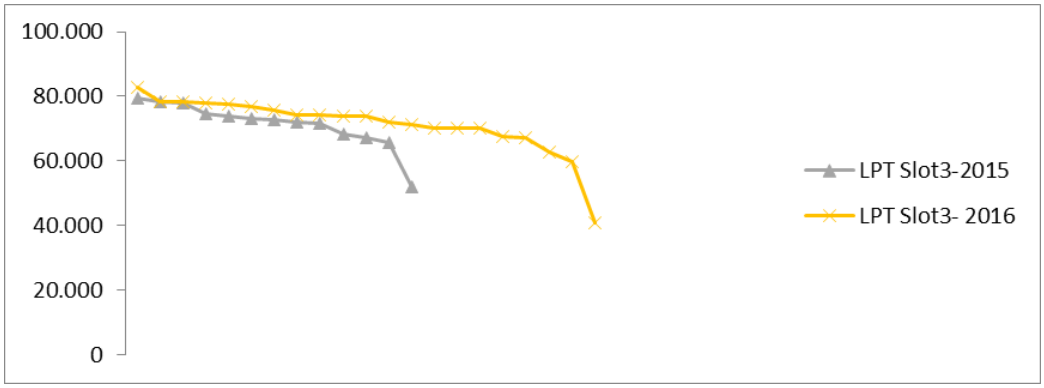


Figure 36: Laptops Slot 3: ABSA2015 and ABSA2016 Accuracy scores

The results were slightly better also in Slot 2 and Slots1&2 jointly in SE-ABSA16 as compared to SE-ABSA15 as presented in Fig. 37-38. An interesting observation is that, unlike SE-ABSA15, Slot1 (aspect category detection) attracted significantly more submissions than Slot2 (OTE extraction); this may indicate a shift towards concept level ABSA approaches.

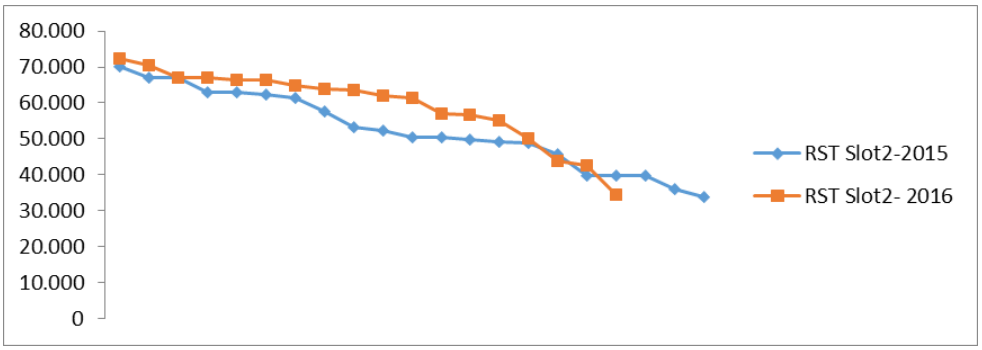


Figure 37: Restaurants Slot 2: ABSA2015 and ABSA2016 F-1 scores

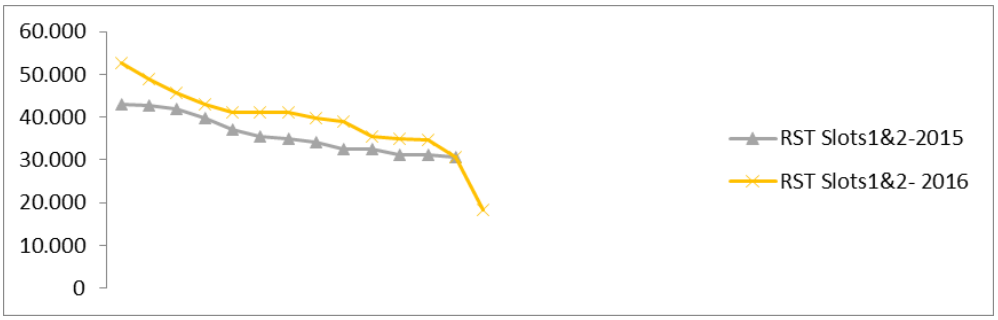


Figure 38: Restaurants Slots 1&2: ABSA2015 and ABSA2016 F-1 scores

Fig. 39-42 below present a comparison of the scores achieved in all slots and each subtask in the restaurants domain per language. The best scores in all slots and subtasks were achieved for the English language. This may be due to the fact that it

was the second year or that there were more available resources for the particular language. The maximum difference between the best best score and the worst best score in both slots for all languages is almost 10%. It seems that aspect category detection presents the same level of difficulty for Dutch, French, Russian and Turkish. Furthermore, as it is indicated by the results, sentiment polarity classification is easier than aspect category detection in all languages, and the best scores were achieved for English, Spanish and Turkish.

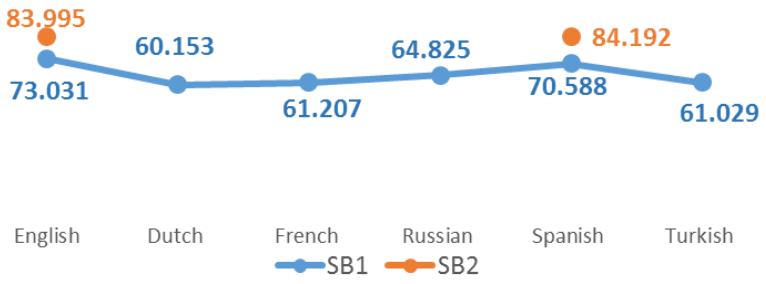


Figure 39: ABSA16 Slot1 Best F-1 scores for Restaurants

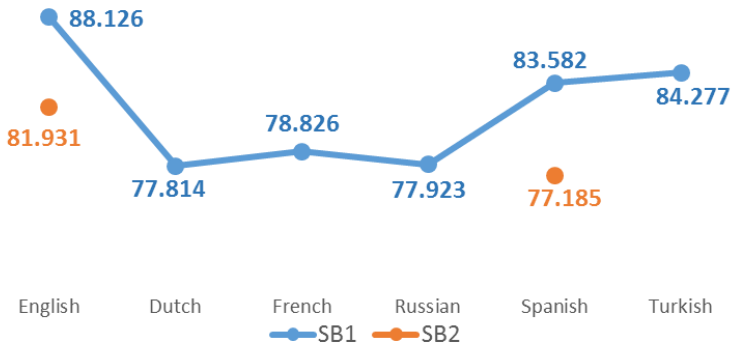


Figure 40: ABSA 2016 Slot3 Best Accuracy Scores for Restaurants

In SB1, the teams with the highest scores for Slot1 and Slot2 achieved similar F-1 scores in most cases (e.g. en/rest, es/rest, du/rest, fr/rest), which shows that the two slots have a similar level of difficulty. Opinion target extraction (Slot 2) appeared to be harder for Dutch and Russian; this may be due to different linguistic structures or language specific phenomena. Even though systems achieve high scores when detecting aspect categories and opinion target expressions separately, when they have to link these two types of information there is a drop in performance by almost 20% (Fig. 42).

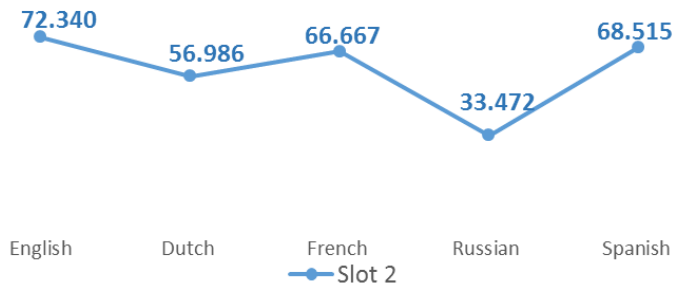


Figure 41: ABSA 2016 Slot 2 Best F-1 Scores for Restaurants

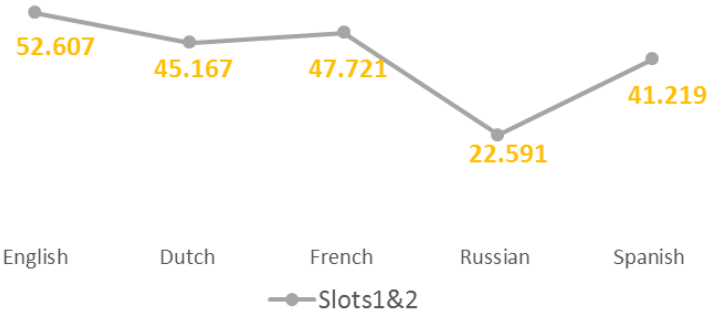


Figure 42: ABSA 2016 Slots 1&2 Best F-1 Scores for Restaurants

As for the other domains, the best scores are presented below in Fig. 43-44. As expected the performance in aspect category detection is significantly lower in the consumer electronic domains (laptops, cameras, phones) due to the complexity of the classification schema as compared to the restaurants one. Again, sentiment polarity classification is easier than aspect category detection in all languages and domains.

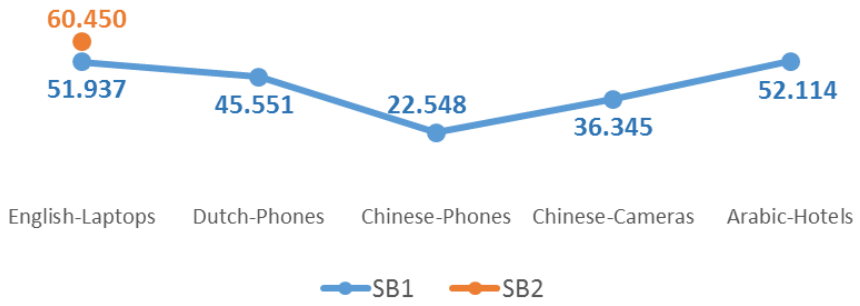


Figure 43: ABSA 2016 Slot1 Best F-1 Scores for Other Domains

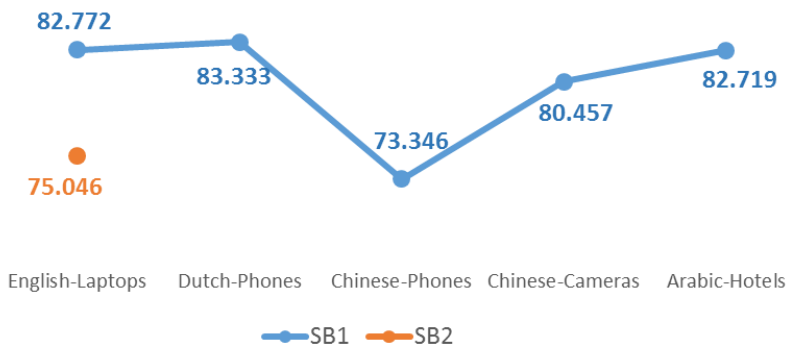


Figure 44: ABSA 2016 Slot3 Best Accuracy Scores for Other Domains

As regards the results for text-level ABSA (SB2), if we compare the results for SB1 and SB2, we notice that the SB2 scores for Slot1 are significantly higher (e.g. en/lapt, en/rest, es/rest) even though the respective annotations are for the same (or almost the same) set of texts. This is due to the fact that it is easier to identify whether a whole text discusses an aspect category than finding all the sentences in the text discussing the particular aspect. On the other hand, for Slot3, the SB2 scores are lower (e.g., en/rest, es/rest, ru/rest, en/lapt) than the respective SB1 scores. This is observed mainly because an aspect may be discussed at different points in a text and often with a different sentiment. In such cases a system has to identify the dominant sentiment, which usually is not trivial.

The participating teams experimented with a variety of features, techniques, and resources. Starting with aspect category detection, the best score was achieved by the deep learning approach of the NLANGP team (Toh and Su, 2016) that enhanced their SE-ABSA2015 submission (Toh and Su, 2015) with additional features learnt from neural networks achieving 73.031% in the restaurants and 51.937% in the laptops domain. In particular, they employed a binary classifier for each category using a single layer feedforward network algorithm (Vowpal Wabbit tool) and Deep Convolutional Neural Networks (CNNs) with n-grams, syntactic parsing, name lists and word clusters learnt from Amazon and Yelp data, word embeddings (word2vec, GloVe) and the CNN probability output as additional feature. The task attracted other deep learning approaches (e.g. Ruder, Ghaffari, and Breslin, 2016; Tamchyna and Veselovská, 2016). As for the more traditional methods, they performed also well. For example, the AUEB team (Xenos et al., 2016) achieved approximately 3% lower score from the first team and was ranked 4th and 2nd in the restaurants and laptops domain, respectively, using one classifier for each E#A, multiple ensembles based on Support Vector Machine (SVM) classifiers and unigrams, bigrams, POS tags, lexicons from training data and word embeddings (word2vec).

As for the OTE slot, most of the teams modelled it as a sequential labelling task using CRFs with a variety of features. Again the best score (72.34%) was achieved by a deep learning approach (Toh and Su, 2016) that used CRFs along with Recurrent Neural networks based on features derived from name lists, word clusters and double propagation for the detection of candidate OTEs. The second best score (70.44%) was achieved by a traditional CRF method (Xenos et al., 2016) that relied on a variety of morphological, lexical and contextual features.

Finally, in the sentiment polarity classification slot the XRCE team (Brun, Perez, and Roux, 2016) achieved best scores in the restaurants domain not only in English (88.126%) and also for French (78.826%) with an Interactive feedbacked ensemble method pipeline (CRFs & Elastic Net regression models) that relied on robust syntactic and semantic parsing (i.e. a full processing chain including tokenization, morpho-syntactic analysis, POS tagging, Named Entity Detection, chunking, extraction of dependency relations, extraction of semantic information about aspect targets and their polarities (lexical information about word polarities and semantic classes, and aspect categories as features). Another effective cross-lingual approach was the one of the IIT-TUDA team (Kumar et al., 2016) that submitted results for 7 languages and 4 domains achieving best scores in many cases: 82.772% (English laptops, ranked 1st), 86.729% (English restaurants, ranked 2nd), 83.582% (Spanish restaurants, ranked 1st), 72.222% (French restaurants, ranked 4th), 73.615% (Russian restaurants, ranked 3rd), 76.998% (Dutch restaurants, ranked 2nd), 84.277% (Turkish restaurants, ranked 1st), 81.72% (Arabic hotels, ranked 2nd), 82.576% (Dutch phones, ranked 2nd). Their method combined SVMs with domain dependency and distributional semantics; polarity lexica (for each language) were constructed using seeds and external corpora, n-grams, and E#A pairs as a binary feature for sentiment classification. More details about the submitted systems and the evaluation results are available at the task overview paper (Pontiki et al., 2016).

3. VERBAL AGGRESSION AS AN INDICATOR OF ONLINE XENOPHOBIC ATTITUDES

This chapter presents the second part of this thesis that is the fine-grained SA framework designed and implemented in the context of the XENO@GR project for examining VA as an indicator of xenophobic attitudes in Greek Social Media. Starting with the background and the research strand of this work (3.1), section 3.1.1 provides a comprehensive overview of the key concepts, types, causes and effects of offline and online VA (3.1.1.1), and the related state-of-the-art computational methods (3.1.1.2). The relation of xenophobia to VA along with an overview of the historical context of this phenomenon in Greece, and the research goals of this thesis are presented in sections 3.1.2.1 and 3.1.2.1, respectively. Section 3.2 presents the methodology followed for building the VA analysis framework, and subsequently, section 3.3 discusses the analysis results with regard to the specific RQs that this thesis aims to address focusing on the amount, the type and the content of the aggressive messages. This chapter concludes (3.4) with a discussion of the most interesting observations regarding the quantitative and the qualitative analysis. It also presents some further insights about the nature of online xenophobic behavior in Greece (3.4.1). The limitations and the possible future research directions of the presented work are discussed in 3.4.2.

3.1 BACKGROUND AND RESEARCH STRAND

3.1.1 VERBAL AGGRESSION

3.1.1.1 KEY CONCEPTS, TYPES, CAUSES, AND EFFECTS

The concept and the content of VA have been studied within the scope of psychology and communication studies (Hamilton and Hamble, 2011; Infante and Wigley, 1986; Kinney, 1994) in different contexts (e.g. marriage, workplace, parental relations). VA is expressed when language is used to "*inform another that she or he is bad, possesses negative qualities, or is not meeting some internal or external standard*" (Kinney, 1994). It is also defined as a personality trait that "*predisposes persons to attack the self-concepts of other people instead of, or in addition to, their positions on topics of communication*" (Infante and Wigley, 1986) or as a stance that "*involves using messages to attack other people or those aspects of their lives that are extensions of*

their identity" (Hamilton and Hample, 2011). The underlying concept of VA is that of aggressive communicative behavior. According to Infante (1987) a communicative behavior is aggressive "*if it applies force...symbolically in order, minimally, to dominate and perhaps damage, or maximally, to defeat and perhaps destroy the locus of attacks*". VA is also viewed as a subset of hostility, meaning that "*all verbal aggressiveness is hostile, but not all hostility involves verbal aggression*" (Infante and Rancer, 1996).

Furthermore, as Feshbach (1970) states, "*aggression is a social act*". This implies that the act of hurting others requires social engagement. Caprara and Pastorelli (1989) argue that "*aggression is always a phenomenon which develops along a sequence of interpersonal exchanges and in a social context*." This perspective highlights the social nature of aggression and places particular emphasis on the roles that communication and relationships play in the process of harming others. Furthermore, the concept of VA presupposes the speech act theory performative approach to language, which addresses speaking as intentionally doing things with words (Austin, 1976). Moving a step forward in the perception of verbal aggression, the intentional use of language can be associated with the social construction of aggression; thus, in terms of social psychology, language can be viewed as a "weapon" (Graumann, 1998). Given that language -when used as a weapon- can be both psychologically and physiologically damaging, the effects of VA have been subject of research over several decades in different types of studies and examined in terms of physical and affective reactions, psychological states as well as relational and behavioral consequences.

Verbally attacked subjects have been found to experience increased heart rates (Glass et al., 1980; Levenson and Gottman, 1983, 1985), blood pressure (Glass et al., 1980; Ewart, Burnett and Taylor, 1983; Kiecolt-Glaser et al., 1993), plasma epinephrine (Glass et al., 1980), skin conductance (Levenson and Gottman, 1983, 1985), and a down-regulated immune system (Ewart, Burnett and Taylor, 1983; Kiecolt-Glaser et al., 1993). In addition to physiological findings such as increased heart rates and diastolic blood pressure, other studies found that participants who were insulted during laboratory tasks experienced also aggression (Rule and Hewitt, 1971) and self-reports of anger (Gentry, 1972), respectively. Such physiological and affective findings suggest that verbal attacks are perceived as threats eliciting a variety of distress forms, which -depending on the type and the severity of the attack- may range from anxiety and upset to depression and physical illness (Kinney, 1994).

Empirical studies have shown a positive relationship between VA and emotional exhaustion (Karatepe, Yorganci and Haktanir, 2009; Yaratan and Uludag, 2012), cynicism and reduced professional efficacy as burnout dimensions (Yaratan and Uludag, 2012). Receiving verbal attacks in a frequent and consistent basis turns to verbal abuse and can lead to low self-regard, since the receivers tend to be cynical, unhappy, and troubled (Hoffman, 1984; Rohner and Rohner, 1980; Vissing et al., 1991). For example, verbally abused children were found to be more aggressive and emotionally unstable, and to have lower self-esteem and self-adequacy than children that were accepted by their parents (Rohner and Rohner, 1980), while women that received sexual harassment at their workplace in the form of aggressive communication experienced loss of confidence, lowered self-esteem, and problems with relationships (Tangri, Burt and Johnson, 1982). More recent studies suggest that marital aggression may be one important factor that contributes to the development of drinking problem (Kelley, Lewis and Mason, 2015).

The link between alcohol and aggression was so far well-established mainly by studies that consider alcohol consumption as a causal factor for aggression, rather than the other way around (e.g. Parrott and Giancola, 2006). Other studies focusing on specific contexts, associate VA also with factors such a history of violence and previous drug use in the case of impatient verbal aggressive behavior on psychiatric wards (Stewart and Bowers, 2013). In the context of communication studies, the causes of VA are summarized by Infante and Wigley (1986) as follows: a) *Frustration* (having a goal blocked by someone, having to deal with a disdained other), b) *Social Learning* (individuals are conditioned to behave aggressively and this can include modelling where the person learns the consequences of a behavior vicariously by observing a model such as a character in a television program), c) *Psychopathology* (involves transference where the person attacks with verbally aggressive messages those people who symbolize unresolved conflict), and d) *Argumentative skill deficiency* (individuals resort to VA because they lack the verbal skills for dealing with social conflict constructively). Generally, research findings suggest that VA is augmented by disturbing life events (Day and Hamblin, 1964), viewing violent films (Sebastian et al., 1978), drugs (Haward, 1958), and brain damage (Vondráček, Horvai and Študent, 1964) and inhibited by positive events such as argumentation training (Infante and Rancer, 1996).

Depending on the research type and goals, the approach, and the social context (e.g. marriage, workplace, etc.) several types of VA have been proposed. For example, Infante (1987) and Infante et al. (1990) suggest a ten-way classification schema in the context of marital disputes (character attacks, competence attacks, background

attacks, physical appearance attack, Malediction, Teasing, Ridicule, Threats, Swearing, Nonverbal emblems (e.g. facial expressions, gestures, eye behaviors used to attack one's self-concept)). Kinney (1994) suggests an inductively derived typology of VA based on the domains of attacks as follows: a) group membership attacks (messages that associated or placed one into a negatively evaluated group), b) personal failings attacks (messages that pointed out personal deficits), and c) relational failings attacks (messages that described one's social or interpersonal relationship deficits). According to Kinney (1994) there is correspondence with Infante's et al. (1990) classification involving *background attacks, character attacks, competence attacks,* and *physical appearance attacks*, and the fact that *maledictions, teases, ridicules, threats* and *swears* did not surface in his typology suggests that they may represent methods of attack rather than targets of attack. In the context of impatient verbally aggressive behavior on psychiatric wards –where staff members were the most frequent target of aggression– incidents of verbal aggression were categorized in order of prevalence as follows (Stewart and Bowers, 2013): abusive language (defined as swearing, use of foul language, insults, use of sexually inappropriate language or more generic terms such as "verbal abuse" or "abusive"), shouting (defined as screaming, yelling, making loud noises or being noisy), threats (verbal threats of violence, damage to property, against self or other actions), expressions of anger (anger was coded when the notes mentioned the patient being angry, cross or in a heated confrontation), and racist comments.

All the above "traditional" studies examine VA in face-to-face communications. However, emerging technologies have expanded the boundaries of VA to the digital world creating new forms and also propagating the effects to a large scale affecting billions of people. The increase of the number of the users and online interaction gives rise to aggression incidents and related events such as ***flaming***, ***cyberbullying*** and ***hate speech***. An important issue with online VA seems to be the absence of responsibility because of the possibility of maintaining an anonymous profile while posting offensive written content (Lee and Kim, 2015). Online interaction differs from face-to-face communication, especially because anonymity and pseudonymity enable a more disinhibited self (Bandura, 2004); the fact that individuals can mask their identity or operate anonymously seems to influence online disinhibition, namely the tendency to say things in cyberspace that would not be said or done in person (Vandebosch and Cleemput, 2009). In particular, Suler (2004) distinguishes two main behavioral categories that fall under the ***online disinhibition*** effect; *benign disinhibition* that is behavior in which people might self-disclose more on the internet than they would in real life, or go out of their way to help someone or show kindness, and *toxic disinhibition* that includes rude language, threats, and visiting places of

pornography, crime, and violence on the internet–places the person might not go to in real life.

Given that behavior includes language and actions, the term "toxic" is also used to describe language that is hurtful as a synonym to aggressive language[39]. The term ***toxicity*** has been adopted also in the context of online content; for example, the Google Perspective Model[40] aims to score toxicity of abusive comments in online platforms on a scale from 0 ("healthy") to 1 ("very toxic"), where toxic refers to "*rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion*". Other typologies (Laineste, 2012) consider three general types of online VA based on its intensity as follows: a) Strong aggression: when a text expresses straightforward violence, displays nationalist or racist slogans, calls for physical actions against "others", and praises historical violence, b) Medium aggression: uses or introduces new negative stereotypes about the "other", swearing, accusing of stupidity, naming and slurs, c) Mild aggression: when jokes and other forms of humour are used, the target is presented in a negative context, or as possessing negative influence, racist viewpoints are referred to or a previous flame is cited without any counter-arguments. Work on online aggressive/abusive language has spread across several overlapping fields; this can cause some confusion as different works may tackle specific aspects of abusive language, define the term differently, or apply it to specific online domains only (Nobata et al., 2016). Here is an overview of the three main types of online VA found in the literature: flaming, cyberbullying, and hate speech.

- **Flaming.** Flaming is the most commonly observed manifestation of toxic online disinhibition (Voggeser, Singh and Göritz, 2018; Johnson, Cooper and Chin, 2009; Alonzo and Aiken, 2004), and is often used as a synonym to online VA (Tereszkiewicz, 2012). It has been characterized in the literature as "verbal aggression, blunt disclosure, and nonconforming behavior" (Parks and Floyd, 1996), "hostile verbal behavior" (Thompsen and Foulger, 1996), "a form of social aggression" (Colomb and Simutis, 1996), or "emotional outbursts" (Korenman and Wyatt, 1996). It is typically represented as "rude or insulting messages" (Schrage, 1997) or "vicious attacks" (Dvorak, 1994) and can include swearing, name calling, threats, and insults (Kiesler, Siegel and Mcguire, 1984; Dubrovsky, Kiesler and Sethna, 1991; Weisband, 1992; Kayany, 1998), among others. According to Kayany's (1998) definition, such messages contain both hostility and lack of restrain and can target a person, his/her character, religion, race,

[39] https://www.academia.edu/4389928/Toxic_Language

[40] https://www.perspectiveapi.com/#/

intelligence, and physical or mental ability. Several studies (Lea et al., 1992; Postmes, Spears and Lea, 2000; Spears and Lea, 1992) recognize the importance of social categories as sources of social influence irrespective of the message content. In this context, O'Sullivan and Flanagin (2003) define flames as "*intentional (whether successful or unsuccessful) negative violations of (negotiated, evolving, and situated) interactional norms*". The topic of discussion, the familiarity with the group members, and the confidence in the provision of anonymity have been found among the conditions supporting it (Kerlinger, 1986; Pinsonneault and Heppel, 1998). Possible causality factors of flaming incidences include demographic e.g. men tend to flame more than women (Aiken and Waller, 2000), psychological and behavioral variables e.g. hostility (Reinig, Briggs and Nunamaker, 1998), disinhibition seeking, anxiety, and assertiveness (Alonzo and Aiken, 2004).

Flaming has been found to reduce productivity in the work place (Reinig, Briggs and Nunamaker, 1998) and also to contribute to loss of business (Cosentino, 1994); it can be defamatory with serious consequences to an organization's products, services, and good-will (Alonzo and Aiken, 2004). Flames posted on discussion groups can disrupt the well-being of online communities; a major issue arises when users influence each other's communication behavior (Papacharissi, 2004; Anderson et al., 2013), since a single user's incivility may be sufficient to initiate a "flame war" that is a major user-on-user group-conflict within a community (Voggeser, Singh, and Göritz, 2018). Users group into factions with strong opinions on polarized topics and attack the other faction(s) with violent language (Johnson, Cooper and Chin, 2009). Similarly, a "shit storm" may occur when a large group of people voice their discontent with one entity (a single person or any form of organization) using different social media platforms, which along with "flame wars" can derail societal and political discourse and hinder consensus finding (Voggeser, Singh, and Göritz, 2018).

- **Cyberbullying.** Online bullying is defined as the deliberate use of electronic communication tools through which harm or disturbance is intentionally and repeatedly delivered, targeting a specific individual or group of individuals (Patchin and Hinduja, 2006; Ang and Goh, 2010). This type of VA is rooted in somewhat undetermined social and communicative norms linked to social media, since the comments posted online are often ambiguous and may be interpreted either as humorous or as hostile (Livingstone and Smith, 2014). Relationship problems like break-ups, envy, intolerance, and ganging up (Hoff and Mitchell, 2009) and victimisation (Bauman, 2010; Walrave and Heirman, 2011) are among

the causes of this type of aggressive behavior. According to Willard (2005) cyberbullying occurs in the form of flaming (angry or rude messages), harassment (recurring offensive messages), denigration (harmful, false, cruel statements), masquerade (pretending to be someone else to make that person look bad), outing (sharing others' private information), cyberstalking (threats of harm or intimidation), trickery (tricks to solicit embarrassing information), and exclusion (intentional exclusion from an online group). A more detailed list of cyberbullying aggression avenues is available by Notar, Padgett and Roden (2013). Studies examining the content of aggressive messages used by adolescents in cyberbullying situations (Rachoene and Oyedemi, 2015; Simão et al., 2018) identified VA associated with attacks on intelligence and physical appearance, insults, threats, and outing.

Cyberbullying constitutes an increasingly serious public mental health problem with devastating consequences for the victims (Srabstein, Berkman and Pyntikova, 2008; Ybarra and Mitchell, 2004) ranging from withdrawal from school activities, school absence, and school failure, to eating disorders, substance abuse, depression, and even suicide (Chibbaro, 2007; Klomek, Brunstein and Gould, 2011). Furthermore, it undermines the freedom of youth to use and explore online resources (Hinduja and Patchin, 2009). There are also long-term implications for bullies (Notar, Padgett and Roden, 2013) such as exhibiting typically higher levels of antisocial, violent and/or criminal behavior in adulthood (Patchin and Hinduja, 2012; Kulig et al., 2008). Thus, identifying the content that is used by adolescents online and understanding its effects on these individuals is crucial because these impacts, both for the bully victim and the bully, create ongoing social and economic costs for the community (Notar, Padgett and Roden, 2013) and the social and cultural implications may be very destructive (Rachoene and Oyedemi, 2015).

- **Hate speech.** Hate speech is a general term covering a broad spectrum of extremely negative discourse stretching from hatred and incitement to hatred, to abusive expression and vilification, as well as to extreme forms of prejudice and bias (Jacobs and Potter, 1998). Several definitions have been proposed covering different approaches and perspectives e.g. "*bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics*" (Almagor, 2011) or "*any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic*" (Nockleby, 2000). According to the Council of Europe, hate

speech covers all forms of expressions, which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin (Council of Europe, 1997). Hate speech can target on different bases; depending on the approach these bases may vary somehow e.g. religion, race, gender, and disability (Seglow, 2016), religion, disability, social status, politics, race, sex and gender issues, plus others (Del Vigna et al., 2017), homophobic, racist, sexist, anti-Semitic and against disability (Musto et al., 2016).

Hate speech on disability refers to the incitement against the physical and mental conditions of a person; it is more common for disables than the able-bodied people and is mainly due to the perception of disability by the violator rather than the actual disability of a person (Hollomotz, 2013). Gendered hate speech also known as sexist hate speech is a kind of social shaming which intends to disrespect women, introduce fear and insecurity among women in the society (Chetty and Alathur, 2018). This kind of harassment affects personal lives and professional careers of women (Simons, 2015). Racist hate speech refers to expressions towards the appearance (e.g. skin color) of a person or group. The frequency and impact of this type depends on the intention and perception of the government of a particular nation and varies from one leadership to another leadership (Chetty and Alathur, 2018), given that racism as a system involves "*cultural messages and institutional policies and practices as well as the beliefs and actions of individuals*" (Tatum, 2011). Social networks have a significant role in racism; while racist behavior is primarily motivated by bias, social media have the potential to foster such bias; for instance, some psychologists suggest that social media can deepen users' preexisting biases (Bozdag, 2013). Religious hate speech refers to expressions of hatred against religions (e.g. Islam, Hindu, Christian). Terrorist attacks (e.g. Paris, Tunisia, and Woolwich) trigger increased islamophobic attacks; online anti-Muslim abuse is expressed through negative attitudes, stereotypes, bullying, discrimination, harassment and treats indicating offline violence (Chetty and Alathur, 2018). Muslims are used as a model to depict homogeneous out-group which is involved in conflict, violence and extremism (Törnberg and Törnberg, 2016).

The current surge in online hate speech in Europe has been linked to the ongoing refugee crisis (Ross et al., 2016). Hateful content has become a major problem for online platforms that host user-generated content, since it can alienate users and it can also support radicalization and incite violence (Allen, 2013). Hate speech

poses threat to the dignity of individuals, to personal liberties, to the social fabric of democracies (Waltman and Haas, 2010) as well as to the security of societies; it contributes to a general climate of intolerance which in turn makes attacks more probable against those given individuals or groups, and may also give rise to wider scale conflict and violence. Platform operators and lawmakers are increasingly aware of the problem and are developing approaches to deal with it. In 2016 the European Commission and four major social media platforms (Facebook, Twitter, YouTube and Microsoft) announced a Code of Conduct on countering illegal online hate speech. It included a series of commitments by internet companies to combat the spread of such content in Europe by promising to remove illegal messages within 24 hours after they are reported, making users aware about what is banned by each company, and training staff to let them better spot and respond to online hate speech.

To sum up, VA can manifold in a multitude of ways in different contexts with somewhat different intentions and various effects on individuals, communities and social cohesion. While VA predated the Internet, the extent and the nature of online communication tools amplifies incidents of aggression affecting billions of people. In addition to the severe effects of the verbal attacks per se, there is also a norm of reciprocity that operates for VA; in other words, "*VA begets the same*" (Infante et al., 1990). Furthermore, several studies have directly examined VA as a catalyst toward physical violence (Goldstein and Rosenbaum, 1985). For example, Infante et al. (1990) suggest that a latent hostile disposition, combined with an argumentative skill deficiency, makes VA particularly instigative of violence since little else is available for defense of self. Online VA may also escalate to physical violence; for example, hateful language delivered in the media resulted in massive violence in Kenya before and after the elections in 2007 and 2008 (Benesch, 2018). In some cases aggressive content may even constitute illegal content. A typical example is hate speech which challenges the limits of free speech; it may be considered synonym to hate crime and is handled by different regulations in different countries. An overview of the legal frameworks on hate speech is available in (Chetty and Alathur, 2018). Hence, the detection of online user-generated aggressive content has become a task of critical importance; an overview of the up-to-date approaches, methods and techniques employed to tackle online aggressive content is provided below in section 3.1.1.2.

3.1.1.2 RELATED WORK AND CHALLENGES

A variety of classification methods and algorithms have been used for the detection of online aggressive content. The majority of the related studies adopt supervised learning approaches using *SVMs* (Warner and Hirschberg, 2012; Chen et al. 2012; Xiang et al., 2012; Badjatiya et al., 2017; Davidson et al. 2017; Del Vigna et al., 2017; Jha and Mamidi, 2017; Saleem et al., 2017), *Naïve Bayes* (Razavi et al., 2010; Kwok and Wang, 2013; Bourgonje et al., 2017; Davidson et al. 2017; Saleem et al., 2017), *Logistic Regression* (Xiang et al., 2012; Waseem and Hovy, 2016; Badjatiya et al., 2017; Bourgonje et al., 2017; Davidson et al. 2017; Saleem et al., 2017), *Random Forest* (Xiang et al., 2012; Davidson et al., 2017) or *Decision Trees* (Davidson et al., 2017; Bourgonje et al., 2017). Some works employ also semi-supervised methods using *bootstrapping* to generate automatically lexical resources or additional data (Xiang et al., 2012; Gitari et al., 2015; Waseem and Hovy, 2016). Finally, there are also *deep learning approaches* (Badjatiya et al., 2017; Del Vigna et al., 2017). For example, Del Vigna et al. (2017) use both *SVM* and *LSTM* classifiers and they achieve slightly better accuracy with SMVs (80.6%) as compared to LSTM (79.81%).

A variety of text representation methods and features have been also used to improve classification performance including simple surface features such as *Bag-of-Words* (Kwok and Wang, 2013; Badjatiya et al., 2017; Bourgonje et al., 2017) and *n-grams* (Kwok and Wang, 2013; Nobata et al., 2016; Badjatiya et al., 2017; Davidson et al., 2017; Del Vigna et al., 2017), word clustering such as *Brown clustering* (Warner and Hirschberg, 2012; Malmasi and Zampieri, 2018), and *LDA* (Saleem et al., 2017; Xiang et al., 2012), distributed representations (based on neural networks) i.e. *word embeddings* (Badjatiya et al., 2017; Del Vigna et al., 2017) and *paragraph embeddings* (Warner and Hirschberg, 2012; Nobata et al., 2016), syntactic features such as *Part-Of-Speech tagging* (Xu and Zhu, 2010; Nobata et al., 2016; Silva et al., 2016; Waseem and Hovy, 2016; Davidson et al., 2017; Del Vigna et al., 2017; Saleem et al., 2017) or *typed dependency relationships* (Xu and Zhu, 2010; Saleem et al., 2017), *lexicon-based* features (Xu and Zhu, 2010; Razavi et al., 2010; Xiang et al., 2012; Gitari et al., 2015; Silva et al., 2016; Davidson et al., 2017; Del Vigna et al., 2017), and *SA-based* features (Gitari et al., 2015; Davidson et al., 2017; Del Vigna et al., 2017). Other types of features include also *token features* like capitalization, non-alpha numeric characters present in tokens, punctuation information as well as information about URL mentions (Chen et al., 2012; Nobata et al., 2016).

Unigrams and larger n-grams are often reported to be highly predictive (Schmidt and Wiegand, 2017), however as it is reported (Davidson et al., 2017) bag-of-words

approaches tend to have high recall but lead to high rates of false positives since the presence of offensive words can lead to the misclassification of tweets as hate speech (Kwok and Wang 2013; Burnap and Williams, 2015). In many studies n-gram features are combined with a large selection of other features. Linguistic aspects also play an important role; for example, some studies (Burnap and Williams, 2015) report significant performance improvement using dependency relations. Lexical resources (e.g. lists of offensive, hateful, etc. words) constitute also an important resource for the detection of online aggressive content; despite their effectiveness, as it is often reported (Nobata et al., 2016) they are insufficient as stand-alone features, since contextual factors should also be taken into consideration. SA, in particular negative polarity detection is also used often as an auxiliary classification (Sood, Antin and Churchill, 2012; Gitari et al., 2015; Malmasi and Zampieri, 2018); for example, Malmasi and Zampieri (2018) report that the inclusion of sentiment features helps in detecting profanity and hate speech. The use of word embeddings that have been successfully applied to a variety of NLP tasks, including SA (e.g. Tang et al., 2016), is reported to have limited effectiveness (Nobata et al., 2016) as compared to paragraph embeddings which are internally based on word embeddings (Djuric et al., 2015). Gupta and Waseem (2018) report that domain-agnostic word-embeddings perform slightly worse as compared to domain-specific, though domain-specific are apt at dealing with class embeddings, while Hasanuzzaman, Dias and Way (2017) report that word embeddings that incorporate demographic (Age, Gender, and Location) information significantly improve over the classification performance of demographic-agnostic models.

Other studies report also that non-linguistic features like the gender of the author can help improve classification (Dinakar et al., 2012; Waseem and Hovy 2016). Given that what is considered aggressive/offensive/hate/etc. speech can be highly dependent on world knowledge, it is expected that the detection of such complex phenomena might benefit from including information on non-language related aspects. In this context, some approaches exploit also demographic and geographic information (Waseem and Hovy 2016) or behavioral characteristics (Pitsilis, Ramampiaro and Langseth, 2018) to boost performance, while Dinakar et al. (2012) employ automatic reasoning over world knowledge focusing on anti-LGBT hate speech. A more detailed overview of the features used is available by Schmidt and Wiegand (2017).

However, it is difficult to draw safe conclusions about the effectiveness of the different methods, techniques and features, since the results are not directly comparable. Each study employs *different definitions* and addresses somewhat *different aspects of online VA*; *flames* (e.g. Razavi et al., 2010), *profanity-related*

offensive content (Xiang et al., 2012; Sood, Antin and Churchill, 2012), *cyberbullying* (Dinakar et al. 2012; Dadvar et al., 2014; Hee et al., 2015; Hee et al., 2018), *hate speech* (e.g. Warner and Hirschberg, 2012, Kwok and Wang, 2013; Gitari et al. 2015), or *abusive language in general* (Chen et al., 2012) in different social media and on line communities e.g. *Twitter* (Kwok and Wang, 2013; Waseem and Hovy, 2016; Davidson et al., 2017), *Facebook* (Ben-David and Matamoros-Fernandez, 2016), *YouTube* (Chen et al. 2012), Yahoo! (Sood, Antin and Churchill, 2012; Djuric et al., 2015; Nobata et al., 2016), and *Reddit* (Saleem et al., 2017), among others.

For example, Chen et al. (2012) employ a supervised classification approach using SVMs with features including n-grams, automatically derived blacklists, manually developed regular expressions and dependency parse features to detect offensive language in YouTube comments aiming at shielding adolescents. Their method achieves a performance of 98.24% and 94.34% in terms of precision and recall, respectively, on the task of inflammatory sentence detection. Even though –as they point out- they do not have a strict definition of offensive language in mind, their tool can be tuned by the use of a threshold which can be set by parents or teachers so online material can be filtered out before it appears on a web browser. On the other hand, Sood, Antin and Churchill (2012) work on detecting (personal) insults, profanity and user posts that are characterized by malicious intent. They employ a list-based system that achieves an F-measure score of 45.7% and they discuss the limitations of such approaches arguing that profanity detection is not a simple task since it requires also domain, community and context specific knowledge. Xu et al. (2012) further look into jokingly formulated teasing in Twitter messages that represent (possibly less severe) bullying episodes. They experimented with four text classifiers (Naive Bayes, SVM(linear), SVM(RBF) and Logistic Regression) and explored combinations of n-gram features with POS-information-enriched tokens achieving the best accuracy (81.6%) with SVM linear kernel. In another line of online aggression research, Burnap and Williams (2015) focus on hate speech and in particular on *othering language*, characterized by an us-them dichotomy in racist communication. They combined n-gram hateful terms features and typed dependencies and experimented with three classifiers (Bayesian Logistic Regression, SVMs, and a Random Forest Decision Tree rule based classifier), while they also implemented an ensemble classifier where a combination of all three was used to make a final classification decision achieving an overall F-measure of 95% in the detection of offensive or antagonistic Tweets in terms of race ethnicity or religion.

This diversity and lack of consensus in terminology is also apparent in the related *shared tasks* that have been organized recently. For example, the TRAC 2018[41] shared task on aggression detection distinguishes three levels of textual aggression: *overtly aggressive* (an expression of aggression directly with specific words or keywords), *covertly aggressive* (subtly aggression such as indirect attack or with more polite expressions) and *non-aggressive*. The GermEval [42] 2018 shared task on the identification of offensive language included two levels of Tweets classification: binary (*offensive - other*), and fine-grained (*profanity, insult, abuse, other*), while the Kaggle [43] challenge on online toxic comment classification provides a six-way schema: *toxic, severe toxic, obscene, threat, insult, and identity hate*.

Focusing on *hate speech* that is more close to the work presented in this thesis, some studies adopt a binary classification schema aiming to distinguish hateful from non-hateful content (Djuric et al., 2015; Burnap and Williams, 2015; Köffer et al. 2018), while other studies attempt to differentiate hate speech and offensive language (Nobata et al., 2016; Davidson et al., 2017; Malmasi and Zampieri, 2018). For example, Nobata et al. (2016) propose a supervised classification method for detecting hate speech in user comments found on Yahoo! Finance and News. They employ a generic definition of hate speech (language which attacks or demeans a group based on race, ethnic origin, religion, disability, gender, age, disability, or sexual orientation/gender identity) and examined it along with other two types of abusive language; derogatory speech (language which attacks an individual or a group, but which is not hate speech) and profanity (language which contains sexual remarks or profanity). Their method combines n-grams, linguistic (e.g. token features and punctuation) and syntactic features, and distributional semantics (three types of embeddings-derived features) and achieves F-scores up to 81% in detecting abusive content. Davidson et al. (2017) distinguish hate speech from other types of offensive language by limiting their definition of hate speech to language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group and they test a variety of models (e.g. logistic regression, SVMs, decision trees, Naive Bayes, Random Forests) to classify Tweets into three categories, namely hate speech, offensive language, or neither.

Another line of research focuses on specific types/categories of hate speech. For example, Warner and Hirschberg (2012) approach hate speech as offensive language that makes use of stereotypes to express an ideology of hate focusing on hateful

[41] https://sites.google.com/view/trac1/shared-task

[42] https://projects.fzai.h-da.de/iggsa/

[43] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

language directed towards a minority or disadvantaged group and more specifically on anti-Semitic hate rather than abusive/offensive language in general; they designed an anti-Semitic/not anti-Semitic axis to annotate a corpus of websites (identified as offensive by the American Jewish Congress) and user comments (Yahoo! news group posts that readers had found offensive). They report an F-score of 63% using a supervised classification method by first targeting certain words that could either be hateful or not using a template-based strategy presented in (Yarowsky, 1994), and then using Word Sense Disambiguation techniques to determine the polarity of the word. Waseem and Hovy (2016) propose a character n-gram based approach that exploits also demographic (e.g. gender) and geographic information to detect hateful/offensive Tweets in terms of racist and sexist slurs based in critical race theory reaching F-scores up to 74%. Pitsilis, Ramampiaro and Langseth (2018) employ a recurrent neural network-based approach composed of multiple Long-Short-Term-Memory (LSTM) based classifiers, and utilize user behavioral characteristics such as the tendency towards racism or sexism to boost performance on the dataset of Waseem and Hovy (2016).

As Davidson et al. (2017) point out, some approaches conflate hate speech with offensive language making it difficult to ascertain the extent to which they are really identifying hate speech. Furthermore, the lack of consensus and the overlap between subtasks –which is apparent in the variety of the labels used- resulted in contradictory annotation guidelines. For example, as Waseem et al. (2017) observe, some messages considered as hate speech by Waseem and Hovy (2016) are only considered derogatory and offensive by Nobata et al. (2016) and Davidson et al. (2017). Similarly, Hee et al. (2015) identify discriminative remarks (racist, sexist) as a subset of "insults" in the context of cyberbullying events, whereas Nobata et al. (2016) classify similar remarks as "hate speech" or "derogatory language". In this context, Waseem et al. (2017) propose a two-fold typology of abusive language that synthesizes these different subtasks considering whether (i) the abuse is directed at a specific target (distinguishing between abuse directed at individuals or online communities such as cyberbullying, and abusive expressions towards generalized groups such as racial categories and sexual orientations), and (ii) the degree to which it is explicit (distinguishing between unambiguous explicit abusive language such as racial and homophobic slurs, and implicit language that is characterized by the use of ambiguous terms, sarcasm, lack of profanity or hateful terms, and other indirect means).

Other studies put effort on notions of hate speech that can be operationalized. For example, Saleem et al. (2017) propose a community-driven model of hateful speech

referring to speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity. This approach is based on a deep sociological literature that acknowledges that communities both form, and are formed by, coherent linguistic practices (Bucholtz and Hall, 2005). It is tested on Reddit which the authors consider an attractive testbed for work on hateful speech both because the community spaces are well-defined (i.e., they have names, complete histories of threaded discussions) and because it has been a major online home for both hateful speech communities and supporters for their target groups. In another line of research, the recent work of Sanguinetti et al. (2018) deals with the creation of an Italian Twitter corpus annotated for hate speech against immigrants. Besides hate speech, the proposed annotation scheme includes also the categories aggressiveness, offensiveness, irony, stereotype, and intensity as different factors that can contribute to the definition of a hate speech notion.

To sum up, the related literature review findings suggest that the detection of online aggressive content is not a trivial task; verbal attacks are shaped differently depending on individuals' intentions and strategic choices in language use ranging from expressions of negative emotions, name calling, swearing, threatening, and insulting to the use of humor, sarcasm and irony or even paralanguage (e.g. capitalization, punctuation, emoticons) (Tereszkiewicz, 2012). Profanity and physical threats are perceived as more aggressive than criticism (Greenberg, 1976). However, depending on the social context the intensity of aggression may vary. For example, swearing may indicate high (Infante et al., 1992) or medium aggression (Laineste, 2012). In addition, in some cases VA can only be interpreted in a specific situational context, whilst aggressive language is not always abusive. For example, profanity that is key indicator of flaming can also be a marker of relationship closeness between friends; hence as O'Sullivan and Flanagin (2003) point out "*the determination of whether a message is considered a flame is often based upon an outside observer's perspective – that of the commentator, the researcher, or a coder*". Several studies observed considerable percentage of disagreement between human expert annotators having the same definition of flames (Razavi et al., 2010). Similarly, work on online hate speech has revealed that identifying hate speech consistently is difficult and often yields the paradox that each person seems to have an intuition for what hate speech is, but rarely are two people's understandings the same (Saleem et al., 2017). For example, the study of Kwok and Wang (2013) demonstrated a mere 33% agreement between human annotators from different races tasked with labelling racist tweets.

Hence, given the above challenges, the lack of consensus in terminology and the overlap between several subtasks there is a need for clear and operational definitions

and focused frameworks for the examination of VA depending on each case. Furthermore, the data source also matters when examining aggression since different online contexts (e.g. social media, hate-promoting websites, etc.) do differ in the level of aggression (Laineste, 2012). Sood, Antin and Churchill (2012) have also found differing amounts of profane language, insults, and directed insults (insults directed at other members of the community) in different communities of Yahoo! Buzz. A final, yet important, observation is that detecting and classifying an aggressive message is not enough; the target of the attack as well as the attacker constitute crucial elements of a verbal attack. For example, as Chetty and Alathur (2018) point out, an effect of hate speech depends on the originator, the content and the targeted one. However, only few studies incorporate these elements in their research; Sood, Antin and Churchill (2012) aside from classifying insulting messages also predict whether such messages are directed at an author of a previous comment or at a third party, Xu et al. (2012) look at the entire bullying event and assign roles to actors involved in the event, while Silva et al. (2016) provide an analysis of the main hate target groups on Twitter and Whisper concluding that in both platforms people are mostly bullied for their ethnicity, behavior, physical characteristics, sexual orientation, and class or gender.

3.1.2 XENOPHOBIA

3.1.1.2 KEY CONCEPTS AND RELATION TO VA

Xenophobia is typically described as a universal phenomenon and broadly defined as intense dislike, hatred or fear of those perceived to be strangers (Crush, 1996; Master and Roy, 2000). According to the definition employed by the International Organization for Migration and Office for High Commissioner for Human Rights "*Xenophobia describes attitudes, prejudices and behavior that reject, exclude and often vilify persons, based on the perception that they are outsiders or foreigners to the community, society or national identity*" (International Labour Organization 2001)". The underlying assumption in this category of definitions is that xenophobia denotes behavior specifically based on the perception that the "Other/stranger" is foreign, or originates from outside the nation or the (imagined) community. This is more explicit in the definition established by the South African Human Rights Commission (SAHRC) that defines xenophobia as "*the deep dislike of non-nationals by nationals of a recipient state*" (Bekker and Carlton, 2010).

The conceptualization of xenophobia is often based on the semantics of the term itself involving two Greek words, "xenos" and "phobos" meaning "foreigner/stranger" and

"fear", respectively (Makgopa, 2013). Hence, it is often defined as "*distrust, unreasonable fear, or hatred of strangers, foreigners, or anything perceived as foreign*" (Yakushko, 2009). Sentiment constitutes an important aspect in the conceptualization of xenophobia, since emotional involvement is crucial in the understanding of the phenomenon. In particular, xenophobia as a "*psychological state of hostility or fear towards outsiders*" (Reynolds and Vine, 1987) is associated with feelings of dominance (implying superiority) or vulnerability (implying the perception of threat), respectively (van der Veer, 2013). As a disposition, xenophobia can be the basis of racism, fascism, and nationalism (Delanty and O'Mahony, 2002), since it is often rooted in (cultural, religious, racial, etc.) prejudices or driven by ideology. For example, with the rise of the extreme right ideologies in Greece and Europe during the past years a spike in anti-immigrant violence has been witnessed[44].

The relationship between xenophobia and racism constitutes a distinct chapter in the literature; even though they are considered two distinct phenomena, the limits between racism and xenophobia are not always easy to delineate; for example Tafira (2011) proposes the deconstruction of the term xenophobia and seeing it as culturally-based racism that is "*heavily entrenched in cultural differences enunciated by dissimilarities in nationality, ethnicity, language, dress, customs, social and territorial origins, speech patterns and accents",* given the conceptual and definitional limitations of the term xenophobia in describing the complex social realities occurring in South African black communities. Other scholars point out that it is important to bear in mind that xenophobia is "*the basis for both overt racist actions and more subtle forms of exclusion hidden in the discourse of society*" (Hjerm, 1998).

In this line of research and focusing mainly on the effects and the consequences of xenophobia in social life -rather than its conceptual formulation- Delanty and O'Mahony (2002) describe xenophobia as "*rooted in the symbolic violence of everyday life*", while Bronwyn (2002) suggests that xenophobia is more than just an attitude towards foreigners; it can also take shape as a practice, and in particular as a violent practice. Interdisciplinary empirical studies on the roots of xenophobia and aggression (Wahl, 2002) suggest that some patterns of behavior and extreme emotions preceding later xenophobic violence (e.g. anger, hate, hyperactivity, aggression, anxiousness, fear, grief) can be detected early in childhood. In the field of psychological and behavioral studies (Holloway, 1974; Marler, 1976) the term

[44] Human Rights Watch reports that racist and xenophobic violence in Greece has reached "alarming proportions, with gangs regularly attacking migrants and asylum seekers" (http://www.hrw.org/europecentral-asia/greece)

"*aggressive xenophobia*" is used to denote aggressive acts and postures, attributed to the fear of the "unfamiliar" or the "stranger". In the field of political and social sciences "*violent xenophobia*" is considered a regular feature xenophobic expression especially in South Africa (Crush and Ramachandran, 2014) that has a unique brand of xenophobia and xenophobic mobilization linked to its specific apartheid history. For example, Hågensen (2014) combined key informant interviews with relevant organizations, analysis of published work (reports and an article) and xenophobia literature review findings in order to explore the causes and nature of xenophobia in South Africa focusing on the case of De Doorns -where 3000 Zimbabweans were chased out of their homes in November 2009, which were subsequently looted and destroyed- as an example of a xenophobic incident that went beyond xenophobic attitudes to manifest in violent behavior towards African migrants.

However, xenophobia is a phenomenon of global dimensions; numerous studies (e.g. the European Social Survey, 2002 – 2003; Coenders, Lubbers and Scheepers, 2004; The Eurobarometer Survey, 2000; Crush and Ramachandran, 2009; Geschiere, 2009) in different countries document increasing intolerance, xenophobia, ethnic exclusionism and opposition to immigration and diversity. The United Nations High Commissioner for Refugees (UNHCR) increasingly recognizes that xenophobia's various manifestations represent protection threats to its Persons of Concern (PoC): refugees, stateless persons, asylum seekers and internally displaced persons.

As far as the actors are concerned, according to Misago, Freemantle and Landau (2015): "*If we consider violence as overt, explicit and due to conscious complicity, almost everyone is guilty of xenophobic behavior: ordinary residents, community leaders, public servants, political officials, bureaucrats and law enforcement agents. Government officials and political leaders often make xenophobic pronouncements that shape or reinforce public opinion and behavior; public servants deny 'outsiders' access to services they are entitled to; law enforcement agents are particularly known for extortion, harassment, arbitrary detention and selective enforcement of the laws while 'members of the public' often engage in, or condone, collective violence against foreigners*". As with other forms of collective violence, xenophobic violence manifestations include murder, assaults causing bodily harm, looting and vandalism, robbery, arson attacks, burning of property, immolation, displacement, intimidation and threats, eviction notices, etc. Furthermore, as it is pointed out by Gerring (2009), in some circumstances, intimidation and the threat of violence cause substantial socio-economic damage and are thus of the same order as overt physical attacks.

In this context, and given the background on VA discussed in the previous section, it can be argued that online VA constitutes an important component in the study of xenophobia; aggressive messages targeting foreigners can be indicative of xenophobic attitudes. For instance, VA involves using messages to attack other people or those aspects of their lives that are extensions of their identity (Hamilton and Hample, 2011) and the forms of aggression are manifold and vary from expressions of disgust and contempt, to threats, slander, insults, and hatred (Rösner and Krämer, 2016). For example, in an attempt to map xenophobia on the Estonian Internet by describing the use of verbal aggression directed against some more common groups in Estonia, Laineste (2012) describes the main objects of online flaming and the social and contextual background of the target choice. The close relation of online VA with xenophobia is also demonstrated by the hate speech literature and especially by approaches that focus on xenophobia-related types of hate speech like racist (Kwok and Wang, 2013; Waseem and Hovy, 2016) and hate speech directed to immigrants (Sanguinetti et al., 2018) or specific ethnic groups (Warner and Hirschberg, 2012), even though they do not make an explicit reference to xenophobia.

As it has been already discussed in previous sections, traditionally, xenophobia is examined using empirical and statistical methods; xenophobic attitudes are being measured using data coming from focus groups, interviews, and public sentiment polls using standard questions in order to capture opinions, emotions, perceptions and beliefs (e.g. Eurobarometer). However, despite the numerous research efforts in automatically detecting and analyzing online sentiment, VA and hate speech, the user-generated content has been scarcely explored from the xenophobia monitoring and measuring standpoint in a large scale by making use of use of computational social science approaches and big data analytics. Such an effort is the recent UNHCR project (UN Global Pulse, 2017) to build xenophobia monitors and situation awareness monitors in order to enable the United Nations Refugee Agency staff to routinely monitor and analyze relevant social media feeds in six different languages: Arabic, Farsi, English, Greek, German and French. The project used three categories to classify geo-referenced Tweets in terms of xenophobic content: a) *Xenophobic* for tweets that express negative attitude, prejudice, or hostile sentiment that vilifies PoC, b) *Non-Xenophobic* for tweets that express explicit support, positive attitude, or friendly sentiment towards PoC, c) *Neutral* for tweets that describe facts about PoC (for example, news articles) but that do not express a strong sentiment or any sentiment at all, and d) *Irrelevant* for tweets that are not related to PoC. The situation awareness monitors intended to gauge responses to the terrorist attacks, and how these might be related to PoC in the global Twittersphere using the following categories: 1) *Blame*: tweets that explicitly blame PoC for the incident, 2) *Don't Blame*: tweets that

advocate for not blaming PoC for the incident, or at least that attempt to deattach them, 3) *No reference to PoC*: tweets that describe facts about the incident, but that do not mention PoC, and 4) *Irrelevant*: tweets that mention PoC, but that are not related to the incident, 5) *Off-topic*: tweets that are neither related to PoC, or the incident.

The monitors rely on combinations of keywords based on the use of logical operators (AND/OR/NOT); the same vocabulary was employed for each language to enable a relative degree of comparison between monitors. For example:

Xenophobia English ((*migrant* OR *refugee* OR *refugees* OR *immigrants*) AND (*Greece* OR *Greeks* OR *fear* OR *hatred* OR *racism* OR *xenophobia* OR *foreigners* OR *arrivals* OR *Syrians*))

Xenophobia Greek ((*μετανάστης* OR *πρόσφυγας* OR *πρόσφυγες* OR *μετανάστες*) AND (*φόβος* OR *μίσος* OR *ρατσισμός* OR *ξενοφοβία* OR *ξένοι* OR *αφίξεις* OR *Σύριοι*))

Even though the project does not explicitly refer to VA towards PoC, the negative perceptions are examined in terms of racists, extremist or xenophobic comments from host communities in their native language, negative sentiment and feelings towards refugees and migrants. Focusing on the Greek monitor, only 5% of the tweets retrieved by the Xenophobia Greek monitor were classified as xenophobic as compared to 15% in the Xenophobia English monitor. According to the project report, although the monitors retrieved a larger number of posts in Greek, with the sample retrieved, there were more xenophobic posts in English than in Greek for this particular geographic location.

To the best of our knowledge, the only up-to-date research effort that examined xenophobia as a violent practice using computational social science and big data techniques is the XENO@GR project that aimed to examine xenophobia over time in Greece (see below section 3.1.1.2).

3.1.1.2 XENOPHOBIA IN GREECE AND THE XENO@GR PROJECT

Early studies carried out at a European level aiming to provide an overview of the phenomenon in the EU member states reported tolerant, xenophilic and generally free of racial prejudice attitudes of Greeks towards ethnic or religious minorities stressing also out that the Greek Jewish community had "generally not suffered" from Greek anti-Semitism and that the country's historical legacy discouraged the rise of right wing extremism of the revival of fascist/nazist movements (Evrigenis Report, 1985). In a special Eurobarometer survey in 1989 Greeks were included among the most

tolerant and least xenophobic Europeans. Two years later, in 1991 a European Parliament Report stressed as alarming the activities of certain marginal extreme right groups (United Nationalist Movement/ENEK, National Political Union/EPEN) that participated in activities against Jews, Roma and Muslims; however, xenophobic attitudes were reported to be limited in Greek society given the small number of foreigners residing in Greece at the time and targeted only against the Muslim minority in Thrace. A year later, the Eurobarometer findings drew a very different picture in relation to attitudes towards immigrants; 35% of Greeks opted for the restriction of rights of migrants vs. 14% that were in favor of their extension. In a recent PEW survey (July 2016) that explored the threat perceptions crystallized in Europe in relation to the refugee crisis and to the terrorist attacks that affected a number of European countries, Greek public opinion has been found to perceive the refugees primarily as "*a burden to our country because they take our jobs and social benefits*" and secondarily as a potential link to terrorism or to the rise of criminality.

These findings can be examined in relation to the historical changes and events that affected the country during the last thirty years; the *migration flows to Greece* and the *economic crisis in Greece*. The geopolitical changes that took place after the collapse of socialist regimes in Central Eastern Europe in the post-1989 period resulted in a migration wave which in early and mid-1990s consisted mostly of Albanians, and in second half of the 1990s included also immigrants from Balkan countries and former USSR countries. At that time, Greece's political stability and democratic regime, EC membership and its (relative) economic prosperity were among the factors that made the country an attractive destination for economic immigrants (Triandafyllidou, 2010). A second migration wave took place in the 2000s, when there was an increase in migration flows from Asian countries (Pakistan, India, Bangladesh, Iraq, and Afghanistan to Europe. Finally, the recent ongoing conflicts and violence around the world led over 1.4 million people to seek refuge in Europe between 2015 and the first part of 2017 leading to the (ongoing) refugee crisis; an estimated[45] 362,000 refugees and migrants risked their lives crossing the Mediterranean Sea in 2016, with 181,400 people arriving in Italy and 173,450 in Greece.

Even though the migration flows do not directly imply an association with the rise of xenophobia and as it is suggested in the related literature "*the relationship between immigration and extremism is unclear and complex … so we need to explore how, when and to what effects immigration is translated into a political issue*" (Mudde, 2010), the recent refugee/immigrant crisis in Europe gave burst to anti-immigrant

[45] https://www.unhcr.org/europe-emergency.html

sentiments, attitudes and practices across Europe ranging from individual reactions to official state policies (e.g. closing borders).

Research on xenophobia in Greece was stimulated by the significant increase of migration flows to the country in the 1990s. According to research findings (Voulgaris et al., 1995) the construction of otherness was primarily based on nationality, while religion and language were not considered of primary importance unless associated with national/ethnic difference. Thus, the image of foreigner in Greek society was mainly associated with Balkan, and mainly Albanian, nationality. Other researchers associated the phenomenon of xenophobia with the development of Greek nationalism and Greek national identity and also correlated xenophobic attitudes with negative attitudes towards economic development (Michalopoulou et al., 1998). An analysis of the 138 Special Eurobarometer in the early 00s survey concluded that negative attitudes towards minority groups in Greece are above the EU average; "*the respondents in Greece claim that they are not very willing to accept refugees and that they are afraid of unemployment and insecurity because of these minority groups*".

At this point, it is worth noting that the historical roots of xenophobia constitute an important aspect that should be taken into consideration, since historical events can create a legacy of xenophobia which conditions a society's perceptions and attitudes in future situations. In particular, Baldwin-Edwards (2014) goes back to the Asia Minor Catastrophe and to the mass influx of refugees in the 1920s to the Greek territory arguing that the hostile popular response against this migration wave was "structurally important" for the reception of Balkan (mainly Albanian) migration in the 1990s explaining in addition how the negative attitudes of Greek society, politicians and mass media constructed and reproduced the stereotype of the "dangerous Albanian".

Another line of research associates the rise of xenophobia with the emergence of the economic crisis. The starting point of this period could be placed symbolically in 2009, when the issue of the economic crisis was raised in the public debate, since it "*signifies a turning point for Greek national discourse and also a rupture in terms of national self-image and memory practices*" (Lialiouti and Bithymitris 2017). Several studies explored the current expressions of xenophobia in relation to the migrant flows from Asian and African countries and the resulting problems in urban centers due to the shortages in public policies. For example, Chtouris et al. (2014), building on the literature that links scarcity of resources with the intensity of threat perceptions or with instrumental social reactions against immigrants, suggest that the association

of xenophobia with the economic crisis is confirmed based on findings that correlated high perception of threats with high presence of immigrants or the status of unemployment.

The economic crisis and the bailout agreements that involved painful austerity measures were framed by the government and by the media as a crisis of national sovereignty. Teperoglou and Tsatsanis (2014) argue "*the main results of the crisis was to repoliticize a number of divisive issues that had themselves been simmering in the background such as questions of relations to the European Union, social peace, as well as national identity and immigration*". As popular indignation grew, the various discourses focusing on the role of external enemies in the crisis, such as the IMF, Germany or the EU, flourished in the Greek public sphere. In particular, the popularity of anti-German attitudes is attested by a series of public opinion findings: In a VPRC survey in 2012, for almost one third of respondents Germany denoted "Hitler/3rd Reich/Nazism". In the Pew Global Attitudes Project (2012) 78% of Greeks held an "unfavorable view" of Germany, while 83% felt that "German/EU power over our economy" was a "major threat to the country's economic well-being".

An important aspect of contemporary research on xenophobia in Greece is related to the study the neo-NAZI party of Golden Dawn whose appeal is often interpreted as a tangible proof for the growth of xenophobia. For example, Ellinas (2013) emphasizes the emergence of anti-system and anti-immigrant sentiments as a result of the economic crisis and GD's ability to capitalize on these sentiments precisely by constructing an anti-system and anti-immigrant political profile for itself, with anti-Semitism, Holocaust denialism and conspiracy theories as important elements of the party's discursive practices. In fact, the role of anti-Semitism in the current Greek political culture has attracted attention after a series of opinion poll findings and most importantly after the rise of Golden Dawn (Georgiadou, 2015). According to the ADL Global 100[46] survey, which elaborated an index of anti-Semitism based on the strength of anti-Semitic stereotypes, Greece was the most anti-Semitic country in Europe scoring 69%.

A detailed overview and review of the state of the art empirical methods and research findings about xenophobia in Greece is available by Georgiadou et al. (2017). With the exception of the recent UNHCR project (UN Global Pulse, 2017) that included also a Greek xenophobic monitor as discussed in the previous section, the only up-to-

[46] http://global100.adl.org/public/ADL-Global-100-Executive-Summary.pdf

date research effort that examined xenophobia in Greece using computational methods and big data analytics was the XENO@GR[47] project.

XENO@GR is the short name of an interdisciplinary project entitled "Examining Xenophobia in Greece during the economic crisis: A Computational perspective". The basic aim of this research effort was to examine the evolution of the phenomenon of xenophobia in the contemporary Greek society from the 1990s onward focusing on whether (or not) the phenomenon of xenophobia is an outcome of the recent financial crisis or it comprises a long-lasting social perception deeply rooted in the Greek society. The research hypothesis was that, given the common perception that xenophobia is a deeply-rooted social phenomenon that reasonably escalates under circumstances of severe economic crisis, xenophobia should have been raised in Greece after the outburst of the economic crisis in 2009. The research goal was to formulate adequate responses to the following Research Questions (RQs):

- How have the prejudices and stereotypes about the 'other' been shaped in a historical perspective in Greece taking as a reference point the 1990s when there was a substantial wave of xenophobic tensions against immigrants in Greece?

- How have the economic crisis, spread in Greece from 2009 onwards, affected this sort of xenophobic attitudes and beliefs?

- Does the effect of the economic crisis comprise the basic factor of the rise (or fall) of xenophobic sentiments among Greeks or can we support the hypothesis that this phenomenon has deep roots in the Greek society and the economic crisis has negligible or minor impact on the way Greeks behave against "others" and/or immigrants?

The notion of "other(s)" was limited to specific Target Groups (TGs) of interest, which were defined based on a number of criteria (e.g. population of the specific ethnic groups in Greece, dominant prejudices in Greece about the specific groups). In particular, the project focused on the following 10 TGs: TG1: PAKISTANI, TG2: ALBANIANS, TG3: ROMANIANS, TG4: SYRIANS, TG5: MUSLIMS/ISLAM, TG6: JEWS, TG7: GERMANS, TG8: ROMA, TG9: IMMIGRANTS, TG0: REFUGEES.

IMMIGRANTS and REFUGEES are considered two generic TGs and examined separately due to the different connotations and implications of these two lexicalizations; the research hypothesis was that people framed as "*immigrants*" are more likely to

[47] http://xenophobia.ilsp.gr/?lang=en

receive xenophobic behaviors rather than those framed as "*refugees*". In addition, there are legal protection differences between *immigrants* and *refugees*; refugees are specifically defined and protected by international law, particularly regarding refoulement[48].

In order to achieve the above goals, the project focused on the violence aspect of xenophobia and performed a large-scale multi-source study based on the use of advanced computational social science approaches. Looking beyond traditional empirical approaches of social science research, the project aimed at analyzing and providing an in-depth understanding of the evolution of the phenomenon of xenophobia as a violent practice in the Greek society drawing on social computational methods and big data analytics. In particular, two principal data analytics workflows were employed:

- **Event Analysis** using news data aiming to capture physical attacks (e.g. violent attacks, sexual attacks, attacks against properties) against the predefined TGs of interest.

- **Sentiment Analysis** using Twitter data aiming to detect verbal attacks targeting the predefined TGs of interest.

The work presented in this thesis constitutes the fine-grained SA framework that was designed and implemented in the context of this project. In particular, the research activity was directed towards a data-driven and linguistically-inspired conceptual and computational framework for the analysis of VA expressed against the predefined TGs of interest aiming to address the following three RQs focusing on the amount, the type and the content of the verbal attacks, respectively:

- RQ1: Who are the main targets of Twitter verbal attacks?
- RQ2: Which are the main types of Twitter verbal attacks?
- RQ3: Are there stereotypes and prejudices against foreigners rooted deeply in the Greek society?

This notion of VA is closely related to hate speech, however, given the lack of a universally agreed definition as well as the legal implications of the term hate speech, the general term VA is used instead for explicitly stated verbal attacks targeting specific groups of foreigners in Greece. The ultimate goal was to build a KD that

[48] Expel or return a refugee to the territories where her/his life or freedom would be threatened on the account of race, religion, nationality, membership of a particular social group or political opinion. UNHCR (1977).

would help to formulate adequate responses to the above RQs. To this end, a five-step methodology was followed (see below 3.2).

3.2 METHODOLOGY

The overall workflow for building the VA analysis framework included a five-step process presented below in Fig. 45. At the first step, data was gathered related to the predefined TGs of interest (e.g. JEWS, MUSLIMS, ALBANIANS, etc.). At the second step, samples of the collected data were explored in order to identify different aspects of VA related to the predefined targets of interest. Next, based on data observations and literature review findings, a linguistically-driven VA framework was designed according to which the VA messages (VAMs) were classified into distinct categories based on specific criteria (described below in Section 3.2.3). The fourth step included: i) the design and the development of the resources (e.g. lexical resources, linguistic patterns) and the models/algorithms needed for the computational treatment of the VA framework (VA analyzer), and ii) the automatic processing of the data collections with the VA analyzer. At the fifth step, the output was visualized in various ways in order to obtain a better understanding of the data and the results of the VA analysis.

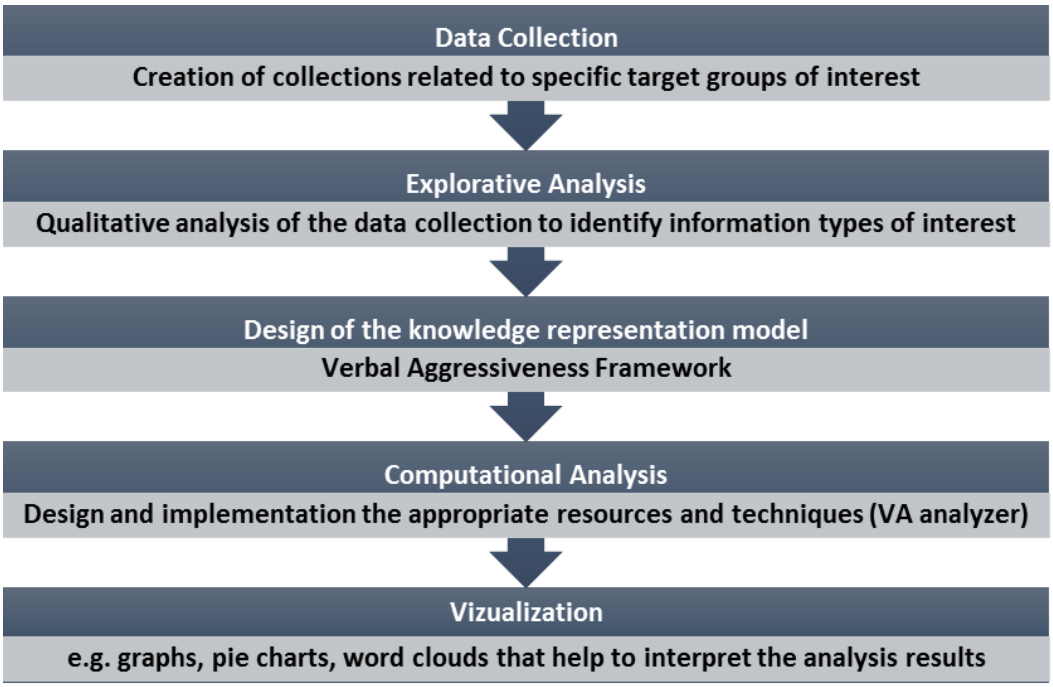


Figure 45: Workflow for building the VA framework

3.2.1 DATA COLLECTION

For each TG of interest relevant Tweets were retrieved from the Twitter data source using related queries/keywords (e.g. "ισλάμ" (="islam"), "Πακιστανός" (="Pakistani"), "Ρουμάνος" (="Romanian"), etc.). Given that the search function in the database configuration is stemmed, the queries returned also tweets containing morphological variations of the selected keywords (e.g. "*ισλαμοποίηση*" for "*ισλάμ*"); the search resulted in 10 collections (1 per TG) containing in total **4.490.572** Tweets (see Table 10) covering the period 2013-2016. The per-year amount of Tweets that were retrieved for each TG is illustrated in Fig. 46.

Target Group (TG)	Number of Tweets
TG1: PAKISTANI	**66.307**
TG2: ALBANIANS	**199.095**
TG3: ROMANIANS	**74.270**
TG4: SYRIANS	**299.350**
TG5: MUSLIMS/ISLAM	**546.880**
TG6: JEWS	**101.262**
TG7: GERMANS	**1.097.597**
TG8: ROMA	**182.974**
TG9: IMMIGRANTS	**672.009**
TG0: REFUGEES	**1.250.828**

Table 10: Data collection per TG

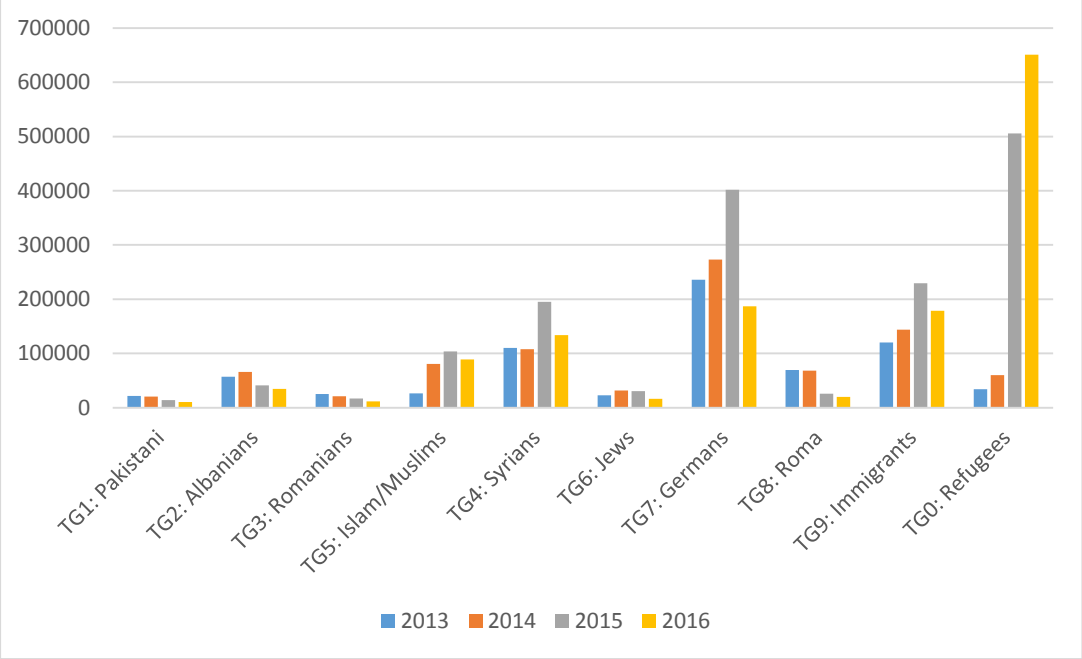


Figure 46: Per-year number of Tweets collected for each TG

The most discussed TGs are the ones of REFUGEES and GERMANS. In particular, the peak in the mentions of refugees during 2015 and 2016 coincides with the refugee crisis in Europe, whilst GERMANS are continuously in the limelight since, along with the IMF and the EU, they have a central role in the Greek crisis. The next most discussed TGs are IMMIGRANTS and SYRIANS, who are also related with the refugee crisis. MUSLIMS/ISLAM follow in the 5th place, with a peak from 2014 onward which coincides with the rise of ISIS. However the number of Twitter mentions is not necessarily indicative of the amount of the verbal attacks against each TG. In fact, the VA analysis results (see Section 3.3) indicate that the most discussed/mentioned TGs in Twitter are not always the most attacked ones as well.

3.2.2 EXPLORATIVE ANALYSIS

Data exploration is an integral part of the methodology, since the proposed approach is data-driven and sets out to incorporate human-in-the-loop; it helps to understand and obtain a broader view of the whole dataset and is crucial for filtering the data and clustering them into targeted collections that can be used for development and training purposes. Furthermore, valuable insights are extracted helping to finalize the knowledge representation framework. To this end samples of the collected data were explored using the ILSP Palomar Data Analysis and Modeling Platform (Papanikolaou et al., 2016). In particular, the Tweets were examined by a computational linguist (the author) and a political scientist focusing on the content of the verbal attacks (i.e. different aspects of VA, emerging stereotypes and themes discussed per TG) against the predefined TGs as well as on the types of the linguistic devices/weapons used for the attacks (i.e. linguistic instantiations of VAMs).

In a first phase a random selection of 1000 Tweets for each TG (10000 Tweets in total) was explored. The queries started as simple word or phrase queries (e.g. μουσουλμάνος) and resulted to complex ones using Boolean operators (e.g. μουσουλμάνος AND φανατισμός AND ...) (see Fig. 47). Based on initial observations and findings more Tweets were explored with more focused queries. This was an iterative procedure, as simultaneously the VA Framework was modified and improved, until it was finalized. The outcome of this phase was VA oriented data collections that were used for the development of the VA analyzer (approximately 1000 Tweets per TG) and valuable insights about the content and the types of verbal attacks and weapons that can be summarized as follows:

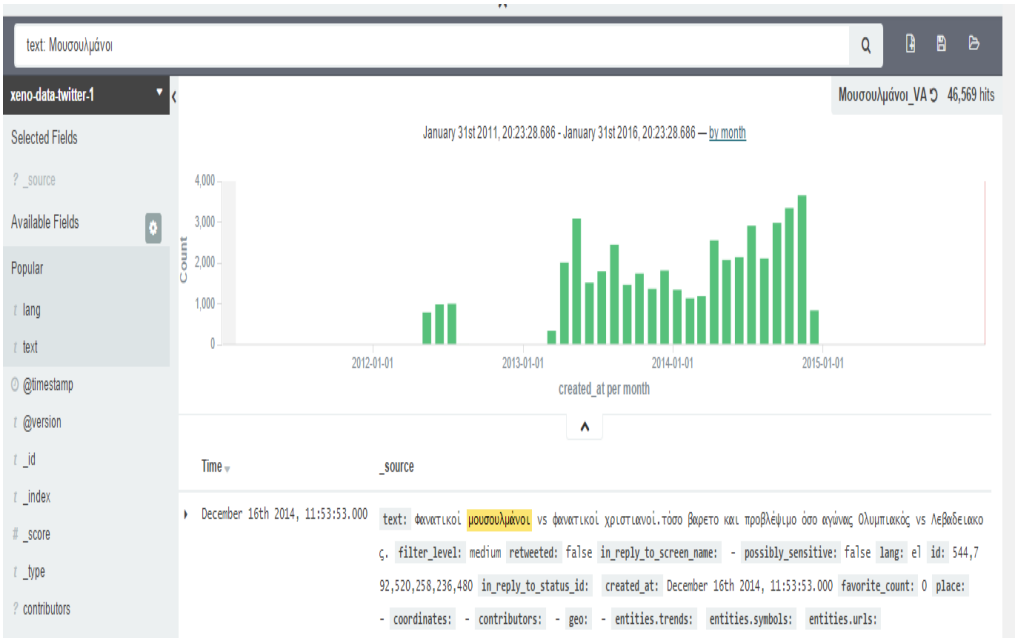


Figure 47: Exploring the Twitter collection, retrieving documents using the query "MUSLIM"

- The content and the type of the verbal attacks may vary; some attacks criticize or demean particular attributes of the targets often designating specific stereotypes and themes discussed per TG such as the problematic personal hygiene of PAKISTANI in (a) "*Αλήθεια δεν ξέρω τον λόγο που ο Πακιστανός στην γωνία του λεωφορείου μυρίζει σαν σάπιο κουνάβι*", the brutal violence stereotype for MUSLIMS in (b) "*Τι πολιτισμένα χτυπούν τις γυναίκες τους οι Μουσουλμάνοι!!!*" or the literate/cultural inferiority of ROMA in (c) "*Περιοδικό για τσιγγάνες μανάδες: Το πιντί μου κι εγκώ*".
- Other types of attacks are expressed in the form of direct or indirect threats and calls for different types of actions like ouster in (d) "*ΑΠΛΥΤΟΙ ΕΒΡΑΙΟΙ τραπεζικά καθάρματα φύγετε από τη χώρα*" or physical extinction in (e) "*Μισές δουλειές έκανε ο #Hitler με τους Εβραίους*".
- There are also attacks that convey the aggressor's threat perception like in (f) "*Δεν ανησυχώ! Γεννάνε αβέρτα οι Πακιστανές, οι αλβανίδες θα κάνουμε μιά Οθωμανική αυτοκρατορία Super.*"
- Focusing on the different types of sentiments that drive or underlie the verbal attacks, the explorative analysis reveals mainly feelings of (extreme) dislike, fear, and anger.
- Verbal attacks may be instantiated with different ways; explicitly or implicitly using a variety of linguistic devices and structures such as vulgar/obscene language, evaluative language, irony, metaphors, humor and jokes.

The explorative analysis points out also verbal attacks expressed in the context of a pro-immigrants/refugees discourse, especially in the case of SYRIANS. In addition, the findings indicate that it is more likely to verbally attack groups of people framed as IMMIGRANTS rather than REFUGEES due to the different connotations and implications of these two lexicalizations. All the above observations led to a VA framework (section 3.2.3). This framework remains to be confirmed or disproved (section 3.3) by large-scale data analysis using computational methods (section 3.2.4).

Finally, it is worth to be noted that the data exploration revealed also another type of xenophobic attitude, namely self-reports which in many cases constitute explicit xenophobic identity statements (e.g. *Είμαι σε όλα τα είδη ρατσίστρια, Αλβανοί, Πακιστανοί κλπ!*), and which are, however, out of the scope of this thesis.

3.2.3 VA FRAMEWORK

Based on literature review and explorative analysis findings a linguistically-driven VA framework is designed where VAMs are classified into distinct categories based on specific linguistic criteria. A data-driven approach is employed focusing on explicitly stated aggressive messages/expressions towards the TGs of interest. Given a collection of Tweets, the goal is to identify different types of verbal attacks against the TGs following the typology presented below in Fig. 48.

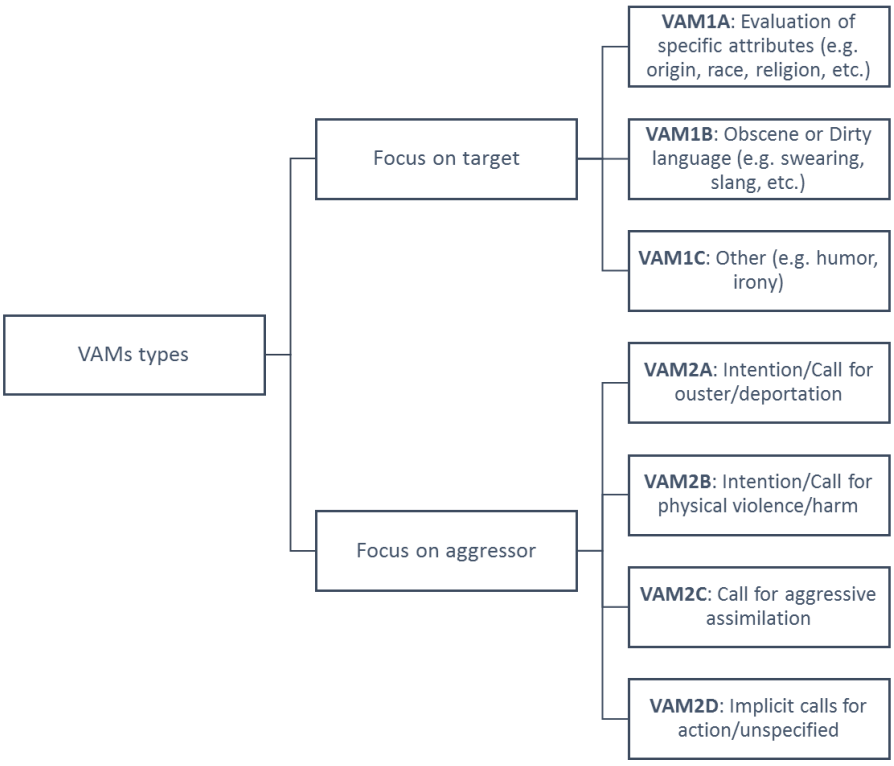


Figure 48: Typology of VAMs

As illustrated above in Fig. 48, VAMs are classified into distinct categories based on:

- Their focus (i.e. distinguishing between VA utterances focusing on the target of the attack and VA utterances focusing on the attacker).
- The type of the linguistic weapon used for the attack (e.g. formal evaluations, obscene/dirty language, humor).
- The content of the attack (e.g. threats/calls for physical violence or for deportation).

In particular, we consider two main types of VAMs (VAM1 and VAM2) that are further categorized in specific subtypes:

■ **VAM1:** Messages of this type focus on (the attributes of) the target (e.g. physical appearance, religion, etc.) and are further classified into subcategories based on the type of the linguistic devices (weapons) used by the aggressor to attack the target:

- **VAM1A**: Formal evaluations of specific attributes (e.g. origin, race, religion, etc.) e.g.

  "*και κάτι που ξέχασα να προσθέσω είναι ότι η θρησκεία (Μουσουλμάνοι) δεν χαρακτηρίζεται από καινοτομίες..."*

  [*Islam is not characterized by innovation (= meaning forward thinking)…*]

- **VAM1B**: Taboo or dirty language (e.g. swearing, slang, etc.) e.g.

  " *Γαμω τους αλβανούς ρε φίλε...."* [*Fucking Albanians…*]

  Note that messages of this type may also express evaluation about specific attributes (e.g. *dimwit*). Obscene messages are considered a separate category because they can provide different types of insights. For example, as mentioned above in section 3.3.1, depending on the online context, swearing may indicate different levels of aggression. In addition, swearing can act as an in-group solidarity marker, as when a group shares identical swearing norms (Mercury 1995; Allan and Burridge, 2006; Crystal 1995).

- **VAM1C**: Other (e.g. humor, irony) e.g.

  "Ευτυχώς που η φύση κρατάει ισορροπία και σκοτώνονται οι Εβραίοι με τους φανατικούς μουσουλμάνους !!!"

  [*Jews and Muslims are killing each other…fortunately nature keeps a balance!!!*]

■ **VAM2:** Messages of this type focus on the aggressor's intentions providing information about specific types of attack and are further classified into subcategories based on content the of the attack:

- **VAM2A**: Intentions or calls for ouster/deportation (oriented to legal means) e.g.

  "*Να φύγουν όλοι οι Αλβανοί απ την Ελλάδα καιρός είναι*"

  [*It's about time for all Albanians to leave Greece*]

- **VAM2B**: Intentions or calls for physical violence/harm (oriented to physical extinction) e.g.

  "*ΦΡΙΚΤΟΣ θάνατος στο Πακιστανικό κτήνος"*

  [*Murder that Pakistani beast*]

- **VAM2C**: Call for aggressive assimilation e.g.

  "*Να εκχριστιανιστούν οι Μουσουλμάνοι μετανάστες αν θέλουν άδεια εργασίας στην Ελλάδα. Μαθήματα γλώσσας κι ελληνικής ιστορίας."*

  [*Muslims should be baptized if they want job permission*]

- **VAM2D**: Implicit or unspecified call for action e.g.

  "*Θα συνεχίσουμε να κάνουμε τους χαζούς μπροστά στον ισλαμικό κίνδυνο;"*

[*Will we keep pretending that there is no Islamic danger?*]

In order to be able to apply this typology at a large scale, a Data Analytics pipeline was developed for automatic VA analysis as described in the following section.

3.2.4 COMPUTATIONAL ANALYSIS

For the computational treatment of the proposed framework a linguistically-driven VA analyzer has been designed and implemented using as development data the focused collections of Tweets created during the explorative analysis phase. Given an input text (i.e. a Tweet), detects VAMs towards the TGs of interest and classifies them according to the typology presented above in the previous section. The approach is lexicon-based and explores shallow syntactic relations between aggressive terms (i.e. words that are used to express VA) and sequences of Tokens-candidate targets of the attacks using linguistic patterns. The overall architecture for the VA analysis is illustrated in Fig. 49.

The input for the VA analyzer is raw data (Twitter collections). In a first phase the data are processed through a Natural Language Processing (NLP) pipeline that performs tokenization, sentence splitting, part-of-speech tagging, and lemmatization using the ILSP suite of NLP tools for Greek (Papageorgiou et al., 2002; Prokopidis, Georgantopoulos and Papageorgiou, 2011). In the next phase, the pre-processing output is given as input to the Semantic Analysis Unit, which performs VA analysis with a rule-based method that comprises a variety of lexical resources and grammars (sets of linguistic patterns). The VA analyzer is a Finite State Transducers (FST) cascade implemented as a JAPE grammar (Cunningham, Maynard and Tablan, 2000) in the GATE framework. These FSTs process annotation streams with regular expressions to create generalized rules. Moreover, they are ordered in a cascade, so that the output of an FST is given as input to the next transducer.

The method is precision-oriented and focuses on explicitly stated VA; it relies on a set of lexical resources that are built to capture possible linguistic instantiations of VA towards the TGs of interest. In particular, the lexical resources cover five of the seven types of VAMs (1A/B and 2A/B/C) included in the typology (see below Section 3.2.4.1). VAMs that are instantiated through complex linguistic structures and devices (i.e. humor, irony, implicit calls for action), and cannot be captured at the lexical level are out of the scope of the proposed approach. Exceptions are some specific cases of VAMs of types 1C and 2D that were found repeatedly in the data -reproducing some well-known stereotypes towards specific TGs- and were addressed with lexico-syntactic patterns (see below Section 3.2.4.2).

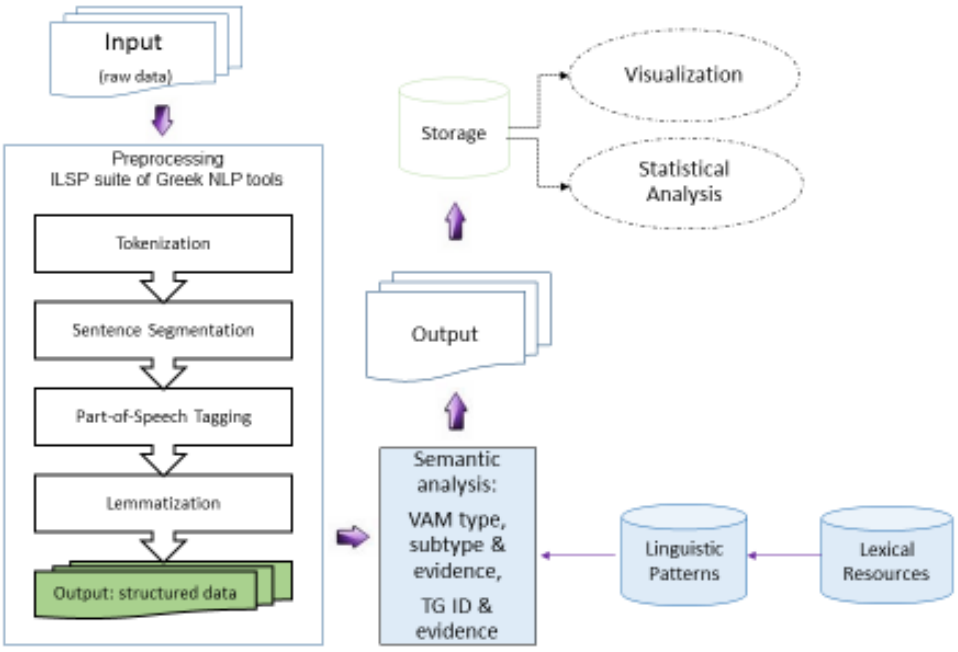


Figure 49: Architecture for VA analysis

In a first phase, the analyzer detects candidate VAMs and candidate targets based on the respective lexical resources; if a token is recognized as a lexicon entry then it is annotated with the respective metadata (lexicon labels). In a subsequent step, the grammars (Section 3.2.4.2) determine which spotted candidate VAMs and targets are correct. The output is recorded in a KD (Section 3.2.4.3) and is, then, used for statistical analysis and visualizations (Section 3.2.5). The experimental evaluation results confirmed the expectations favoring a precision-oriented method (Section 3.2.4.4).

3.2.4.1 LEXICAL RESOURCES

The VA analyzer relies on the following hand-crafted lexical resources:

- **TG Lexicon**. It contains 61 entries, possible lexicalizations of the TGs e.g. "*μουσουλμάνος*", "*ισλαμικός*" for TG5 (MUSLIM/ISLAM), "*Ισραηλίτης*", "*Εβραίος*", "*Σιωνιστικός*" for TG6 (JEWS). Each term is assigned a respective TG id label.

- **TGVA Lexicon**. It contains 72 entries, possible lexicalizations of the TGs that express at the same time VA towards them e.g. racial slurs, derogatory morphological variations of nationality adjectives (e.g. *Πακιστανά, Αλβαναριό, Οβραίοι, Ισλαμοπίθηκος*). All entries of this lexicon are considered by default

verbal attacks in all contexts; each term is assigned a respective TG id label, and a semantic label indicating the VAM type it belongs to.

- **VAM1 Lexicon**. It contains a customized version of EvalLex (Pontiki, Aggelou and Papageorgiou, 2013; Pontiki and Papageorgiou, 2015), an Appraisal Theory (Martin and White, 2005) grounded Lexicon for Evaluative Language that was manually compiled for the Greek language. EvalLex contains 5887 terms in total. Each term is assigned a label according to its category (i.e. adjective (JJ), adverb (RB), noun (NN), or verb (VB)) and its sentiment polarity (i.e. negative (n), positive (p), or both (b)). In addition, the terms are further classified as follows based on the strength degree of their evaluative meaning (EM) and prior polarity (PP): 1) Strong EM with a strong (p/n) PP e.g. "*υπεροπτικός" ("arrogant")*: [JJ1n]. 2) Weak EM with a strong (p/n) PP e.g. "*ώριμος" ("mature")*: [JJ2p]. 3) Strong or weak EM with a weak (p/n/b) or no PP e.g. "*μικρός" ("small")*: [JJ3b].

  A subset of this lexicon is used for the detection of VAMs that focus on the attributes of the target of the verbal attack. In particular, 2434 evaluative terms (968 adjectives and 1455 nouns) with negative orientation were further grouped in two semantic categories -VAM1A for formal evaluations and VAM1B for obscene language- (e.g. "*αναλφάβητος" ("illiterate")*: [JJ2n, VAM1A], "*μπάσταρδος" ("bastard")*: [JJ1n, VAM1B]) and used to detect candidate VAMs of these two types.

- **VAM2 Lexicon**. It contains 80 terms used to express verbal aggression focusing on the aggressor's intentions (VAM2). Each term is assigned a label according to its category (i.e. noun (NN), or verb (VB)) and the aggression type it indicates (e.g. "*διώχνω*" ("*oust*"): [VB, VAM2A], "*θάνατος*" ("*murder*"): [NN, VAM2B], "*αποποιούμαι*" ("*abnegate*"): [VB, VAM2C]). Verbs are further classified according to their syntactic behavior. For example, verbs like "αποχωρώ/φεύγω" (leave, go away) are activated as lexicon entries only when they appear in second or third person, whilst verbs like "διώχνω" (kick out) only when they appear in first and second person.

- **Modifiers Lexicon**. A typical SA resource that contains intensifiers (e.g. "*τελείως*" ("*totally*")), downtoners (e.g. "*κάπως*" ("*somewhat*")), and negators (e.g. "*καθόλου*" ("*not at all*")).

3.2.4.2 GRAMMARS

As already mentioned above, the analyzer detects candidate VAMs and targets based on the respective lexical resources. For example, given as input the following tweets:

(1) "Φρικτός θάνατος στον Πακιστανό της Πάρου!"

[*Horrible death to the Pakistani in Paros](=Murder that Pakistani in Paros)*

(2) Μετανάστες βρήκαν φρικτό θάνατο από ασφυξία.

[*Immigrants found horrible death from suffocation.*]

The analyzer detects the words "*θάνατος*" και "*θάνατο*" (=death) as a candidate attacks and the words "*Πακιστανό*" and "*Μετανάστες*" as candidate targets of the attacks, respectively. Then, in a second phase linguistic grammars impose restrictions in the context around the spotted candidate verbal attacks and targets and determine which of them are correct, since the appearance of aggressive terms and linguistic instantiations of the TGs of interest in a text or text snippet does not necessarily entail a verbal attack against them in all contexts (e.g. sentence (1) vs sentence (2)).

The grammars are the implementation of multi-phase algorithms where the output of each phase is input for the next one. Each phase comprises several modules that contain a variety of contextual lexico-syntactic patterns. The patterns are templates that generate rules in the context around the candidate verbal attacks and targets using shallow syntactic relations. In particular, the analyzer comprises two grammars, one for each basic type of VA:

- **VAM1 Grammar**. It consists of 5 phases and respective modules that contain a total of 59 lexico-syntactic patterns. A first set of rules performs stepwise VA and target detection using combinations between TG Lexicon entries and VAM1 Lexicon adjectives (phase 1) and nouns (phase 2), respectively. For example, pattern (A) below marks sequences of specific types of adjectives and nouns as VAMs and targets, respectively, when they appear in vocative case (e.g. "*αναλφάβητε Αλβανέ*"). The next set of patterns (phase 3) generates rules in the context around the candidate attacks and targets focusing on verbal structures; the patterns exploit shallow syntactic relations between verbs like "είμαι/αποτελώ" (be/constitute) and VAM1 lexicon entries as well as between verbs that express VA (e.g. *γαμώ (fuck), σιχαίνομαι (hate)*) and TG Lexicon entries.

```
(A)If Token[i] ∈ VAM1Lex
            & Token[i].Type ==“JJ2n” & Token[i].POS ==“Aj.*Vo”
               & Token[i+1] ∈ TGLex & Token[i+1].POS == “No.*Vo”
   then Token[i].Label = “VA”
             & Token[i].Class == Token[i].VAM_Type
                & Token[i+1].Label == “TARGET”
                & Token[i+1].Class == Token[i+1].ID
```

Another set of patterns is used to detect VAMs of type 1C (phase 4), namely ironic or humoristic messages that cannot be captured at the lexical level. These patterns are used to capture specific types of messages which are used to reproduce common stereotypes about specific TGs and were found repeatedly during the explorative analysis (e.g. *σαπούνι για τους Εβραίους (soap for JEWS)*, *κάλτσα με πέδιλο για τους Γερμανούς (the usual stylistic choice of GERMAN tourists to wear Birkenstock sandals with socks*)), and not to address complex linguistic phenomena like humor or irony in general. Finally, phase 5 uses propagation rules (Qui et el. 2011) in order to capture VAMs and targets based on already detected ones (e.g. when they appear in conjunction).

- **VAM2 Grammar**. It consists of 3 phases and respective modules that contain a total of 35 patterns. Two sets of patterns perform VA and target detection using combinations between TG Lexicon entries and VAM2 Lexicon nouns (phase 1) and verbs (phase 2), respectively. For example, according to pattern (B), when specific types of verbs that belong to VAM2 Lexicon appear in other than past tenses after a token that is not a negator and they are followed by an article or a personal pronoun and a TG lexicon entry, then they are tagged as VAMs towards the specific TGs. The specific pattern, as most of the patterns, includes also some optional elements (e.g. prepositional modifiers, adjectives or adverbs before or after the lexicon entries, etc.) in between the mandatory ones that are illustrated above, in order to capture as many as possible different syntactic variations of the specific structure. For example, pattern (B) captures expressions like “στείλτε τον (αλήτη) (τον) Εβραίο σπίτι του”, “να διώξουμε (άμεσα) (όλους) τους (παράνομους) μετανάστες”, etc.

```
(B)If Token[i-1] ∉ Negators
          &Token[i] ∈ VAM2Lex
               & Token[i].Type ==“(VAM2A_VB2|VAM2A_VB3|VAM2A_VB4)”
                    & Token[i].POS !~ “Vb.*Pa”
                         & Token[i+1].POS == “(At|Pn)”
                              & Token[i+2] ∈ TGLex
       then Token[i].Label = “VA”
              & Token[i].Class == Token[i].VAM_Type
                  & Token[i+2].Label == “TARGET”
                  & Token[i+2].Class == Token[i+2].ID
```

Finally, as in the case of VAMs of type 1C, another set of patterns (phase 3) is used to address VAMs of type 2D, namely implicit or unspecified calls for action that cannot be captured at the lexical level. These patterns detect specific expressions/slogans found repeatedly during the explorative analysis (e.g. “ξυπνάτε/γρηγορείτε Έλληνες” (Greeks wake up!)).

3.2.4.3 OUTPUT

The Twitter collections described in section 3.2.1 were automatically processed through the Data Analytics pipeline for VA analysis described in the previous section. An example of the VA analysis output in the GATE environment is illustrated below in Fig. 50. For each identified verbal attack, the method returns as a structured output a tuple that contains information about the type and the target of the attack, and their linguistic instantiations. In addition, the tuples contain information about the VA analyzer’s phase, module and rule that captures the verbal attack in each case (this is helpful for development and upgrade purposes).

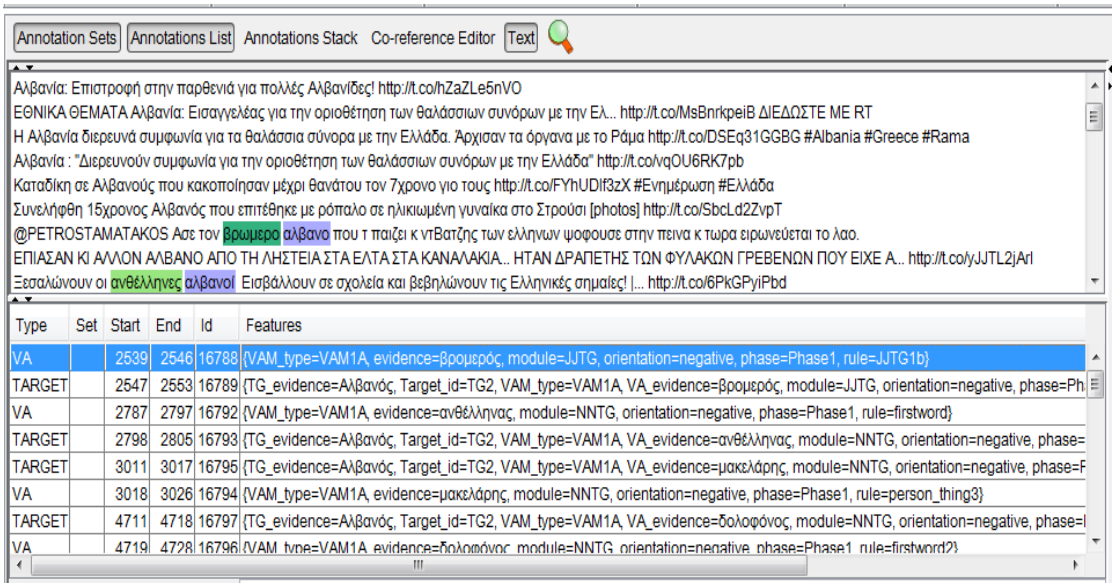


Figure 50: VA analysis output example from GATE

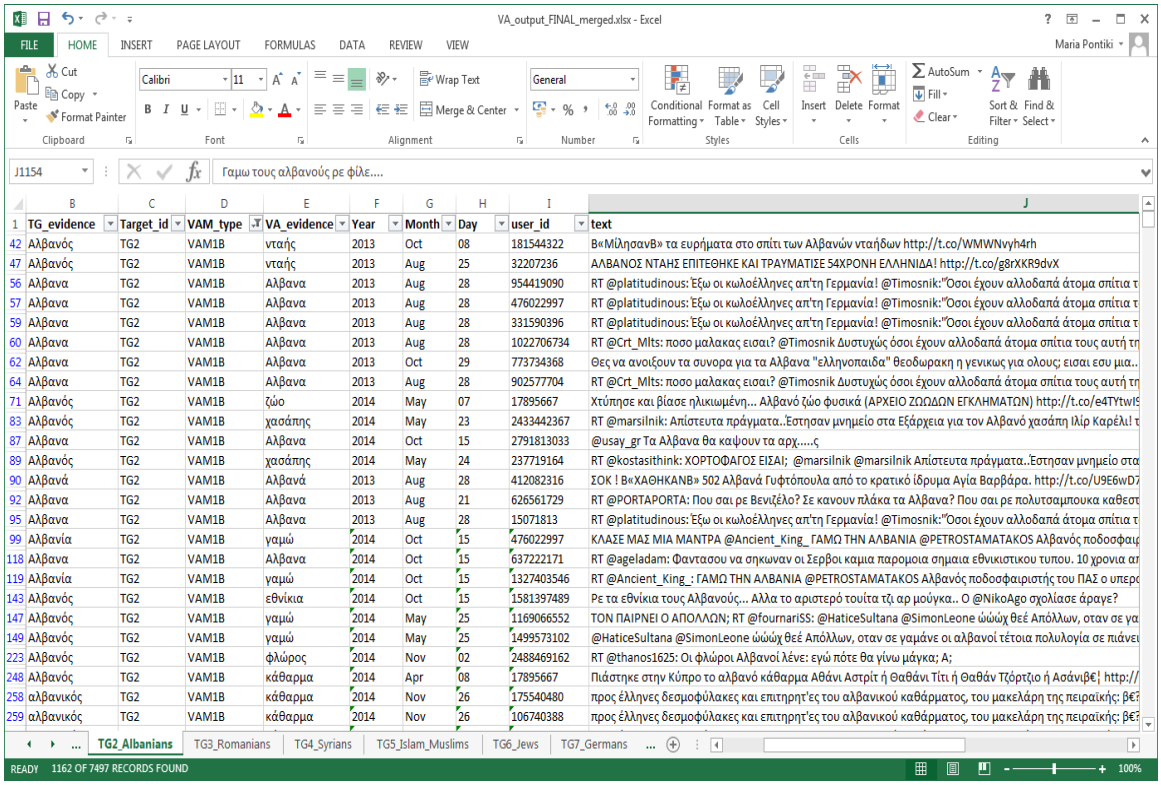
The output is then recorded in the Knowledge Database so that it can be used for statistical analysis and visualizations. For each processed Tweet the Knowledge Database is populated with two types of metadata following the structure described below:

- Annotations derived by the automatic VA analysis:

  - **TG_id** (string variable): the unique ID label that has been assigned for each TG of interest (predefined values: TG0-TG9).
  - **TG_evidence** (string variable): The lexicalization of the TG as referred to in the Tweet.
  - **VAM_type** (string variable): the type of the VAM as it is coded in the typology (predefined values: VAM1A, VAM1B, VAM1C, VAM2A, VAM2B, VAM2C, VAM2D).
  - **VA_evidence** (string variable): The lexicalization of the verbal attack as it appears in the Tweet.

- Twitter metadata:

  - **Tweet timestamp** (numeric variables): Year, Month, Day.
  - **User_id** (numeric variable): The Twitter ID of the user that texted the Tweet.
  - **Text**: The actual Tweet message.



	TG_evidence	Target_id	VAM_type	VA_evidence	Year	Month	Day	user_id	text
42	Αλβανός	TG2	VAM1B	νταής	2013	Oct	08	181544322	Β«ΜίλησανΒ» τα ευρήματα στο σπίτι των Αλβανών νταήδων http://t.co/WMWNvyh4rh
47	Αλβανός	TG2	VAM1B	νταής	2013	Aug	25	32207236	ΑΛΒΑΝΟΣ ΝΤΑΗΣ ΕΠΙΤΕΘΗΚΕ ΚΑΙ ΤΡΑΥΜΑΤΙΣΕ 54ΧΡΟΝΗ ΕΛΛΗΝΙΔΑ! http://t.co/g8rXKR9dvX
56	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	954419090	RT @platitudinous: Έξω οι κωλοέλληνες απ'τη Γερμανία! @Timosnik:"Οσοι έχουν αλλοδαπά άτομα σπίτια τ
57	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	476022997	RT @platitudinous: Έξω οι κωλοέλληνες απ'τη Γερμανία! @Timosnik:"Οσοι έχουν αλλοδαπά άτομα σπίτια τ
59	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	331590396	RT @platitudinous: Έξω οι κωλοέλληνες απ'τη Γερμανία! @Timosnik:"Οσοι έχουν αλλοδαπά άτομα σπίτια τ
60	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	1022706734	RT @Crt_MIts: ποσο μαλακας εισαι? @Timosnik Δυστυχώς όσοι έχουν αλλοδαπά άτομα σπίτια τους αυτή τη
62	Αλβανα	TG2	VAM1B	Αλβανα	2013	Oct	29	773734368	Θες να ανοιξουν τα συνορα για τα Αλβανα "ελληνοπαιδα" Θεοδωρακη η γενικως για ολους; εισαι εσυ μια..
64	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	902577704	RT @Crt_MIts: ποσο μαλακας εισαι? @Timosnik Δυστυχώς όσοι έχουν αλλοδαπά άτομα σπίτια τους αυτή τη
71	Αλβανός	TG2	VAM1B	ζώο	2014	May	07	17895667	Χτύπησε και βίασε ηλικιωμένη... Αλβανό ζώο φυσικά (ΑΡΧΕΙΟ ΖΩΩΔΩΝ ΕΓΚΛΗΜΑΤΩΝ) http://t.co/e4TYtwIS
83	Αλβανός	TG2	VAM1B	χασάπης	2014	May	23	2433442367	RT @marsilnik: Απίστευτα πράγματα..Έστησαν μνημείο στα Εξάρχεια για τον Αλβανό χασάπη ΙΙίρ Καρέλι! τ
87	Αλβανα	TG2	VAM1B	Αλβανα	2014	Oct	15	2791813033	@usay_gr Τα Αλβανα θα καψουν τα αρχ.....ς
89	Αλβανός	TG2	VAM1B	χασάπης	2014	May	24	237719164	RT @kostasithink: ΧΟΡΤΟΦΑΓΟΣ ΕΙΣΑΙ; @marsilnik @marsilnik Απίστευτα πράγματα..Έστησαν μνημείο στα
90	Αλβανά	TG2	VAM1B	Αλβανά	2013	Aug	28	412082316	ΣΟΚ ! Β«ΧΑΘΗΚΑΝΒ» 502 Αλβανά Γυφτόπουλα από το κρατικό ίδρυμα Αγία Βαρβάρα. http://t.co/U9E6wD7
92	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	21	626561729	RT @PORTAPORTA: Που σαι ρε Βενιζέλο? Σε κανουν πλάκα τα Αλβανα? Που σαι ρε πολυτσαμπουκα καθεστ
95	Αλβανα	TG2	VAM1B	Αλβανα	2013	Aug	28	15071813	RT @platitudinous: Έξω οι κωλοέλληνες απ'τη Γερμανία! @Timosnik:"Οσοι έχουν αλλοδαπά άτομα σπίτια τ
99	Αλβανία	TG2	VAM1B	γαμώ	2014	Oct	15	476022997	ΚΛΑΣΕ ΜΑΣ ΜΙΑ ΜΑΝΤΡΑ @Ancient_King_ ΓΑΜΩ ΤΗΝ ΑΛΒΑΝΙΑ @PETROSTAMATAKOS Αλβανός ποδοσφαιρ
118	Αλβανα	TG2	VAM1B	Αλβανα	2014	Oct	15	637222171	RT @ageladam: Φαντασου να σηκωναν οι Σερβοι καμια παρομοια σημαια εθνικιστικου τυπου. 10 χρονια αr
119	Αλβανία	TG2	VAM1B	γαμώ	2014	Oct	15	1327403546	RT @Ancient_King_: ΓΑΜΩ ΤΗΝ ΑΛΒΑΝΙΑ @PETROSTAMATAKOS Αλβανός ποδοσφαιριστής του ΠΑΣ ο υπερ
143	Αλβανός	TG2	VAM1B	εθνίκια	2014	Oct	15	1581397489	Ρε τα εθνίκια τους Αλβανούς... Αλλα το αριστερό τουίτα τζι αρ μούγκα.. Ο @NikoAgo σχολίασε άραγε?
147	Αλβανός	TG2	VAM1B	γαμώ	2014	May	25	1169066552	ΤΟΝ ΠΑΙΡΝΕΙ Ο ΑΠΟΛΛΩΝ; RT @fournariSS: @HaticeSultana @SimonLeone ώώώχ θεέ Απόλλων, οταν σε γα
149	Αλβανός	TG2	VAM1B	γαμώ	2014	May	25	1499573102	@HaticeSultana @SimonLeone ώώώχ θεέ Απόλλων, οταν σε γαμάνε οι αλβανοί τέτοια πολυλογία σε πιάνει
223	Αλβανός	TG2	VAM1B	φλώρος	2014	Nov	02	2488469162	RT @thanos1625: Οι φλώροι Αλβανοί λένε: εγώ πότε θα γίνω μάγκα; Α;
248	Αλβανός	TG2	VAM1B	κάθαρμα	2014	Apr	08	17895667	Πιάστηκε στην Κύπρο το αλβανό κάθαρμα Αθάνι Αστρίτ ή Θαθάν Τίτι ή Θαθάν Τζόρτζιο ή Ασάνι€¦ http://
258	αλβανικός	TG2	VAM1B	κάθαρμα	2014	Nov	26	175540480	προς έλληνες δεσμοφύλακες και επιτηρητ'ες του αλβανικού καθάρματος, του μακελάρη της πειραϊκής: β€?
259	αλβανικός	TG2	VAM1B	κάθαρμα	2014	Nov	26	106740388	προς έλληνες δεσμοφύλακες και επιτηρητ'ες του αλβανικού καθάρματος, του μακελάρη της πειραϊκής: β€?

Figure 51: Snapshot of the Knowledge Database

The Tweet timestamp is split in three separated fields (Day, Month, Year) instead of one (Day/Month/Year) in order to be able to produce more fine-grained visualizations like timelines and thus, to better monitor the evolution of VA in time. The information about the User Id can help to identify highly aggressive users as well as to be exploited for social network analysis in order to spot specific communities that promote xenophobic attitudes. A snapshot of the database is provided above in Fig. 51. Finally, the output was visualized in various ways (see below section 3.2.5) giving a better understanding of the data and the results of the VA analysis.

3.2.4.4 EVALUATION

The performance of the VA analyzer is evaluated using a random selection of 500 Tweets per TG (5000 Tweets in total). The test data was annotated semi-automatically; a computational linguist (the author) inspected and corrected the VA analyzer's output (automatically annotated VA tuples) on this data. Then, the originally machine annotated data was compared to the gold data (human annotated) using the GATE Annotation Diff Tool (Fig. 52).

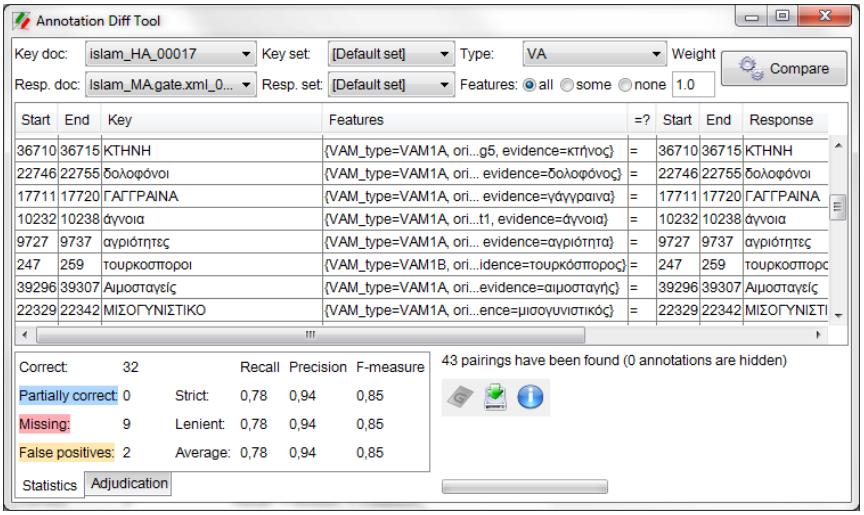


Figure 52: Example of evaluation using the GATE Annotation Diff Tool

The performance of the analyzer is measured in terms of Precision (P), Recall (R) and F-Measure (F-1). P, R and F-1 are defined as follows:

$$P = \frac{|S \cap G|}{|S|} \qquad R = \frac{|S \cap G|}{|G|} \qquad F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

S is the set of the VA tuples that the system returned for all the test Tweets, and *G* is the set of the gold (correct) VA tuples-annotations. F1 score is the harmonic mean of *P* and *R*.

Evaluation is performed not only on the total test set (5000 Tweets), but also separately for each TG-specific sub-collection (500 Tweets per TG) in order to obtain a more fine-grained and in-depth view of the results. As it is presented below in Table 11, the evaluation results confirmed the expectations favoring a precision-orientated method, since -with the exception of the last two TGs (IMMIGRANTS and REFUGEES)- it ranges from 80% to 94%. However, as it was also expected for a precision-oriented method, it suffers in terms of recall (60% overall).

In general, the precision is negatively affected mainly in cases of quoted or reported/indirect speech. For example, the Tweet "*Κούρδισσα Στρατηγός: Θα εξαφανίσουμε το Ισλαμικό Κράτος από το πρόσωπο της γης*" reproduces a threat towards the ISIS expressed by a Kurdish General; the detected VA tuple is not considered correct, since it does not convey a threat expressed by the user of a Greek Twitter account.

	PRECISION	RECALL	F-MEASURE
Overall (total test set)	84%	60%	68%
TG1: PAKISTANI	93%	65%	76%
TG2: ALBANIANS	89%	25%	39%
TG3: ROMANIANS	89%	67%	76%
TG4: SYRIANS	80%	67%	73%
TG5: MUSLIMS/ISLAM	94%	78%	85%
TG6: JEWS	90%	56%	69%
TG7: GERMANS	88%	39%	54%
TG8: ROMA	83%	45%	59%
TG9: IMMIGRANTS	70%	78%	74%
TG0: REFUGEES	67%	80%	73%

Table 11: Evaluation Results of VA Analyzer

The low precision in the case of IMMIGRANTS and REFUGEES (70% and 67%, respectively), is also due to ironic messages, were VA is used in the context of a pro-immigrant/refugees discourse and is directed towards anti-immigrant/refugees perceptions, behavior, practices, etc. For example, the Tweet "*Να διώξουμε τους μετανάστες που μας παίρνουν τις δουλειές των 10€/μέρα μπροστά από τα ντουφέκια των επιστατών*" expresses a call for ouster/deportation of immigrants because they are taking jobs from Greeks by referring to an incident of an attack against a group of strawberry pickers from Bangladesh who were shot by their Greek employers at Manolada farm after demanding unpaid wages. It is worth to be noted that, as regards the specific TGs, not only the recall is better than the precision, but also the VA analyzer achieves the best results in terms of recall in general (78-80%); this may indicate that the coverage of the VA towards these two TGs is better as compared to the rest of TGs and that they probably receive fewer verbal attacks.

Turning to the recall, the error analysis findings indicate that it is negatively affected in the following cases:

- Misspellings/typos e.g. "*θα σας γαμης@@@*", "*Αλβανός μαλ@κας κράτησε ομήρους σε σούπερμάρκετ στου Ζωγράφου*".

- The limitations of the shallow syntactic relations modelling, since long distance dependencies cannot be captured through a window of a limited number of tokens e.g.

  "*Αλβανοί φορούν το χακί του Ελληνικού Στρατού και προκαλούν*".

- Complex linguistic structures and devices that a lexicon-based method is not designed and cannot address (e.g. humor, irony, implicit intentions and calls for action). This is the main reason for the low recall especially in the case of ALBANIANS (25%), GERMANS (39%) and ROMA (45%), where the analyzer fails to capture a great amount of jokes/anecdotes attacking different attributes/deficiencies of specific TGs such as the Greek accent of ALBANIANS (e.g. "*Θέλεις μου λέει ο Αλβανός Κορεάτικα λουκάνικα; Κορεάτικα του λέω; Ναι ρε μου λέει, από το Κωριό*"), the incestuous relationships between ROMA (e.g. "*Γύφτος χωρίζει με τη γύφτισσα φίλη του: Αν τες, μπορούμε να μείνουμε ξαντέρφια...*") or the cultural and intellectual inferiority of GERMANS (e.g. "*Την εποχή που εμείς κάναμε εγκαίνια στην Ακρόπολη οι Γερμανοί είχαν ουρά και πηδούσαν από κλαδί σε κλαδί*").

Overall, the evaluation results and the error analysis findings suggest that the analyzer addresses sufficiently explicitly stated verbal attacks; as for the cases of wrong and worst precision, in most of the them, the identified VAMs constitute indeed verbal attacks, but the method fails to identify and assign the correct target (in the cases of irony) or does not take into account the aggressor (in the cases of quoted/indirect speech), since it currently assumes that the user who tweets is also the actor/opinion holder. The best performance both in terms of precision and recall is reported for MUSLIMS/ISLAM (F-measure 85%), PAKISTANI and ROMANIANS (F-measure 76%); this may indicate that the attacks against these TGs are somewhat more straightforward, as opposed to other TGs (i.e. ALBANIANS, GERMANS, ROMA, JEWS) where language users' creativity in language play, humour and constructing jokes is unlimited and requires more sophisticated methods and techniques for its computational treatment.

3.2.5 DATA VISUALIZATION

The content of the KD is visualized in various ways in order to make the VA results explorable, comprehensible and thus more easily interpretable. The different types of information types that are extracted, allow for many different associations and graphs for both quantitative and qualitative analysis. In particular, the generated visualizations include:

- **Graphs** that display the VA analysis results per year and per TG (e.g. Fig. 53 below). Such graphs provide an overview of the most and the least attacked TGs and can help to monitor xenophobia in time (peak points, discontinuities etc.).

- **Pie charts** that present the distribution of the different VAM types per TG (e.g. Fig. 54 and 57 below). Such charts can help to explore whether different types of VA can be associated with different TGs, in other words to explore if "foreigners" can be framed based on specific VAM types.
- **Word Clouds** that display the unique aggressive terms captured per TG. Clouds of this type make the results understandable and easily usable for the human eye. They are very useful since they can provide access to the different attributes/aspects that are being attacked in each case and, thus, reveal dominant stereotypes per TG (e.g. Fig. 58-60 below).

3.3 RESULTS AND INTERPRETATION

This section presents the actual results of the VA analysis methodology described in the previous section. The detected verbal attacks are discussed both quantitatively and qualitatively with regard to the specific RQs that this thesis aims to address -focusing on the amount (RQ1: main targets of attacks), the type (RQ2) and the content (RQ3: stereotypes and prejudices) of the aggressive messages, respectively-, and are interpreted in the context the XENO@GR project as indicators of xenophobic attitudes in Greece over time.

3.3.1 RQ1: WHO ARE THE MAIN TARGETS OF TWITTER VERBAL ATTACKS?

Overall, the quantitative analysis of the verbal attacks indicates that xenophobic behaviors do not seem to be so dominant in Greek Twitter, since as illustrated below in Fig. 53 the VA rates (VAMs/Tweets) detected in Twitter regarding the specific TGs are low (i.e. the VA rate for the mostly attacked TG is approx. 4%). It should be noted that no data deduplication is performed so redundant or repeated Tweets are included in the analysis. Taking into account the evaluation results regarding the recall of the method (section 3.2.4.3), the actual verbal attacks expressed in Greek Twitter against some TGs (i.e. ALBANIANS, ROMA, GERMANS) may be much more than those that were captured by the VA analyzer, but still the VA rate is likely to be single-digit.

Focusing on the research goals of this thesis, the identity of the targets/victims can provide valuable insights about the xenophobic behavior of Greeks. According to the results of the VA analysis (Fig. 54) the most attacked TGs appear to be JEWS (23%), ALBANIANS (22%), PAKISTANIS (15%), MUSLIMS/ISLAM (14%), and IMMIGRANTS

(10%). As already mentioned in the data collection section, the results indicate that the most mentioned TGs are not always also the most attacked ones. In fact, REFUGEES is the most discussed but least attacked TG. The same holds also for the highly mentioned TGs of IMMIGRANTS and SYRIANS.

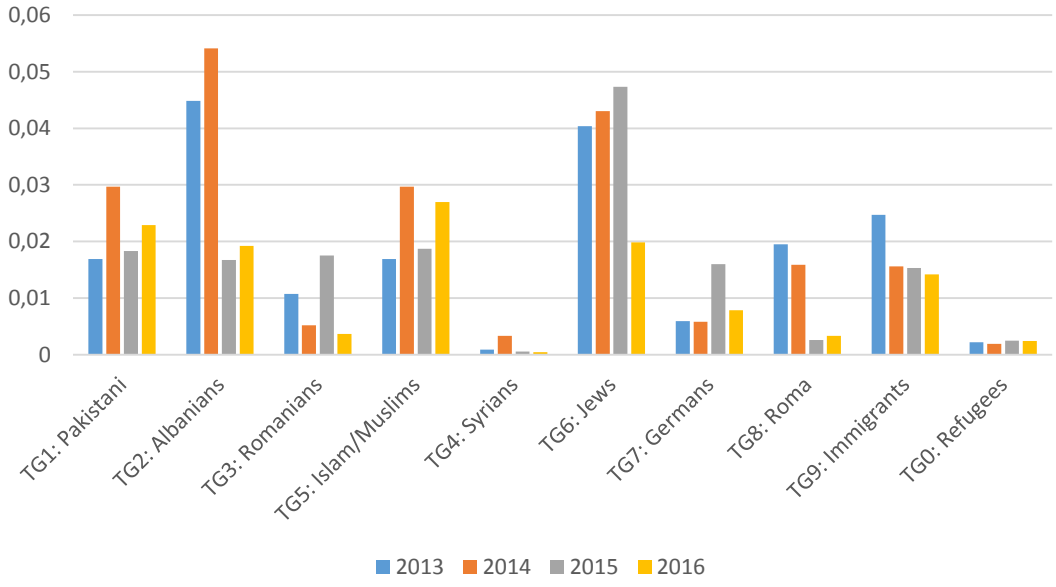


Figure 53: Per-year VA rate (VAMs/Tweets) per TG

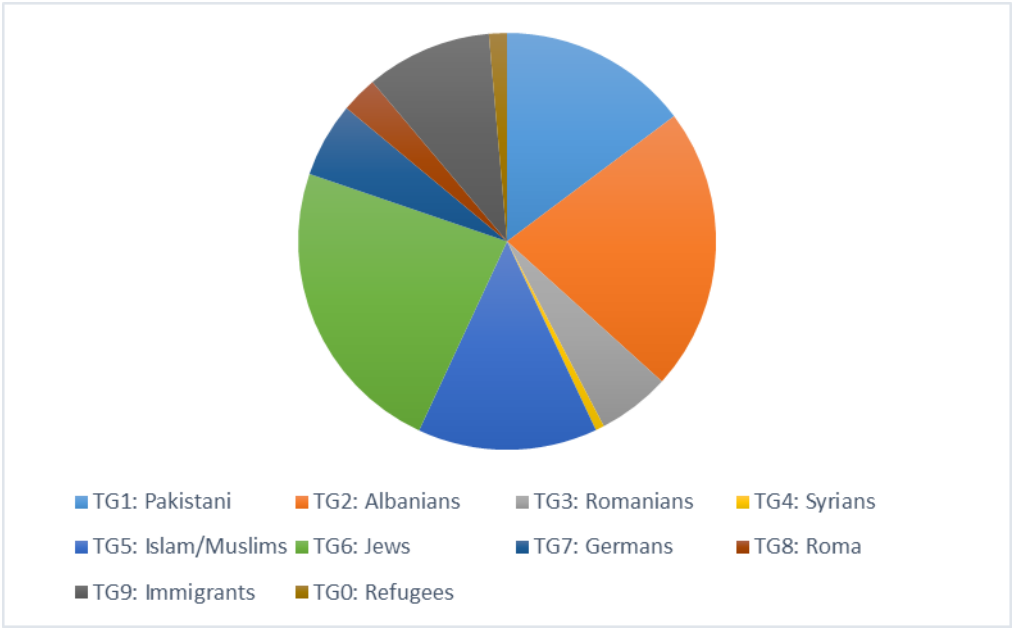


Figure 54: VA rate (VAMs/Tweets) per TG

Anti-Semitism seems to be at the core of the Greek Twitter aggressive discourse against foreigners. This observation can be examined in the context of recent survey findings; according to the ADL Global 100[49] survey, which elaborated an index of anti-Semitism based on the strength of anti-Semitic stereotypes, Greece was the most anti-Semitic country in Europe scoring 69%. The role of anti-Semitism in the current Greek political culture has attracted attention after a series of opinion poll findings and most importantly after the rise of neo-Nazi Golden Dawn, a party with an explicit anti-Semitic discourse (Georgiadou, 2015).

ALBANIANS and PAKISTANIS constitute the largest immigrant populations in Greece. ALBANIANS are perhaps the most established group of foreigners in Greek public discourse. According to research findings (Voulgaris et al., 1995) the image of foreigner as it was constructed in Greece during and after the first wave of migration flow (early 1990s-mid 1990s) was mainly associated with Balkan, and mainly Albanian, nationality. The results of the Twitter VA analysis indicate that almost 25 years later, several stereotypes and prejudices about ALBANIANS are still dominant and keep being reproduced (see below 1.3.3). PAKISTANIS who came to Greece during the 2000s increased migration flow from Asian countries (Pakistan, India, Bangladesh, Iraq, and Afghanistan) face also great hostility and are not so welcome.

MUSLIMS have a long time presence in Greece; actually, the Muslim minority in Thrace is the only explicitly recognized minority in Greece. According to the 2016 PEW survey findings a big majority of Greeks holds an unfavorable view of Muslims (65%), while 78% believes that Muslims residing in Greece "*want to be distinct from the larger society*" and are unwilling to adopt Greek "*customs and way of life*". These perceptions are somehow reflected through specific themes and stereotypes in the detected aggressive messages (e.g. aggressive terms used to debase core Islamic values, principles, practices, etc.). However, it is worth noting that verbal attacks targeted to MUSLIMS/ISLAM are triggered by geopolitical events such as the rise of ISIS or events related to violent practices or sexual abuse of specific population groups (women, children).

As for the generic group IMMIGRANTS, the results confirm that it is more likely to verbally attack groups of people framed as IMMIGRANTS rather than REFUGEES due to the different connotations and implications of these two lexicalizations. According to the 2016 PEW findings[50] Greek public opinion perceives the REFUGEES primarily as

[49] http://global100.adl.org/public/ADL-Global-100-Executive-Summary.pdf

[50] http://www.pewglobal.org/2016/07/11/europeans-fear-wave-of-refugees-will-mean-more-terrorism-fewer-jobs/

"*a burden for our country because they take our jobs and social benefits*" and secondarily as a potential link to terrorism or to the rise of criminality. The Twitter analysis results indicate that when it comes to self-motivated expression of opinions/aggression, the verbal attacks seem to be addressed to specific ethnic groups and not to IMMIGRANTS/ REFUGEES as targets in general.

3.3.2 RQ2: WHICH ARE THE MAIN TYPES OF TWITTER VERBAL ATTACKS?

The overall number of messages that express verbal aggression focusing on the target of the attack (VAM1) is quite bigger than the number of messages focusing on the aggressor's intentions (VAM2); the proportion of the detected VAMs of type 1 and 2 is approximately 89% and 11%, respectively. The distribution/rate of each VAM type per TG is illustrated below in Fig. 55.

Focusing on VAM1 attacks, the results (Fig. 56) indicate that the TGs who are mostly attacked with messages negatively evaluating specific attributes of them (VAM1A) are those of ALBANIANS and JEWS, whilst the TGs that receive the most obscene messages (VAM1B) are PAKISTANIS and IMMIGRANTS. As for the third subcategory (VAM1C), GERMANS receive the most ironic or humoristic messages. The qualitative analysis of the content of these attacks (see below 3.3.3) provides interesting insights about the evaluated attributes and the linguistic weapons used in each case.

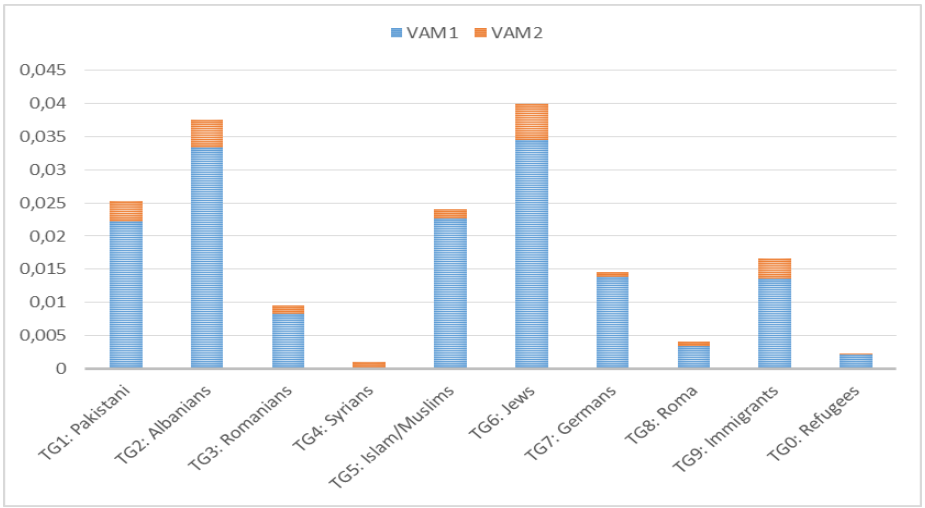


Figure 55: VAM1 and VAM2 distribution/rate per TG

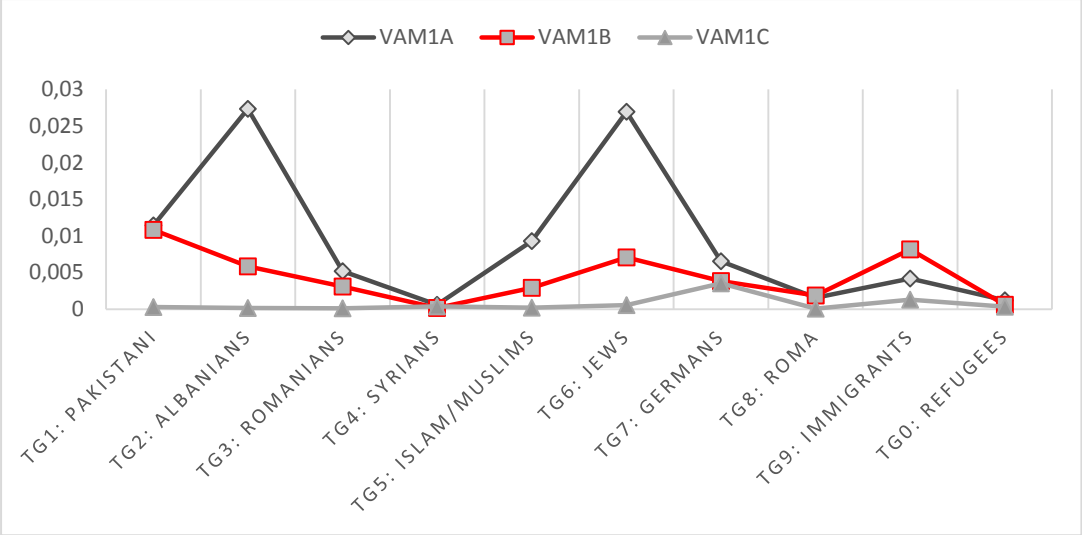


Figure 56: VAM1 subtypes rate per TG

Focusing on VAM2 attacks, given the limited amount of these messages only the overall distribution per TG is presented (Fig. 57). JEWS receive most of the attacks that focus on the aggressor's intentions with ALBANIANS and PAKISTANIS following in the second and third place, respectively.

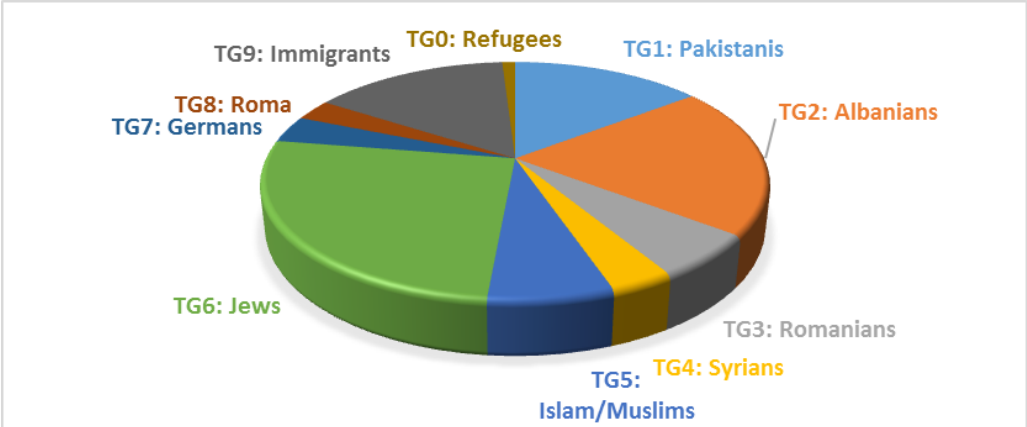


Figure 57: VAM2 distribution per TG

In fact, calls for physical extinction and are much more for JEWS than for any other group e.g. "*Κρεατακι Εβραιου Ο καλυτερος μεζες ΣΚΟΤΩΣΤΕ ΤΟΝ ΑΛΗΤΗ*" (="*Jewish meat…The best meal…Kill that (Jew) Bastard*"). What needs to be noted is that there is not a significant number of JEWS living in Greece as compared to ALBANIANS and PAKISTANI that constitute the largest immigrant populations in this country. Moreover, aggressive messages related to this specific TG reveal the

emergence of threat perception based on biological and cultural terms, as well as the perception of a particular enmity towards the Greek nation (see also below 1.3.3).

Threat perception seems to prevail also for PAKISTANIS, ALBANIANS and IMMIGRANTS, according to the share of VAM2 attacks and in particular the calls for ouster/deportation for the specific groups e.g. "*Άμεση απέλαση... Αφού δεν σέβονται την χώρα...*" (="*Immediately deport the immigrants…they do not respect our country*").

3.3.3 RQ3: ARE THERE STEREOTYPES AND PREJUDICES AGAINST FOREIGNERS ROOTED DEEPLY IN THE GREEK SOCIETY?

Stereotypes and prejudices are examined focusing on the content of the verbal attacks expressed in Twitter. To this end, the linguistic evidence of the aggressive messages is visualized using word clouds that contain the unique aggressive terms found per TG, based on the assumption that the unique linguistic weapons used against each TG may be associated with specific types of attributes or themes discussed per TG. The qualitative analysis of the results confirms the existence of stereotypes and prejudices against specific TGs that are deeply rooted in Greek society. A typical case is the stereotypes about ALBANIANS, perhaps the most established group of foreigners in Greek public discourse. The dominant stereotypes in the construction of the image of ALBANIANS are associated with "*crime*" and "*cultural inferiority*" (Fig. 58).

Figure 58: Word Cloud of unique aggressive terms for "Albanians"

This type of crime stereotypes is used also in the construction of the image of ROMANIANS, in addition to the sexual decadence (lax rules of sexual morality) stereotype for the Romanian women. These results indicate a continuity of the so-called stereotype of the Balkanian criminal (Albanian, Romanian, etc.) that became powerful in media representations of immigrants and in popular discourse in Greece between early 1990s and late 1990s, when migration flows alongside with the breakup of Yugoslavia, the war conflicts in the Balkan region and the emergence of the so-called Macedonian question contributed to the construction of a discourse on perceived threats concerning the country's territorial integrity as well as national and cultural identity (Voulgaris, 2006).

The same types of criminality and inferiority stereotypes are dominant also in the verbal attacks against ROMA, in addition to physical appearance and personal hygiene stereotypes, which are so deeply rooted in Greek society that word "*γύφτος*" (="gipsy") is used as a synonym for poor hygiene conditions, among others.

Inferiority and personal hygiene stereotypes are also dominant in the verbal attacks against PAKISTANIS. An interesting observation, is the limited number of unique aggressive terms for the specific TG (Fig. 59) -as compared to the other TGs (e.g. Albanians, Muslims, Jews, Germans)-, most of which are derogatory morphological variations of the nationality adjective implying inferiority (e.g. "*Πακιστάνια*", "*Πακιστανά*", "*Πακιστάνοι*", "*πάκι*").

Figure 59: Word Cloud of unique aggressive terms for "Pakistani"

In other words, with the exception of some messages focusing on the physical appearance (e.g. "*ανθυποπίθηκος*" (="looking like a monkey")) or the color skin (e.g. "*αράπης*" (="nigger")), PAKISTANIS are mostly evaluated as inferior beings using their nationality name as a linguistic weapon. In addition, it is worth mentioning that most

of the attacks are usually triggered by events related to violent practices or sexual abuse of specific population groups (women, children), with reference to their religion (Islam).

Crime and inferiority stereotypes are dominant also in the verbal attacks against MUSLIMS/ISLAM, but with rather different aspects. In particular, the attacks are often lexicalized through evaluative and dysphemistic terms of insult or abuse to debase core Islamic values, principles, practices, etc. (see Fig. 60). The stereotypes that are derived by the semantics of the unique aggressive terms for the particular TG are associated with brutal violence (e.g. "*αιμοχαρής*" (="bloodthirsty")), cultural inferiority (e.g. "*αναλφαβητισμός*" (="illiteracy"), "*σκοταδισμός*" (="obscurantism")), sexual abuse (e.g. "*παιδοφιλία*" (="paedophilia")), sexist behavior (e.g. "*μισογυνιστικός*" (="misogynistic, woman-hating")), fanaticism (e.g. "*ισλαμοφασίστες*" (= "islamofascists")), terrorism and irrationalism (e.g. "*εξτρεμιστικός*" (="extremist")). Moreover, the lexicalizations of the verbal attacks against MUSLIMS/ISLAM indicate notions of Islamophobia. An interesting observation is the use of biological metaphors most common in the representation of the Islamic threat (e.g. "*γάγγραινα*" (="gangrene")).

Figure 60: Word Cloud of unique aggressive terms for "Muslims/Islam"

It is worth noting that verbal attacks targeted to MUSLIMS/ISLAM are triggered by geopolitical events such as the rise of ISIS or events related to violent practices or sexual abuse of specific population groups (women, children). For example, as it is

illustrated below in Fig. 61 new terms appear in the verbal attacks towards this group with the rise of ISIS in 2014 i.e. "*παραλογισμός*" (= "irrationalism"), "*γάγγραινα*" (= "gangrene"), and "*μάστιγα*" (= "plague"). In addition, there is a significant increase in the use of other e.g. "*ισλαμοφασισμός*" (= "islamofascism"), "*βαρβαρότητα*" (= "barbarism"), and "*προπαγάνδα*" (= "propaganda").

Similarly, the verbal attacks captured against SYRIANS are triggered by specific violent events. In particular, the attacks target the violent practices of the Syrian rebels and ISIS in the context of the Syrian civil war rather than the Syrian refugees. As for the generic TG of REFUGEES, the few verbal attacks that were captured are mostly attempts to challenge their identity implying that they are illegal immigrants. This notion of "illegality" or "lawlessness" is also dominant in the case of the generic TG IMMIGRANTS, where the most frequent terms used to attack it are the words "*λαθρομετανάστες*" and "*λάθρο*". Given the generic nature of the two later TGs, in that they do not constitute specific ethnic groups with individual characteristics, no unique terms that are related to particular stereotypes were found.

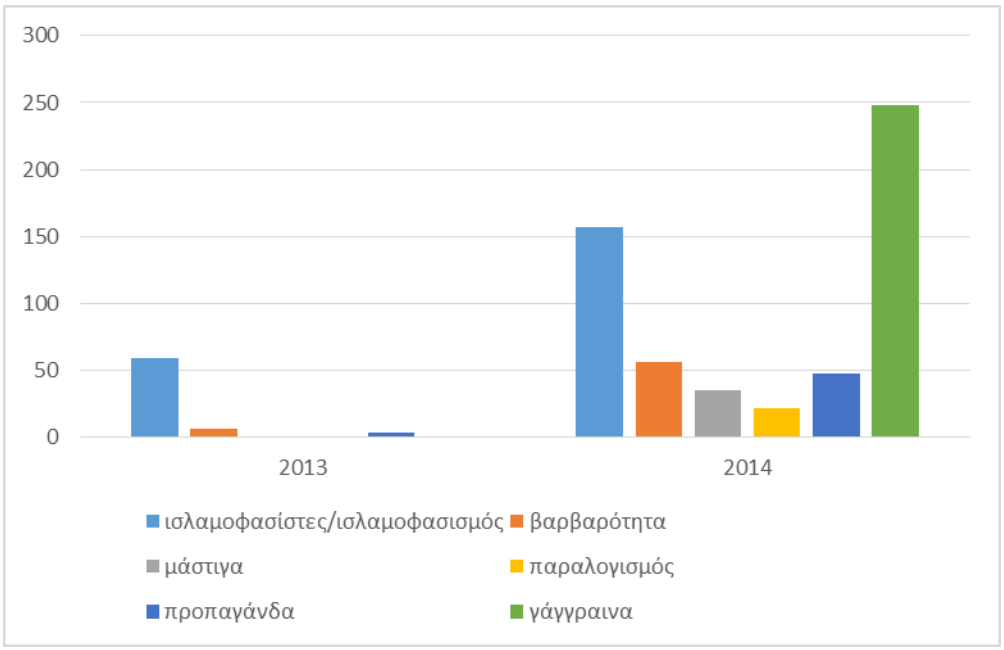


Figure 61: Counts of specific words in the verbal attacks against MUSLIMS/ISLAM for 2013 and 2014

On the other hand, the verbal attacks towards JEWS entail a perception of a particular enmity towards the Greek nation. As illustrated in Fig. 62, "*εχθρότητα*" (="hostility") is the most frequent term for tagging JEWS. Common themes in this group of

messages are the identification with the negative aspects of the banking system and global capitalism, as well as the frequent appeal to conspiracy theory elements (e.g. "*καταχθόνιος* (="sinister"), *δολοπλόκος*" (="conniver"), "*διπλοπροσωπία*" (="duplicitous"), "*καιροσκόπος*" (="opportunist")). In addition, Greece and banks are often tagged as "*Εβραιοκρατούμενη*" (= "owned by Jews"). These observations are in par with the conclusions that were drawn from the survey conducted by Antoniou et al. (2014) who established a correlation between conspiratorial thinking and ethnocentricism, and elaborated an interpretation of Greek anti-Semitism building on aspects of national identity and by employing the concept of victimhood.

Another dominant stereotype that is deeply rooted in Greek society and is reflected also in the verbal attacks against JEWS is the perception that they are avaricious (e.g. "*φραγκοφονιάς*" (="cheeseparing"), "*φιλάργυρος*" (="stingy"), "*τσιγκουνιά*" (="stinginess")). Anti-Semitic attitudes entail also notions of hate-speech e.g. the use of the term "*σαπούνι*" (="soap") in a biting derogatory manner with reference to soap made of Jewish victims by the Nazis.

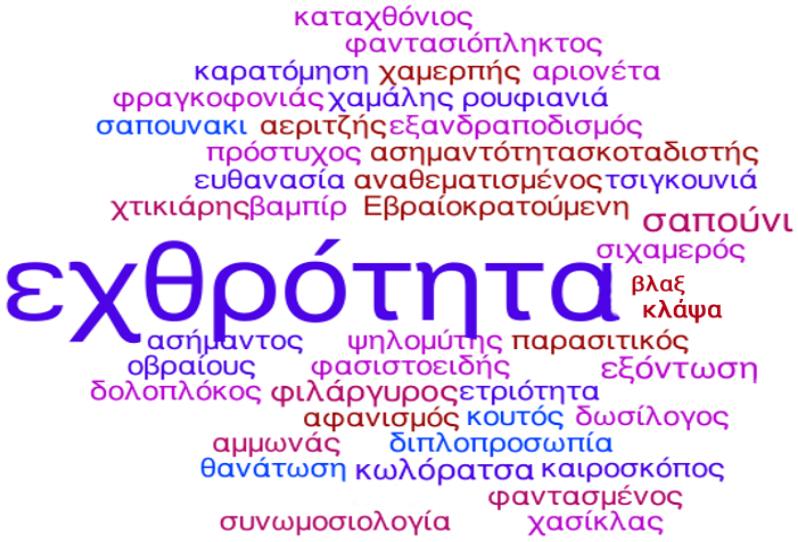


Figure 62: Word Cloud of unique aggressive terms for "Jews"

A perception of a particular enmity towards the Greek nation is also dominant in the verbal attacks against GERMANS, who play a central role in the Greek crisis. The popularity of the anti-German attitudes in Greece is also attested by a series of public opinion findings (Pew Global Attitudes Project, 2012). In the case of Twitter, as illustrated in Fig. 63, a variety of evaluative terms are used to stress out the harshness and hostility of GERMANS against Greeks e.g. "*σαδιστής*" (="sadist"), "*κακοήθεια*" (="malignancy"), "*ανέντιμος"* (="dishonest"), "*εκδικητικός*" (="vindictive"),

"*εμπαθής*" (= "malevolent"), "*μικροψυχία*" (="mean-spiritedness"), "*φρικαλέος*" (="unspeakable"). Memories and symbols of WWII and of Nazi occupation of Greece are also instrumentalized in the context of this victimization repertoire (e.g. "*ναζής*" (="Nazi", "*ράιχ*" (="Reich"), "*Once a Nazi, always Nazi!*"). These findings suggest a resurgence of the anti-German narration in the context of the anti-austerity ("anti-memorandum") discourse. Anti-German narration is considered to be the most prominent formulation of a victimization repertoire based on the foreign enemy concept and on the limited sovereignty discourse (Lialiouti and Bithymitris 2013).

In addition to the blame attribution patterns of the Greek crisis, the semantic analysis of the aggressive Tweets revealed also other types of dominant stereotypes deeply rooted in Greek society like the lack of delight stereotype, in that they do not know how to have fun and enjoy (e.g. "ξενέρωτος" (="killjoy")) and the poor aesthetics (bad taste) stereotype. In particular, the most frequent term in the construction of the image of GERMANS appears to be the word "πέδιλο" (="sandal"). The contentious Tweets comment on the usual stylistic choice of German tourists to wear Birkenstock sandals with socks. This stereotypical picture seems to be revoked and prominent during the crisis in order to condemn the inferiority of the Germans usually implying that they have no right to dictate austerity recipes to Greece which is a far superior culture.

Figure 63: Word Cloud of unique aggressive terms for "Germans"

3.4 DISCUSSION

To sum up, the results of the VA analysis feature JEWS and ALBANIANS at the core of the Greek Twitter xenophobic discourse. These findings are in-line with studies that highlight the strength of Greek anti-Semitism (ADL Global 100[51] survey) and with research findings (Voulgaris et al., 1995) suggesting that the image of foreigner as it was constructed in Greece during and after the first wave of migration flow (early 1990s-mid 1990s) was mainly associated with Balkan, and mainly Albanian, nationality. Among the interesting observations regarding the quantitative analysis of the results are the following:

- the most mentioned TGs in Twitter (e.g. REFUGEES, SYRIANS, GERMANS) are not also the most attacked ones,
- it is more likely to verbally attack groups of people framed as IMMIGRANTS rather than as REFUGEES due to the different connotations and implications of these two lexicalizations,
- the most obscene messages are targeted to PAKISTANIS and IMMIGRANTS,
- attacks that involve calls for physical extinction are far greater for JEWS than for any other group,
- calls for ouster/deportation are mostly targeted to ALBANIANS, PAKISTANIS and IMMIGRANTS,
- verbal attacks targeted to MUSLIMS/ISLAM seem to be mainly triggered by geopolitical events such as the rise of ISIS.

The qualitative analysis of the content of the verbal attacks expressed in Twitter confirms the existence of stereotypes and prejudices that are deeply rooted in Greek society. For example, the dominant stereotypes in the construction of the image of ALBANIANS are associated with "crime" and "cultural inferiority" indicating a continuity of the so-called stereotype of the Balkanian criminal. Crime and inferiority stereotypes are dominant also in the verbal attacks against MUSLIMS/ISLAM, but with rather different aspects. In particular, the attacks are often lexicalized through evaluative and dysphemistic terms of insult or abuse to debase core Islamic values, principles, practices, etc. indicating irrationalism/inferiority, sexist behavior and fanaticism. The inferiority stereotype is also dominant for PAKISTANIS; most of the verbal attacks against them are lexicalized through derogatory morphological variations of the nationality adjective.

[51] http://global100.adl.org/public/ADL-Global-100-Executive-Summary.pdf

In the case of JEWS, the verbal attacks entail a perception of a particular enmity towards the Greek nation and blame attribution patterns of the Greek crisis. Common themes in this group are the identification with the negative aspects of the banking system and global capitalism, as well as the frequent appeal to conspiracy theory elements. These observations coincide with surveys that establish a correlation between conspiratorial thinking and ethnocentricism, and elaborate an interpretation of Greek anti-Semitism building on aspects of national identity and by employing the concept of victimhood. This perception of vulnerability and victimization is reflected also in the verbal attacks against GERMANS, who play a central role in the Greek crisis. As for IMMIGRANTS and REFUGEES, given the generic nature of these TGs, in that they do not constitute specific ethnic groups with individual characteristics, no unique terms that are related to particular stereotypes were found. The content of the specific attacks as well as the effect of the recent refugee crisis on public beliefs is an open question for future research.

3.4.1 FURTHER INSIGHTS

The above presented results can provide some further insights regarding the nature of xenophobic behavior in terms of verbal aggression and also illuminate the possible reasons behind this complex social phenomenon. In particular, the VA analysis results could contribute in addressing the following two (more complex) RQs:

- ***Xenophobia in Greece: Vulnerability-driven or Superiority-based?***

The results illuminate two different dimensions usually correlated with the conceptualization of the phenomenon of xenophobia. On the one hand, verbal attacks directed against TGs like GERMANS and JEWS, who are considered more powerful, are related to the concept of vulnerability, which implies the perception of threat. The deeply established Anti-Semitism witnessed in Twitter is also connected to conspiracy theories and the global banking system. Vulnerability and victimization reflected in VA against GERMANS has mainly two dimensions; first, in relation to the ongoing economic crisis for which GERMANS are held responsible, and second as remnants of the second World War, the Nazis and the German occupation. As for the perception of vulnerability related to MUSLIMS, the verbal attacks that entail notions of Islamophobia are mostly triggered by geopolitical events such as the rise of ISIS and do not seem to constitute a core component of the Greek xenophobia, at least currently. On the other hand, dominance is directed against TGs who are thought of as inferior in socio-economic or cultural perspectives. Dominance implies superiority, in terms that people consider themselves superior in various possible ways to others. In

our case, TGs like ALBANIANS and PAKISTANI – the largest immigrant populations in Greece – are deemed inferior comparing to Greeks.

■ ***Xenophobia in Greece: Culturally-rooted or crisis-driven?***

Xenophobia in Greece, when examined in terms of Twitter VA towards specific predefined groups of interest, seems to be culturally-rooted and not crisis-driven. The results of the qualitative analysis of the VA twitter messages argue in favor of a continuity of deeply rooted stereotypes about specific TGs (e.g. ALBANIANS, JEWS, ROMANIANS, ROMA). These findings are in par with other surveys (Antoniou et al., 2014; Antoniou, Dinas, and Kosmidis, 2017) that interpret for example Greek anti-Semitism as a historically rooted and socially mediated narrative and not as a result of the economic crisis. The importance of the historical roots of xenophobia has been highlighted in studies like the one Baldwin-Edwards (2014) that emphasizes the elements of continuity between the various peaks in the expression of the phenomenon; critical historical events can create a legacy of xenophobia which conditions a society's perceptions and attitudes in future situations. In particular, for the case of Greece, Baldwin-Edwards goes back to the Asia Minor Catastrophe and to the mass influx of refugees in the 1920s to the Greek territory arguing that the hostile popular response against this migration wave was "structurally important" for the reception of Balkan (mainly Albanian) migration in the 1990s. Moreover, he explains how the negative attitudes of Greek society, politicians and mass media constructed and reproduced the stereotype of the "dangerous Albanian" (Baldwin-Edwards, 2014). However, the results indicate also the emergence of attacks that are associated with blame attribution patterns of the Greek crisis (e.g. GERMANS, JEWS). In other words, xenophobic attitudes may not be crisis-driven, but the economic crisis encourages the development of defensive nationalism and the perception of vulnerability.

A final, yet important remark concluding the VA analysis discussion is that the results coincide with those of the analysis of the physical attacks (Event Analysis) in the context of the XENO@GR. In other words, physical and verbal aggression -as indicators of xenophobic attitudes- seem to be addressed to the same targets (Pontiki et al., 2018). In particular, four out of five TGs that are mostly attacked both verbally and physically, are the same, namely JEWS, ALBANIANS, PAKISTANI and IMMIGRANTS. This observation is very interesting, as it may indicate possible correlations between verbal aggression and physical violence.

3.4.2 LIMITATIONS AND FUTURE WORK

The limitations of the approach and the methodology presented in this thesis can be summarized as follows:

- ***Validity***. The evaluation results (section 3.2.4.4) confirmed the expectations favoring a precision-oriented method, since it achieves high precision in all types of the extracted information. In other words, the great majority of the identified verbal attacks and targets of the attacks are correct. However, as it was also expected for a precision-oriented method, it suffers in terms of recall. In addition, the method is lexicon-based and not designed to address complex linguistic structures and devices (e.g. humor, irony, implicit intentions and calls for action). Hence, a significant amount of verbal attacks expressed in Greek Twitter has not been detected and explored.

- ***Representativeness***. Despite the large amount of the datasets that have been analyzed (almost 4.5 million Tweets), they constitute only a snapshot of the Greek Twitter for the period 2013-2016, and do not cover all Greek social media (e.g. Facebook, Instagram). Hence, the results presented in this thesis, as a single platform study, cannot capture the wider social ecology and diffusion (Tufekci, 2014). Furthermore, there is a broader issue regarding the representativeness of Social Media in general, given that they are used by a growing social science literature to study political and social phenomena (e.g. election forecasting, tracking political conversations). According to recent research findings (Mellon and Prosser, 2016), Twitter and Facebook users differ substantially from the general population in terms of demographics, political attitudes and political behavior. In the same vein, Blank and Lutz (2017) suggest that no social media platform is representative of the general population, hence social media data cannot be used to generalize to any population other than themselves.

- ***Completeness***. Xenophobia is a complex social phenomenon that reflects a deep-rooted form of fear and hostility towards the "other", who is perceived as a "stranger" (in Greek "xenos") to the group oneself belongs to. In the approach presented in this thesis the notion of "xenos" is limited to people with other than Greek nationality or origin, and further restricted to ten predefined TGs of interest based on specific criteria. In addition, xenophobia is examined as a violent practice in terms of verbal aggression that constitutes only one aspect of xenophobic attitudes. Hence, the results presented in this thesis do not reflect all

(aspects/types of) xenophobic attitudes expressed in Greek Twitter against all foreigners.

Given the above limitations, it is worth noting that computational methods (in this case SA) and social media data are not to replace traditional political and social science methods (e.g. polls), but rather to complement them by providing valuable insights that can contribute in an in-depth understanding of complex social phenomena and addressing specific RQs. As for the future work, it could be directed towards the following three directions:

- Test on more data, experiment with other types of techniques to deal with the limitations of the current method and improve the recall, design and implement the appropriate components/modules for dealing with more complex linguistic phenomena like metaphors, irony and humor.
- Identify the actors of the verbal attacks; a network analysis of the users that Tweet aggressive messages could help to spot specific groups or communities that promote xenophobic behavior.
- Model, capture and analyze other types of xenophobic attitudes expressed in Twitter and other social media platforms (e.g. self-reports, hate speech).

4. CONCLUSIONS

SA constitutes a key data analytics tool in many contexts and domains, since it helps to automatically detect and analyze public opinions, emotions, attitudes, and needs in massive amounts of unstructured data using NLP and Text Mining methods. The research activity of this PhD thesis focused on two types of opinionated user-generated content; evaluations expressed by customers about products and services and their aspects (ABSA) in particular domains of interest (restaurant and laptop reviews), and verbal attacks against predefined TGs of interest (e.g. refugees, immigrants) in the context of CSS, covering an industrial and a humanitarian use case of SA, respectively. This section provides a concluding summary of the research goals and the outcome of this thesis work in each case; a principled unified ABSA knowledge representation framework and respective English benchmark datasets (4.1), and a conceptual and computational framework for examining VA as an indicator of xenophobic attitudes in Greek Social Media (4.2).

4.1. ABSA

ABSA extends the typical SA setting with a more realistic assumption that negative or positive polarity is associated with specific aspects (or product features) rather than the whole text unit (Ma, Peng and Cambria, 2018). An ABSA method can analyze large amounts of unstructured texts and extract information not included in the user ratings that are available in some review sites. Within the last decades, several ABSA systems of this kind have been developed in a variety of domains. Depending on the approach, aspect could be a synonym for both fine- and coarse grained types of information. The basic definitions are summarized below:

- Aspects are coarse predefined categories (i.e. concept names) similar to rateable aspects (e.g. Ganu, Elhadad and Marian, 2009; McAuley, Leskovec and Jurafsky, 2012).
- Aspects are opinion targets i.e. all the targets towards which opinion can be expressed (e.g. Qiu et al., 2011).
- Aspects or features (Hu and Liu, {2004a, 2004b}) or facets (Mei et al., 2007) denote components/ parts, subcomponents of the target entity, and attributes of the target entity or its components (Liu, 2010).

For example, given sentence (1) from a customer review about a particular restaurant,

(1) "*The pizza was delicious but do not come here on an empty stomach.*"

the output of an ABSA method would be as follows for each of the above representations respectively:

- [FOOD: positive & FOOD: negative] or [FOOD: conflict]
- "pizza": positive
- pizza [+5], size [-3] [u][52]

This diversity in the decomposition of ABSA, as an information extraction task, resulted in different computational approaches –generating different types of output even for the same domains–, which were not directly comparable. In this context, publicly available ABSA benchmark datasets adopted different annotation schemes within different tasks. In addition, the available datasets were constructed to feed specific (types of) algorithms in each case. The annotations were typically presented as numbers of training and testing instances; no qualitative information was provided (e.g. main annotation problems, if and how they have been resolved), since no annotation guidelines of how to build a benchmark ABSA dataset were available. In other words, the computational framework conquered and somehow determined the conceptual framework.

In this setting, the research activity of this PhD thesis focused on the review of the scientific literature and the datasets in the field of ABSA as well as on collecting user-generated data about particular target entities of interest (laptops and restaurants). The aim of the research was to decompose ABSA as an information extraction task focusing on the intended meaning of the text and how it could be formalized into a conceptual knowledge representation framework, as well as to perform a systematic annotation study examining the different ways in which aspects are linguistically instantiated. The ultimate goal was to compile a set of detailed annotation guidelines and to construct gold-standard annotations fostering ABSA research towards more structured and meaningful output as well as to provide a common test bed (evaluation framework) for computational methods and techniques. To achieve these goals the research activity was decomposed in the following phases:

- *Building an annotation framework for ABSA.*

The first step was to examine how existing definitions are applied to datasets for both fine- and coarse-grained ABSA. The starting point was the restaurants and laptop reviews datasets of Pavlopoulos (2014). The laptops dataset contained annotations with ATE and ATP. The restaurant dataset was a subset of the dataset of Ganu,

[52] [u] denotes feature/aspect not appeared in the sentence (Hu and Liu, {2004a, 2004b}).

Elhadad and Marian (2009) that included annotations for coarse aspect categories (ACE) and overall sentence polarities; the dataset was modified to include annotations for aspect terms (ATE) occurring in the sentences, aspect term polarities (ATP), and aspect category-specific polarities (ACP). For example, sentence (1) was annotated as follows:

{ATE ="pizza", ATP ="conflict"} & {ACE =" FOOD", ACP ="conflict"}

The inspection of all the annotations revealed inconsistencies having to do, for example, with the boundaries of multi-word aspect terms (especially when they appeared in conjunctions or disjunctions) or with sentiment polarity ambiguity cases. Based on the systematic annotation study, a data-driven codebook was compiled in order to resolve problematic cases and to achieve consistency by providing definitions of each information unit that should be annotated along with examples and exceptions. Then, the datasets were extended with new unseen sentences and were annotated from scratch according to the guidelines. The restaurants reviews dataset consists of a total of 3841 sentences containing 4827 ATE&ATP annotations and 4738 ACE&ACP annotations. The laptops reviews dataset consists of a total of 3845 sentences containing 3012 ATE&ATP annotations. The proposed annotation guidelines and the annotated datasets were adopted from the SE-ABSA14 shared task (Pontiki et al., 2014) providing for the first time a common evaluation framework for (both coarse- and fine-grained) ABSA.

- *Redefining ABSA.*

Based on data analysis findings and the lessons learnt during the annotation process, at a second phase the ABSA problem has been formalized into a principled unified framework in which all the basic constituents of the expressed sentiments/opinions (i.e., aspects (= concepts), opinion target expressions, sentiment polarities) meet a set of specifications and are linked to each other within tuples. Within this new framework an aspect category is defined as a combination of an entity type E and an attribute type A. For example, sentence (1) is represented as:

{{Aspect category="FOOD#QUALITY", OTE="pizza", polarity="positive"}
{Aspect category="FOOD#STYLE_OPTIONS[53]", OTE=" pizza",
polarity="negative"}}

[53] According to the annotation schema opinions evaluating the food quantity (e.g. portions size) are assigned the label "FOOD#STYLE_OPTIONS"

This definition of aspect makes more explicit the difference between entities and the particular facets that are being evaluated. E can be the reviewed entity *e* itself (e.g. laptop), a part/component of it (e.g. battery or customer support), or another relevant entity (e.g. the manufacturer of *e*), while A is a particular attribute (e.g., durability, quality) of E. E and A are concept names (classes) from a given domain ontology and do not necessarily occur as terms in a sentence. In contrast to previous ABSA representations, in the current framework aspect terms correspond to explicit mentions of the entities E (e.g. service, pizza) or attributes A (e.g. price, quality) if any. A set of aspect category inventories and detailed annotation guidelines have been compiled in order to apply this new ABSA framework into the two domains (restaurants and laptops) that have been studied. Given that, correctly identifying the E, A pairs of a sentence and their polarities often requires examining a wider part or the whole review, new datasets (customer review texts) for each domain were collected and manually annotated following the new framework and guidelines. The reviews were annotated with sentence level but context-aware ABSA tuples. In particular, the restaurants dataset consists of a total of 440 review texts containing 3366 annotated ABSA tuples, and the laptops dataset consists of a total of 530 review texts containing 3710 annotations. The proposed new framework, the annotation guidelines and the datasets were adopted from the SE-ABSA15 shared task (Pontiki et al., 2015).

- *Extending ABSA.*

The third and last part of this thesis work in the context of ABSA was to extend the new framework towards text-level annotations, in order to provide also information of the overall rating of a text towards each discussed aspect; the sentence-level annotations were aggregated at the text level and with the appropriate modifications at the sentiment polarity labels –when needed–, the restaurants dataset was enriched with 1839 and the laptops dataset with 2627 text-level ABSA tuples. In addition, the proposed ABSA framework was extended to new domains (digital cameras, mobile phones, museums, telecommunications) and other than English languages (Arabic, Chinese, Dutch, French, Spanish, Russian, and Turkish) in the context of the SE-ABSA16 shared task (Pontiki et al., 2016). The annotation codebook was extended with annotation examples in more languages and domains. The use of the same annotation guidelines for domains addressed in different languages provided the opportunity also for the development and testing of cross-lingual or language-agnostic approaches.

The SE-ABSA task that was organized and ran in parallel with this thesis research activity (the author was one of the task organizers) for three years provided training datasets and a common evaluation framework for ABSA methods and attracted significant number of participants that contributed with a large number of submissions and system description papers. Furthermore, the annotation guidelines and the respective datasets that were generated during the three phases of the research activity of this thesis have been used also outside the SemEval challenge; the proposed annotation guidelines have been adopted for the creation of benchmark datasets in the same or new domains in other languages, while the generated datasets are being used for training and testing purposes by numerous researchers constituting the standard benchmarks for ABSA (details about the contribution are provided in section 1.2.1).

4.2. VA AND XENOPHOBIA

VA can manifold in a multitude of ways (e.g. flaming, cyberbullying, hate speech) in different contexts with somewhat different intentions and various effects on individuals, communities and social cohesion. Among others, it constitutes an important component in the study of xenophobia, since verbal attacks targeting foreigners can be indicative of xenophobic sentiments, attitudes and perceptions. Despite the numerous research efforts in automatically detecting and analyzing online VA, the user-generated content has been scarcely explored from the xenophobia standpoint at a large scale. Traditionally, xenophobia is examined using empirical and statistical methods; xenophobic attitudes are being measured using data coming from focus groups, interviews, and public sentiment polls using standard questions in order to capture opinions, emotions, perceptions and beliefs (e.g. Eurobarometer). The user-generated content available online constitutes a valuable source of information not only in terms of quantity (massive amounts of data), but also in terms the content itself, since the online disinhibition allows also aggressive forms of expression that cannot be captured by traditional methods that use face to face communications.

However, detecting online aggressive content is not a trivial task, since verbal attacks are shaped differently depending on individuals' intentions and strategic choices in language use (Tereszkiewicz, 2012). In addition, detecting and classifying an aggressive message is not enough; for example, an effect of hate speech depends on the originator, the content and the targeted one (Chetty and Alathur, 2018). Furthermore, there is a general diversity and lack of consensus in terminology of online VA that often results in overlap between several subtasks with the need for clear and operational definitions being stressed by several researchers (e.g. Waseem et al., 2017).

In this setting, this thesis presented a comprehensive overview of the key concepts, types, causes and effects of (offline and online) VA, and focused on verbal attacks expressed in Twitter against specific predefined TGs of interest proposing a data-driven and linguistically-inspired SA framework for examining VA as an indicator of online xenophobic attitudes. This notion of VA is closely related to hate speech, however, given the lack of a universally agreed definition as well as the legal implications of the term hate speech, the general term VA is used instead for explicitly stated verbal attacks targeting specific groups of foreigners in Greece. The ultimate goal was to build a KD that would help to formulate adequate responses to specific research questions concerning the nature and the evolution of the phenomenon of xenophobia as a violent practice in the Greek society in the context of the XENO@GR[54] project. In particular, the thesis goal was to address the following three RQs focusing on the amount, the type and the content of the verbal attacks, respectively:

- RQ1: Who are the main targets of Twitter verbal attacks?
- RQ2: Which are the main types of Twitter verbal attacks?
- RQ3: Are there stereotypes and prejudices against foreigners rooted deeply in the Greek society?

To this end, the research activity was decomposed to the following phases:

- *Data Collection*. For each TG of interest relevant Tweets were retrieved using related queries/keywords (4.490.572 Tweets in total covering the time period 2013-2016).
- *Explorative Analysis*. Samples of the collected data were explored focusing on the types of the verbal attacks (i.e. different aspects of VA) against the TGs as well as on the types of linguistic weapons used for the attacks (i.e. linguistic instantiations of VA messages).
- *Design of the VA Framework*. Based on literature review and data explorative analysis findings a data-driven VA framework was designed, where VAMs are classified into distinct categories based on specific linguistic criteria:
    - Their focus (i.e. distinguishing between VA utterances focusing on the target of the attack and VA utterances focusing on the attacker).
    - The type of linguistic weapon used for the attack (e.g. formal evaluations, obscene/dirty language, humor).
    - The content of the attack (e.g. threats/calls for physical violence or for deportation).

[54] http://xenophobia.ilsp.gr/?lang=el

- *Computational Analysis*. A rule-based VA analyzer was employed for the computational treatment of the proposed framework. The VA analyzer is a FST cascade implemented as a JAPE grammar in the GATE framework and comprises of a variety of lexical resources and grammars (sets of linguistic patterns).
- *Data Visualization*. The content analysis results, having been revised, were visualized in different ways (e.g. word clouds, graph charts) making them explorable, comprehensible and interpretable.

The quantitative analysis of the detected verbal attacks (RQ1) indicates that Antisemitism seems to be at the core of the Greek Twitter aggressive discourse against foreigners, while ALBANIANS and PAKISTANIS -that constitute the largest immigrant populations in Greece- constitute also the second and third most attacked TGs, respectively. In contrast, it seems that verbal attacks targeted to MUSLIMS/ISLAM are mainly triggered by geopolitical events such as the rise of ISIS or events related to violent practices or sexual abuse of specific population groups (women, children). In addition, the results confirm that it is more likely to verbally attack groups of people framed as IMMIGRANTS rather than REFUGEES due to the different connotations and implications of these two lexicalizations.

As for the types of the verbal attacks (RQ2), the TGs that are mostly attacked with messages negatively evaluating specific attributes of them are those of ALBANIANS and JEWS, the TGs that receive the most obscene messages are PAKISTANIS and IMMIGRANTS, while GERMANS receive the most ironic or humoristic messages. Calls for physical extinction and are far greater for JEWS than for any other group, while threat perception seems to prevail also for PAKISTANIS, ALBANIANS and IMMIGRANTS, according to the share of the calls for ouster/deportation for the specific groups.

The qualitative analysis of the results confirms the existence of stereotypes and prejudices against specific TGs that are deeply rooted in Greek society (RQ3) e.g. crime and cultural inferiority stereotypes in the construction of the image of ALBANIANS and ROMANIANS, inferiority and personal hygiene stereotypes in the case of PAKISTANI, brutal violence, sexist behavior, fanaticism and terrorism stereotypes for MUSLIMS/ISLAM, conspiracy theory elements for JEWS. In the case of JEWS and GERMANS the verbal attacks entail also a perception of a particular enmity towards the Greek nation and blame attribution patterns of the Greek crisis.

As for the further insights derived from the qualitative analysis of the results, they illuminate two different dimensions usually correlated to the conceptualization of the phenomenon of xenophobia; vulnerability and victimization are reflected in VA

against GERMANS and JEWS, who are considered more powerful, while dominance is directed against TGs who are thought of as inferior in socio-economic or cultural perspectives (e.g. ALBANIANS and PAKISTANI). Finally, the results indicate that xenophobia in Greece, when examined in terms of Twitter VA towards specific predefined TGs of interest, seems to be culturally-rooted and not crisis-driven.

Overall, the VA analysis results seem to coincide with other research findings (i.e. the results of the Event Analysis workflow, findings from empirical surveys). However, taking into account also the limitations discussed above (3.4.2) it is worth noting that SA and social media data are not to replace traditional political and social science methods, but rather to complement them by providing valuable insights that can contribute in an in-depth understanding of complex social phenomena and addressing specific RQs. In this setting, the work presented in this thesis constitutes a tangible example of how a carefully designed fine-grained SA approach can serve as a complementary research instrument in the context of CSS. Taking a step further from typical SA approaches, this thesis linked the analysis results to specific RQs including the critical step of their interpretation and presented an interdisciplinary end-to-end fine-grained SA approach. The resulting KD provides valuable quantitative and qualitative insights helping to study the formulation of VA in relation to specific TGs, and to measure and to monitor different aspects of VA as an important component of the manifestations of xenophobia in Greece.

Given the high correlation between verbal and physical aggression (Berkowitz, 1993; Hamilton and Hample, 2011; Laineste, 2012) -in that VA may escalate to physical violence- and the fact that physical and VA -as indicators of xenophobic attitudes- in the context of the XENO@GR project seem to be addressed to the same targets (Pontiki et al., 2018) the proposed framework could provide valuable insights also to policy makers. Furthermore, the proposed framework could be extended and applied to other languages enabling cross-countries studies and cross-cultural comparisons as well as to other targets (with the appropriate modifications) in order to capture on line VA in other contexts such as sexist or homophobic cyber-attacks.

REFERENCES

Aiken, M. and Waller, B. (2000). Flaming among first-time group support system users. *Information & Management*, 37(2), pp. 95–100.

Akhtar, M. S., Ekbal, A. and Bhattacharyya, P. (2016). Aspect based sentiment analysis in hindi: Resource creation and evaluation. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 23-28 May, 2016.

Akhtar, M. S., Sawant, P., Sen, S., Ekbal, A. and Bhattacharyya, P. (2018). Improving Word Embedding Coverage in Less-Resourced Languages through Multi-Linguality and Cross-Linguality. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 18(2), pp. 1–22.

Al-Ayyoub, M., Gigieh, A., Al-Qwaqenah, A., N. Al-Kabi, M., Talafhah, B. and Alsmadi, I. (2018). Aspect-Based Sentiment Analysis of Arabic Laptop Reviews. In: *Proceedings of the International Arab Conference on Information Technology*, Yassmine Hammamet, Tunisia, 22-24 December, 2017.

Alghunaim, A., Mohtarami, M., Cyphers, S. and Glass, J. (2015). A Vector Space Approach for Aspect Based Sentiment Analysis. In: *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, Denver, Colorado, 4-5 June, 2015, pp. 116–122.

Allen, C. (2013). Passing the Dinner Table Test. *SAGE Open*, 3(2).

Almagor, R. C. (2011). Fighting hate and bigotry on the internet. *Policy & Internet*, 3(3), pp. 1–28.

Alonzo, M. and Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems*, 36(3), pp. 205–213.

Al-Smadi, M., Qawasmeh, O., Talafha, B. and Quwaider, M. (2014). Human annotated Arabic dataset of book reviews for aspect based sentiment analysis. In: *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud (FiCloud)*, Rome, Italy, 24-26 August, 2015, pp. 726–730.

Anderson, A., Brossard, D., Scheufele, D., Xenos, M. and Ladwig, P. (2013). The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication*, 19(3), pp. 373–387.

Ang, R. and Goh, D. (2010). Cyberbullying among Adolescents: The Role of Affective and Cognitive Empathy, and Gender. *Child Psychiatry & Human Development*, 41(4), pp. 387–397.

Antoniou, G., Dinas, E., Kosmidis, S. and Saltiel, L. (2014). *Antisemitism in Greece: Evidence from a Representative Survey*. [Online] Available at: http://www.greekpublicpolicyforum.org/2014/09/antisemitism-in-greece-evidence-from.html

Antoniou, G., Dinas, E. and Kosmidis, S. (2017). Collective Victimhood and Social Prejudice: A Post-Holocaust Theory of Anti-Semitism. *SSRN Electronic Journal*.

Apidianaki, M., Tannier, X. and Richart, C. (2016). A Dataset for Aspect-Based Sentiment Analysis in French. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 23-28 May, 2016, pp. 1122–1126.

Austin, J. (1976). *How to do things with words*. Oxford: Clarendon Press.

Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. (2017). Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3-7 April, 2017, pp. 759–760.

Baldwin-Edwards, M. (2014). *Immigrants, Racism and the New Xenophobia of Greece's Immigration Policy*. [Online] Available at: https://ec.europa.eu/migrantintegration/librarydoc/immigrants-racism-and-the-new-xenophobia-ofgreeces-immigration-policy.

Bandura, A. (2004). Selective exercise of moral agency. In Thorkildsen, T.A. and Walberg, H.J. (Eds.) *Nurturing Morality* (pp. 35–57). Boston, MA: Kluwer Academic.

Bauman, S. (2009). Cyberbullying in a Rural Intermediate School: An Exploratory Study. *The Journal of Early Adolescence*, 30(6), pp.803–833.

Bekker, S. and Carlton, D. (1996). *Racism, xenophobia and ethnic conflict*. Durban: Indicator Press.

Ben-David, A. and Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, 10, pp. 1167–1193.

Benesch, S. (2014). *Countering dangerous speech to prevent mass violence during Kenya's 2013 elections*. [Online] Available at: https://dangerousspeech.org/kenya-2013/

Berkowitz, L. (1993). *Aggression*. New York: McGraw-Hill.

Blank, G. and Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61(7), pp. 741–756.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3, pp. 993–1022.

Blitzer, J., Dredze, M. and Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, 25-27 June, 2007, pp. 440–447.

Bourgonje, P., Moreno-Schneider, J., Srivastava, A. and Rehm, G. (2017). Automatic classification of abusive language and personal attacks in various forms of online communication. In: *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, Berlin, Germany, 13-14 September, 2017, pp. 180–191.

Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In: *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, 2-4 June, 2010, pp. 804–812.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3), pp. 209–227.

Bronwyn, H. (2002). A new pathology for a new South Africa? In D. Hook and G. Eagle (Eds), *Psychopathology and Social Prejudice* (pp. 169–184), Cape Town: University of Cape Town Press.

Brun, C., Popa, D. N. and Roux, C. (2014). XRCE: Hybrid Classification for Aspect-based Sentiment Analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 838–842.

Brun, C., Perez, J. and Roux, C. (2016). XRCE at SemEval-2016 Task 5: Feedbacked Ensemble Modeling on Syntactico-Semantic Knowledge for Aspect Based

Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 277–281.

Bucholtz, M. and Hall, K. (2005). Identity and interaction: a sociocultural linguistic approach. *Discourse Studies*, 7(4-5), pp. 585–614.

Burnap, P. and Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2), pp. 223–242.

Cambria, E., Schuller, B., Xia, Y. and Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, 28(2), pp. 15–21.

Caprara, G. V. and Pastorelli, C. (1989). Toward a reorientation of research on aggression. *Personality and Social Psychology*, 9(3), pp. 272–279.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In: *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, Amsterdam, Netherlands, 3-5 September, 2012, pp. 71–80.

Chernyshevich, M. (2014). IHS RD Belarus: Cross-domain extraction of product features using CRF. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 309–313.

Chetty, N. and Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, 40, pp. 108–118.

Chibbaro, J. S. (2007). School Counselors and the Cyberbully: Interventions and Implications. *Professional School Counseling*, 11(1), pp. 65–68.

Choi, Y. and Cardie, C. (2010). Hierarchical sequential learning for extracting opinions and their attributes. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, 11-16 July, 2010, pp. 269–274.

Chtouris, S., Zissi, A., Stalidis, G. and Rontos, K. (2014). Understanding Xenophobia in Greece: A Correspondence Analysis. *European Journal of Sociology*, 55(01), pp. 107–133.

Coenders, M., Lubbers, M. and Scheepers, P. (2003). *Majorities' Attitudes toward Minorities in European Union Member States*. Results from the Standard

Eurobarometers 1997-2000-2003. European Monitoring Centre on Racism and Xenophobia.

Colomb, G.C. and J.A. Simutis. (1996). Visible Conversation and Academic Inquiry: CMC in a Cultural Diverse Classroom. In Herring, S.C. (Ed.) *Computer-mediated Communication: Linguistic, Social and Cross-Cultural Perspectives* (pp. 203–222). Amsterdam: J. Benjamins.

Cosentino, V.J. (1994). Virtual Legality. *Byte*, 19(3).

Crush, J. (1996). A Bad Neighbour Policy? Migrant Labour and the New South Africa. *Southern African Report* 12 (1), pp. 3–5.

Crush, J. and Ramachandran, S. (2009). *Xenophobia, International Migration and Human Development*. United Nations Development Program Human Development Reports. [Online] Available at: http://goo.gl/OL1Pmb

Crush, J. and Ramachandran, S. (2014). *Migrant entrepreneurship, collective violence and xenophobia in South Africa*. Waterloo, ON: Southern African Migration Programme (SAMP).

Cunningham, H., Maynard, D. and Tablan, V. (2000). *JAPE: A Java annotation patterns engine*. Technical report, University of Sheffield, Department of Computer Science.

Dadvar, M. Trieschnigg, D. Ordelman, R. and de Jong, F. (2013). Improving cyberbullying detection with user context. In: *Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR'13)*, Moscow, Russia, 24-27 March, 2013, pp. 693–696.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In: Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM 2017), Montreal, Canada, 15-18 May, 2017, pp. 512–515.

Day, R. C. and Hamblin, R. L. (1964). Some Effects of Close and Punitive Styles of Supervision. *American Journal of Sociology*, 69(5), pp. 499–510.

Delanty, G. and OMahony, P. (2002). *Nationalism and social theory: Modernity and recalcitrance of the Nation*. London: Sage.

Dilawar, N., Majeed, H., Beg, M. O., Ejaz, N., Muhammad, K., Mehmood, I. and Nam, Y. (2018). Understanding Citizen Issues through Reviews: A Step towards Data Informed Planning in Smart Cities. *Applied Sciences*, 8(9).

Dinakar, K., Jones, B., Havasi, C., Lieberman, H. and Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems*, 2(3), pp. 1–30.

Ding, X., Liu, B. and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining* (WSDM '08), Palo Alto, California, 11-12 February, 2008, pp. 231–240.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V. and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In: *Proceedings of the 24th International Conference on World Wide Web (WWW 2015)*, Florence, Italy, 18-22 May, 2015, pp. 29–30.

Do, H. H., Prasad, P., Maag, A. and Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, 118, pp. 272–299.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M. and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, Baltimore, Maryland, 22-27 July, 2014, pp. 49–54.

Dong, X. and de Melo, G. (2018). A Helping Hand: Transfer Learning for Deep Sentiment Analysis. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia, 15-20 July, 2018, pp. 2524–2534.

Dubrovsky, V.J., Kiesler, S. and Sethna, B.N. (1991). The Equalization Phenomenon: Status Effects in Computer-mediated and Face-to-face Decision-making Groups. *Human Computer Interaction* 6(2), pp. 119–146.

Dvorak, J.C. (1994). The Flaming of Madison Ave. *Marketing Computers* 14, 22.

Ellinas, A. A. (2013). The Rise of Golden Dawn: The New Face of the Far Right in Greece. *South European Society and Politics*, 18, pp. 543–565.

Ewart, C. K., Burnett, K. F. and Taylor, C. B. (1983). Communication Behaviors That Affect Blood Pressure. *Behavior Modification*, 7(3), pp. 331–344.

Farra N., McKeown K. and Habash N. (2015). Annotating targets of opinions in Arabic using crowd-sourcing. In: *Proceedings of the 2nd Workshop on Arabic Natural Language Processing*, Beijing, China, 30 July, 2015, pp. 89–98.

Feshbach, S. (1970). Aggression. In Mussen, I. H. (Ed.), *Cannichael's manual* of *child psychology* (pp. 159-259). New York Wiley.

Galariotis, G., Papanikolaou, K., Georgiadou, V., Kafe, A., Lialiouti, Z., Papageorgiou, H., Pontiki, M. and Pappas, D. (2016). Xenophobia in Greece: A Computational Social Science Approach. *Poster presented at the 3rd Computational Social Science Winter Symposium 2016*, Cologne, Germany.

Ganu, G., Elhadad, N. and Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In Proceedings of the *12th International Workshop on the Web and Databases (WebDB 2009),* Providence, Rhode Island, USA, 28 June, 2009, pp. 1–6.

Gentry, W. D. (1972). Biracial Aggression: I. Effect of Verbal Attack and Sex of Victim. *The Journal of Social Psychology,* 88(1), pp. 75–82.

Georgiadou, V. (2015). *Antisemitism in Greece: Concerns and considerations*. In: Antisemitism in Greece. Athens: British Embassy.

Georgiadou, V., Lialiouti, Z., Kafe, A., Galariotis. I. and Voulgaris, Y. (2017). *The historical evolution of Xenophobia in Greece: A computational approach*. (Project deliverable). [Online] Available at: http://xenophobia.ilsp.gr/wp-includes/customContent/deliverables_pdf/Deliverable_2.1.pdf.

Gerring, J. (2009). *A Decoupage of Violence: The Harmonization of Collective Violence Theories* (M.A. Thesis). The University of North Carolina at Chapel Hill.

Geschiere, P. (2009). *The perils of belonging: Autochthony, citizenship, and exclusion in Africa and Europe*. Chicago: Univ. of Chicago Press.

Gitari, N. D., Zhang, Z., Damien, H. and Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering,* 10(4), pp. 215–230.

Glass, D. C., Krakoff, L. R., Contrada, R., Hilton, W. F., Kehoe, K., Mannucci, E. G. and Elting, E. (1980). Effect of Harassment and Competition upon Cardiovascular and Plasma Catecholamine Responses in Type A and Type B Individuals. *Psychophysiology,* 17(5), pp. 453–463.

Goldstein, D. and Rosenbaum, A. (1985). An evaluation of self-esteem of martially violent men. *Family Relations,* 34, pp. 425-428.

Graumann, C. F. (1998). Verbal Discrimination: A Neglected Chapter in the Social Psychology of Aggression. *Journal for the Theory of Social Behaviour*, 28(1), pp. 41–61.

Greenberg, B. S. (1976). The effects of language intensity modification on perceived verbal aggressiveness. *Communication Monographs*, 43(2), pp. 130–139.

Gunes, O. (2016). Aspect term and opinion target extraction from web product reviews using semi-markov conditional random fields with word embeddings as features. In: *Proceedings of the 6th International Conference on Web-Intelligence, Mining and Semantics*, Nimes, France, 13-15 June, 2016.

Gupta, S. and Waseem, Z. (2018). *A Comparative Study of Embeddings Methods for Hate Speech Detection from Tweets*. [Online] Available at: http://noisy-text.github.io/2018/pdf/31_abstract.pdf

Hågensen, L. (2014). *Understanding the Causes and the Nature of Xenophobia in South Africa: A Case Study of De Doorns*. (Thesis). University of Stellenbosch. Thesis.

Hamdan, H., Bellot, P. and Bechet, F. (2015). Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 753–758.

Hamilton, M. and Hample, D. (2011). Testing Hierarchical Models of Argumentativeness and Verbal Aggressiveness. *Communication Methods and Measures*, 5(3), pp. 250–273.

Hasanuzzaman, M. Dias, G. and Way, A. (2017). Demographic Word Embeddings for Racism Detection on Twitter. In: *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Taipei, Taiwan, 27 November – 1 December, 2017, pp. 926–936.

Hasib, T. and Rahin, S. A. (2017). *Aspect-based sentiment analysis using SemEval and Amazon datasets* (Thesis Report), BRAC University, [Online] Available at: http://dspace.bracu.ac.bd/xmlui/handle/10361/9542.

Haward, L. R. C. (1958). The effect of chlorpromazine on verbal aggression. *Indian Journal of Physiology and Pharmacology*, 2, pp. 367–373.

Hee, C. V., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., Pauw, G.D., Daelemans, W. and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *Plos One,* 13(10).

Hee, C. V., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W. and Hoste, V. (2015). Automatic detection and prevention of cyberbullying. In: *Proceedings of the 2015 International Conference on Human and Social Analytics*, St. Julians, Malta, 11-16 October, 2015, pp. 13–18.

Hinduja, S. and Patchin, J. W. (2009). *Bullying beyond the schoolyard*. Thousand Oaks, CA: Corwin.

Hjerm, M. (1998). National Identities, National Pride and Xenophobia: A Comparison of Four Western Countries. *Acta Sociologica*, 41(4), pp. 335–347.

Hoff, D. L. and Mitchell, S. N. (2009). Cyberbullying: Causes, effects, and remedies. *Journal of Educational Administration,* 47(5), pp. 652–665.

Hoffman, P. (1984). Psychological Abuse of Women by Spouses and Live-In Lovers. *Women & Therapy*, 3(1), pp. 37–49.

Hollomotz, A. (2013). Disability, oppression, and violence: Towards a sociological explanation. *Sociology*, 47(3), pp. 477–493.

Holloway, R. (1974). *Primate aggression, territoriality, and xenophobia: A comparative perspective*. New York: Academic Press.

Hu, M. and Liu, B. (2004a). Mining opinion features in customer reviews. In: *Proceedings of the 19th National Conference on Artificial Intelligence*, San Jose, California, 2004, 25-29 July, 2004, pages 755–760.

Hu, M and Liu, B. (2004b). Mining and summarizing customer reviews. In: *Proceedings of 10th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, 22-25 August, 2204, pp. 168–177.

Infante, D. A. and Wigley, C. J. (1986). Verbal aggressiveness: An interpersonal model and measure. *Communication Monographs,* 53(1), pp. 61–69.

Infante, D. (1987). Aggressiveness. In McCroskey, J.C. and Daly, J.A. (Eds.), *Personality and interpersonal communication* (pp. 157–192). Sage, Newbury Park.

Infante, D. A., Sabourin, T. C., Rudd, J. E. and Shannon, E. A. (1990). Verbal aggression in violent and nonviolent marital disputes. *Communication Quarterly*, 38(4), pp. 361–371.

Infante, D. A., Riddle, B. L., Horvath, C. L. and Tumlin, S. A. (1992). Verbal aggressiveness: Messages and reasons. *Communication Quarterly,* 40(2), pp. 116–126.

Infante, D. A. and Rancer, A. S. (1996). Argumentativeness and verbal aggressiveness: A review of recent theory and research. In Burleson, B. K. (Ed.), *Communication yearbook* (pp. 319–351). Thousand Oaks, CA: SAGE.

Jakob, N. and Gurevych, I. (2010). Extracting opinion targets in a single and cross-domain setting with conditional random fields. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, MIT, Massachusetts, 9-11 October, 2010, pp. 1035–1045.

Jacobs, J. B. and Potter, K. (1998). *Hate Crimes: Criminal Law and Identity Politics*. New York: Oxford University Press.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In: *Proceedings of the 2nd Workshop on NLP and Computational Social Science*, Vancouver, Canada, 3 August, 2017, pp 7–16.

Jin, W. and Ho, H.H. (2009). A novel lexicalized hmm-based learning framework for web opinion mining. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, Quebec, Canada, 14-18, June 14 - 18, 2009, pp. 465–472.

Jo, Y. and Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In: *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, Hong Kong, China, 9-12 February, 2011, pp. 815–824.

Johnson, N. A., Cooper, R. B. and Chin, W. W. (2009). Anger and flaming in computer-mediated negotiation among strangers. *Decision Support Systems,* 46(3), pp. 660–672.

Kaljahi, R. and Foster, J. (2018). Sentiment Expression Boundaries in Sentiment Polarity Classification. In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, 31 October, 2018, pp. 156–166.

Karatepe, O. M., Yorganci, I. and Haktanir, M. (2009). Outcomes of customer verbal aggression among hotel employees. *International Journal of Contemporary Hospitality Management,* 21(6), pp. 713–733.

Kayany, J. M. (1998). Contexts of uninhibited online behavior: Flaming in social newsgroups on UseNet. *Journal of the American Society for Information Science*, 49(12), pp. 1135–1141.

Kelley, M. L., Lewis, R. J. and Mason, T. B. (2015). Discrepant Alcohol Use, Intimate Partner Violence, and Relationship Adjustment among Lesbian Women and their Same-Sex Intimate Partners. *Journal of Family Violence*, 30(8), pp. 977–986.

Kerlinger, F. N. (1986). *Foundations of Behavioral Research*. 3rd Edition, Holt, Rinehart and Winston, New York.

Kessler, J. S., Eckert, M., Clark, L. and Nicolov, N. (2010). The 2010 ICWSM JDPA sentiment corpus for the automotive domain. In: *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC 2010)*, Washington, DC, 23-26 May, 2010, pp. 162–169.

Kiecolt-Glaser, J. K., Malarkey, W. B., Chee, M., Newton, T., Cacioppo, J. T., Mao, H. Y. and Glaser, R. (1993). Negative behavior during marital conflict is associated with immunological down-regulation. *Psychosomatic Medicine*, 55(5), pp. 395–409.

Kiesler, S., Siegel, J. and Mcguire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10), pp. 1123–1134.

Kim, S. M. and Hovy, E. (2004). Determining the sentiment of opinions. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 23-27 August, 2004, pp. 483–490.

Kinney, T. A. (1994). An Inductively Derived Typology of Verbal Aggression and Its Association to Distress. *Human Communication Research*, 21(2), pp. 183–222.

Kiritchenko, S., Zhu, X., Cherry, C. and Mohammad, S. (2014). NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 437–442.

Klinger, R. and Cimiano, P. (2014). The USAGE Review Corpus for Fine Grained Multi Lingual Opinion Analysis. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 26–31 May, 2014.

Klomek, A., Brunstein, S. A., and Gould, M. S. (2011). Bullying and Suicide: Detection and Intervention. Part of a special report: Suicide: Part 2. *Psychiatric Times*, 28(2), pp. 27–31.

Köffer, S., Riehle, D. M., Höhenberger, S. and Becker, J. (2018). Discussing the Value of Automatic Hate Speech Detection in Online Debates. In: *Proceedings of the 2018 Multikonferenz Wirtschaftsinformatik*, Lüneburg, Germany, 6-9 March, 2018.

Kok, S. D., Punt, L., Puttelaar, R. V., Ranta, K., Schouten, K. and Frasincar, F. (2018). Review-aggregated aspect-based sentiment analysis with ontology features. *Progress in Artificial Intelligence*, 7(4), pp. 295–306.

Korenman, J. and Wyatt, N. (1996). Group dynamics in an e-mail forum. *Pragmatics & Beyond New Series Computer-Mediated Communication*, pp. 225–242.

Kulig, J. C., Hall, B. L. and Kalischuk, R. G. (2008). Bullying perspectives among rural youth: a mixed methods approach. *Rural & Remote Health*, 8(2), pp. 1–11.

Kumar, A., Kohail, S., Kumar, A., Ekbal, A. and Biemann, C. (2016). IIT-TUDA at SemEval-2016 task 5: Beyond sentiment lexicon: Combining domain dependency and distributional semantics features for aspect based sentiment analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 1129–1135.

Kushwaha, A. and Chaudhary, S. (2017). Review highlights: opinion mining on reviews: a hybrid model for rule selection in aspect extraction. In: *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, Liverpool, United Kingdom, 17-18 October, 2017.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*, Bellevue, Washington, 14-18 July, 2013, pp. 1621–1622.

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, Williamstown, MA, USA, 28 June - 1 July, 2001, pp. 282–289.

Laineste, L. (2012). Verbal expressions of aggressiveness on the Estonian Internet. *ESTONIA AND POLAND: Creativity and Tradition in Cultural Communication Jokes and Their Relations*, pp. 205–220.

Lakkaraju, H., Socher, R., and Manning, C. (2014). Aspect Specific Sentiment Analysis using Hierarchical Deep Learning. In: *Proceedings of the 2014 Workshop on Deep Learning and Representation Learning (NIPS 2014)*, Montreal, Canada, 12 December, 2014.

Lambert, P. (2015). Aspect-Level Cross-lingual Sentiment Classification with Constrained SMT. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, Beijing, China, 26-31 July, 2015, pp. 781–787.

Lea, M., O'Shea, T., Fung, P. and Spears, R. (1992). "Flaming" in Computer-mediated communication: Observations, Explanations, Implications'. In Lea, M. (Ed.) *Contexts of Computer-Mediated Communication* (pp. 89–112). New York: Harvester Wheatsheaf.

Lee, S-H. and Kim, H-W. (2015). Why People Post Benevolent and Malicious Comments. *Communications of the ACM*, 58(11), pp.74–79.

Levenson, R. W. and Gottman, J. M. (1983). Marital interaction: Physiological linkage and affective exchange. *Journal of Personality and Social Psychology*, 45(3), 587–597.

Levenson, R. W. and Gottman, J. M. (1985). Physiological and affective predictors of change in relationship satisfaction. *Journal of Personality and Social Psychology*, 49(1), 85–94.

Li, X. and Lam, W. (2017). Deep multi-task learning for aspect term extraction with memory Interaction. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhagen, Denmark, 7-11 September, 2017, pp. 2886–2892.

Li, L., Liu, Y. and Zhou, A. Q. (2018). Hierarchical Attention Based Position-aware Network for Aspect-level Sentiment Analysis. 2018. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018)*, Brussels, Belgium, 31 October - 1 November, 2018, pp. 181–189.

Lialiouti, Z. and Bithymitris, G. (2013). The Nazis Strike Again': the concept of 'the German Enemy', party strategies and mass perceptions through the prism of the Greek economic crisis. In Karner, C. and Mertens, B. (Eds.) *The Use and Abuse of Memory: Interpreting World War II in Contemporary European Politics* (pp. 155–172). New Brunswick & London: Transaction Publishers.

Lialiouti, Z. and Bithymitris, G. (2017). A nation under attack: perceptions of enmity and victimhood in the context of the Greek crisis. *National Identities*, 19, pp. 53–71.

Lin, C. and He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, Hong Kong, China, 2-6 November, 2009, pp. 375–384.

Livingstone, S. and Smith, P. K. (2014). Annual Research Review: Harms experienced by child users of online and mobile technologies: The nature, prevalence and management of sexual and aggressive risks in the digital age. *Journal of Child Psychology and Psychiatry*, 55(6), pp. 635–654.

Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2, pp. 627–666.

Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael: Morgan & Claypool.

Ma, Y., Peng, H. and Cambria, E. (2018). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, New Orleans, Louisiana, 2-7 February, 2018, pp. 5876–5883.

Master, S. D. and Roy, M. K. (2000). Xenophobia and the European Union. *Comparative Politics*, 32 (4), pp. 419–436.

McAuley, J., Leskovec, J. and Jurafsky, D. (2012). Learning attitudes and attributes from multi-aspect reviews. In: *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM 2012)*, Brussels, Belgium, 10-13 December, 2012, pp. 1020–1025.

Makgopa, M. (2013). A critical evaluation of the key theoretical concepts and models on xenophobia as depicted in the drama of MS SerudusNaga ga di etelane: A qualitative approach. *South African Journal of African Languages*, 33(2), pp. 115–123.

Malmasi, S. and Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2), pp. 187–202.

Marler, P. (1976). On animal aggression: The roles of strangeness and familiarity. *American Psychologist*, 31(3), pp. 239–246.

Martin, J.R. and White, P.R.R. (2005). *The Language of Evaluation, Appraisal in English*. Palgrave Macmillan, London & New York.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C. X. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In: *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, 8-12 May, 2007, pp. 171–180.

Mellon, J. and Prosser, C. (2016). Twitter and Facebook are Not Representative of the General Population: Political Attitudes and Demographics of Social Media Users. *SSRN Electronic Journal.*

Michalopoulou, E., Tsartas, P., Yiannisopoulou, M., Kafetzis, P. and Manologlou, E. (1998). *Macedonia and the Balkans: Xenophobia and Development*. Athens, National Centre for Social Research-Alexandreia.

Mikolov, M., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems (NIPS)*, 26, pp. 3311–3119.

Misago, J. P., Freemantle, I. and Landau, L. B. (2015). *Protection rrom Xenophobia*. An Evaluation of UNHCR's Regional Office for Southern Africa's Xenophobia Related Programmes. [Online] Available at: https://www.unhcr.org/55cb153f9.pdf.

Mitchell, M., Aguilar, J., Wilson, T. and Durme, B. V. (2013). Open domain targeted sentiment. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, 18-21 October, 2013, pp. 1643–1654.

Moghaddam, S. and Ester, M. (2011). ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-11)*, Beijing, China, 24-28 July, 2011, pp. 665–674.

Moore, A. and Rayson, P. (2018). Bringing replication and reproduction together with generalisability in NLP: Three reproduction studies for Target Dependent Sentiment Analysis. In: Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New-Mexico, USA, 20-26 August, 2018.

Mudde, C. (2010). *The Relationship between Immigration and Nativism in Europe and North America*. [Online] Available at: https://works.bepress.com/cas_mudde/35/

Musto, C., Semeraro, G., de Gemmis, M. and Lops, P. (2016). Modeling Community Behavior through Semantic Analysis of Social Data: The Italian Hate Map Experience. In: *Proceedings of the 2016 Conference on User Modeling Adaptation*

and Personalization (UMAP 2016), Halifax, Canada, 13-16 July, 2016, pp. 307–308.

Nguyen, T. H. and Shirai, K. (2015). Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisbon, Portugal, 17-21 September, 2015, pp. 2509–2514.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y. (2016). Abusive Language Detection in Online User Content. In: *Proceedings of the International World Wide Web Conference Committee (IW3C2 2016)*, Montreal, Quebec, Canada, 11–15 April, 2016, pp. 145–153.

Notar, C.E., Padgett, S. and Roden, J. (2013). Cyberbullying: a review of the literature. *Univ. J. Educ. Res.*, 1 (1), pp. 1–9.

Nguyen, H. (2018). *A Joint Model of Term Extraction and Polarity Classification for Aspect-based Sentiment Analysis*. (M.A. Thesis). Japan Advanced Institute of Science and Technology.

Nockleby, J. T. (2000). Hate Speech. In *Encyclopedia of the American Constitution* (pp. 1277–1279). New York: Macmillan.

O' Sullivan, P.B. and Flanagin, A. (2003). Reconceptualizing 'flaming' and other problematic communication. *New Media and Society*, 5(1), pp. 67–93.

Ouyang, Z. and Su, J. (2018). Dependency Parsing and Attention Network for Aspect-Level Sentiment Classification. In: *Proceedings of the 7th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC 2018)*, Hohhot, China, 26-30 August, 2018. pp. 391–403.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Pennsylvania, Philadelphia, 6-7 July, 2002, pp. 79–86.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2), pp. 259–283.

Papageorgiou, H., Prokopidis, P., Demiros, I., Giouli, V., Konstantinidis, A., and Piperidis S. (2002). Multi–level XML–based Corpus Annotation. In: *Proceedings of*

the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain, 29-31 May, 2002, pp. 1723–1728.

Papanikolaou, K., Papageorgiou, H., Papasarantopoulos, N., Stathopoulou, T. and Papastefanatos, G. (2016). "Just the Facts" with PALOMAR: Detecting Protest Events in Media Outlets and Twitter. In: *Proceedings of the 10th International AAAI Conference on Web and Social Media*, Cologne, Germany, 17-20 May, 2016.

Parks, M. and Floyd, K. (1996). Making Friends in Cyberspace. *Journal of Communication* 46(1), pp. 80–97.

Parrott, D. J. and Giancola, P. R. (2006). The effect of past-year heavy drinking on alcohol-related aggression. *Journal of Studies on Alcohol*, 67(1), pp. 122–130.

Patchin, J. W. and Hinduja, S. (2012). *Cyberbullying prevention and response: Expert perspectives*. Routledge, Taylor & Francis Group.

Pavlopoulos, J. (2014). *Aspect based sentiment analysis*. (PhD Thesis). Dept. of Informatics, Athens University of Economics and Business, Greece.

Pennington, J., Socher, R. and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, 25-29 October, 2014, pp. 1532–1543.

Pinsonneault, A. and Heppel, N. (1998). Anonymity in Group Support Systems Research: A New Conceptualization, Measure, and Contingency Framework. *Journal of Management Information Systems*, 14(3), pp. 89–108.

Piryani, R., Gupta, V. and Singh, V. K. (2018). Generating Aspect-based Extractive Opinion Summary: Drawing Inferences from Social Media Texts. *Computación Y Sistemas*, 22(1).

Pitsilis, G. K., Ramampiaro, H. and Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, 48(12), pp. 4730–4742.

Pontiki, M., Aggelou, Z. and Papageorgiou, H. (2013). Sentiment Analysis: Building Bilingual Lexical Resources. In: *Proceedings of the 13th International conference on Greek Linguistics*, Rhodes Island, Greece, 26-29 September, 2013.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S. (2014). Semeval-2014 task 4: Aspect Based Sentiment Analysis. In:

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 23-24 August, 2014, pp. 27–35.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I. (2015). SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 486–495.

Pontiki, M. and Papageorgiou, H. (2015). Opinion mining and target extraction in Greek review texts. In: *Proceedings of the 12th International conference on Greek Linguistics*, Berlin, Germany, 16-19 September, 2015.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. and Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp.19–30.

Pontiki, M., Papanikolaou, K. and Papageorgiou, H. (2018). Exploring the Predominant Targets of Xenophobia-motivated behavior: A longitudinal study for Greece. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Natural Language Meets Journalism Workshop III*, Miyazaki, Japan, 7-12 May, 2018, pp. 11 –15.

Poria, S., Cambria, E. and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108, pp. 42–49.

Postmes, T., Spears, R. and Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3), pp. 341–371.

Prokopidis, P., Georgantopoulos, B., and Papageorgiou, H. (2011). A suite of NLP tools for Greek. In: *Proceedings of the 10th International Conference of Greek Linguistics*, Komotini, Greece, 1-4 September, 2011, pp. 373–383.

Qiu, G., Liu, B., Bu, J. and Chen, C. (2011). Opinion Word Expansion and Target Extraction through Double Propagation. *Computational Linguistics*, 37(1), pp. 9–27.

Rabiner, L. R. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. In Waibel, A. and Lee Alex Waibel and K-F. (Eds.) *Readings*

in speech recognition (pp. 267–296). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rachoene, M. and Oyedemi, T. (2015). From self-expression to social aggression: Cyberbullying culture among South African youth on Facebook. *Communicatio*, 41(3), pp. 302–319.

Rahman, M. D. and Dey, E. K. (2018). Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation. *Data*, 3(2), pp. 15-25.

Rana, T. A. and Cheah, Y. (2016). Aspect extraction in sentiment analysis: Comparative analysis and survey. *Artificial Intelligence Review*, 46(4), pp. 459–483.

Razavi, A. H., Inkpen, D., Uritsky, S. and Matwin, S. (2010). Offensive language detection using multi-level classification. In: *Proceedings of the 23rd Canadian Conference on Artificial Intelligence*, Ottawa, Canada, May 31 - June 2, 2010, pp. 16–27.

Reinig, B. A., Briggs, R. O. and Nunamaker, J. F. (1998). Flaming in the Electronic Classroom. *Journal of Management Information Systems*, 14(3), pp. 45–59.

Reynolds, V. and Vine, I. (1987). *The sociobiology of ethnocentrism: Evolutionary dimensions of xenophobia, discrimination, racism and nationalism*. London: Croom Helm.

Rodriguez, N. and Galeano, S. R. (2018). Shielding Google's language toxicity model against adversarial attacks. [Online] Available at: CoRR abs/1801.01828.

Rohner, R. P. and Rohner, E. C. (1980). Antecedents and consequences of parental rejection: A theory of emotional abuse. *Child Abuse and Neglect*, 4, pp. 189–198.

Rösner, L., and Krämer, N. C. (2016). Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Social Media & Society*, 2(3), pp. 69–94.

Ross, B., Rist, M., Carbonell, G. and Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In: *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, Bochum, Germany, 22 September, 2016.

Ruder, S., Ghaffari, P. and Breslin, J. G. (2016). INSIGHT-1 at SemEval-2016 Task 5: Deep Learning for Multilingual Aspect-based Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 330–336.

Rule, B. G. and Hewitt, L. S. (1971). Effects of thwarting on cardiac response and physical aggression. *Journal of Personality and Social Psychology*, 19(2), pp. 181–187.

Ruppenhofer, J., Klinger, R., Struß, J. M., Sonntag, J. and Wiegand, M. (2014). IGGSA Shared Tasks on German Sentiment Analysis (GESTALT). In: *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, Hildesheim, Germany, 6-7 October, 2014, pp. 164–173.

Saleem H. M., Dillon K., Benesch S. and Ruths, D. (2017) A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In: *Proceedings of the 1st Workshop on Text Analytics for Cybersecurity and Online Safety at LREC 2016*, Portorož, Slovenia, 23 May, 2016, pp. 1–10.

Saeidi, M., Bouchard, G., Liakata, M. and Riedel, S. (2016). SentiHood: targeted aspect based sentiment analysis dataset for urban neighborhoods. In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, Osaka, Japan, 11-17 December, 2016, pp. 1546–1556.

Saias, J. (2015). Sentiue: Target and Aspect based Sentiment Analysis in SemEval-2015 Task 12. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 767–771.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V. and Stranisci, M. (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, 7-12 May, 2018. pp. 2798–2805.

Scaffidi, C. (2006). *Application of a Probability-Based Algorithm to Extraction of Product Features from Online Reviews*. (Tech. Report CMU-ISRI-06-111), School of Computer Science, Carnegie Mellon University.

Schmidt, A. and Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In: *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, Valencia, Spain, 3-7 April, 2017, pp. 1–10.

Schouten, K. and Frasincar, F. (2016). Survey on Aspect-Level Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3), pp. 813–830.

Schrage, M. (1997). Mr. Bozo, meet Miss Courtesy Worm. *Computerworld* 31, 37.

Sebastian, R. J., Parke, R. D., Berkowitz, L. and West, S. G. (1978). Film Violence and Verbal Aggression: A Naturalistic Study. *Journal of Communication,* 28(3), pp. 164–171.

Seglow, J. (2016). Hate Speech, Dignity and Self-Respect. *Ethical Theory and Moral Practice,* 19(5), pp. 1103–1116.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F. and Weber, I. (2016). Analyzing the targets of hate in online social media. In: Proceedings of the of the 10th International Conference on Web and Social Media (ICWSM 2016), Cologne, Germany, 17–20 May, 2016, pp. 687–690.

Simão, A. M. V. DV., Ferreira, P., Francisco, S. M., Paulino, P. and Souza, S. B. (2018). Cyberbullying: Shaping the use of verbal aggression through normative moral beliefs and self-efficacy. *New Media & Society,* 20(12), pp. 4787–4806.

Simons, R. N. (2015). Addressing gender-based harassment in social networks: A call to action. In: *Proceedings of iConference 2015*. [Online], Available at: http://hdl.handle.net/2142/73743.

Somasundaran, S., Wiebe, J. and Ruppenhofer, J. (2008). Discourse Level Opinion Interpretation. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, United Kingdom, 18-22, August, 2008, pp. 801–808.

Sood, S., Antin, J. and Churchill, E. (2012). Profanity Use in Online Communities. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12, Austin, Texas, 5-10 May, 2012, pp. 1481–1490.

Spears, R. and Lea, M. (1992). Social influence and the influence of the social in computer-mediated communication. In Lea, M. (Ed.) *Contexts of Computer-mediated Communication* (pp. 30–65). London: Harvester-Wheatsheaf.

Srabstein, J. C., Berkman, B. E. and Pyntikova, E. (2008). Antibullying Legislation: A Public Health Perspective. *Journal of Adolescent Health,* 42(1), pp. 11–20.

Steinberger, J., Brychcín, T. and Konkol, M. (2014). Aspect-Level Sentiment Analysis in Czech. In: *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* WASSA 2014, Baltimore, Maryland, 22-27 July, 2014, pp. 24–30.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S. and Tsujii, J. (2012). BRAT: A Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of

the 13th Conference of the *European Chapter of the Association for Computational Linguistics (EACL 2012)*, Avignon, France, 23 – 27 April, 2012, pp. 102–107.

Stewart, D. and Bowers, L. (2012). Inpatient verbal aggression: Content, targets and patient characteristics. *Journal of Psychiatric and Mental Health Nursing*, 20(3), pp. 236–243.

Suler, J. (2004). The Online Disinhibition Effect. *CyberPsychology & Behavior*, 7(3), pp. 321–326.

Tafira, K. (2011). Is xenophobia racism? *Anthropology Southern Africa*, 34, pp. 114–121.

Tamchyna, A. and Veselovská, K. (2016). UFAL at SemEval-2016 Task 5: Recurrent Neural Networks for Sentence Classification. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 367–171.

Tangri, S. S., Burt, M. R. and Johnson, L. B. (1982). Sexual Harassment at Work: Three Explanatory Models. *Journal of Social Issues*, 38(4), pp. 33–54.

Tatum, B. D. (2001). Defining racism: Can we talk. In *Race, class, and gender in the United States: An integrated study* (pp. 100–107). New York, NY : Worth Publishers.

Teperoglou, E. and Tsatsanis, E. (2014). Dealignment, De-legitimation and the Implosion of the Two-Party System in Greece: The Earthquake Election of 6 May 2012. *Journal of Elections, Public Opinion and Parties*, 24(2), pp. 222–242.

Tereszkiewicz, A. (2012). Do Poles flame? Aggressiveness on Polish discussion groups and social networking sites. *Estonia and Poland. Creativity and tradition in cultural communication*, pp. 221–236.

Thet, T. T., Na, J. and Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6), pp. 823–848.

Thin, D. V., Nguye, V. D., Nguyen, K. V. and Nguyen, N. L. (2018). Deep Learning for Aspect Detection on Vietnamese Reviews. In: Proceedings of the 5th NAFOSTED Conference on Information and Computer Science (NICS), Ho Chi Minh city, Vietnam, 23-24 November, 2018.

Thompsen, P. A. and Foulger, D. A. (1996). Effects of pictographs and quoting on flaming in electronic mail. *Computers in Human Behavior*, 12(2), pp. 225–243.

Titov, I. and McDonald, R. T. (2008a). A joint model of text and aspect ratings for sentiment summarization. In: *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, Columbus, Ohio, 15-20 July, 2008, pp. 308–316.

Titov, I. and McDonald, R. T. (2008b.) Modeling online reviews with multi-grain topic models. In: *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, Beijing, China, 21–25 April, 2008, pp. 111–120.

Toh, Z. and Wang, W. (2014). DLIREC: Aspect Term Extraction and Term Polarity Classification System. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 235–240.

Toh, Z. and Su, J. (2015). NLANGP: Supervised Machine Learning System for Aspect Category Classification and Opinion Target Extraction. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 719–724.

Toh, Z. and Su, J. (2016). NLANGP at SemEval-2016 Task 5: Improving Aspect Based Sentiment Analysis using Neural Network Features. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 282–288.

Toprak, C., Jakob, N. and Gurevych, I. (2010). Sentence and expression level annotation of opinions in user-generated discourse. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, 11-16 July, 2010, pp. 575–584.

Törnberg, A. and Törnberg, P. (2016). Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Discourse, Context & Media*, 13, pp. 132–142.

Triandafyllidou, A. (2010). Dimensions and characteristics of immigration to Greece. In Triandafyllidou, A. and Maroukis, T. (Eds.) *Migration in 21st Century Greece*. Athens: Kritiki.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In: *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM '14)*, Ann Arbor, Michigan, 1-4 June, 2014.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, USA, 7-12 July, 2002, pp. 417–424.

UN Global Pulse (2017). *Social Media and Forced Displacement: Big Data Analytics & Machine-Learning*, UNHCR Innovation Service. [White Paper] Available at: https://www.unhcr.org/innovation/wp-content/uploads/2017/09/FINAL-White Paper.pdf

Vandebosch, H. and Cleemput, K. V. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society*, 11(8), pp. 1349–1371.

Veer, K. V., Ommundsen, R., Yakushko, O., Higler, L., Woelders, S. and Hagen, K. A. (2013). Psychometrically and qualitatively validating a cross-national cumulative measure of fear-based xenophobia. *Quality & Quantity*, 47(3), pp. 1429–1444.

Vicente, I. S., Saralegi, X. and Agerri, R. (2015). Elixa: A modular and flexible ABSA platform. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, 4-5 June, 2015, pp. 748–752.

Vigna, D. F., Cimino, A., Dell' Orletta, F., Petrocchi, M. and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on Facebook. In: *Proceedings of the 1st Italian Conference on Cybersecurity*, Venice, Italy, 17-20 January 2017, pp. 86–95.

Vissing, Y. M., Straus, M. A., Gelles, R. J. and Harrop, J. W. (1991). Verbal aggression by parents and psychosocial problems of children. *Child Abuse & Neglect*, 15(3), pp. 223–238.

Voggeser, B. J., Singh, R. K. and Göritz, A. S. (2018). Self-control in Online Discussions: Disinhibited Online Behavior as a Failure to Recognize Social Cues. *Frontiers in Psychology*, 8.

Vondráček, V., Horvai, I. and Študent, V. (1964). The aggression following the cephalic trauma. *Psychiatrie, Neurologie und Medizinische Psychologie*, 16, pp. 104–107.

Voulgaris, Y., Dodos, D., Kafetzis, P., Lyrintzis, C., Michalopoulou, K., Nikolakopoulos, E. and Tsoukalas, K. (1995). Perceiving and dealing with the "Other" in present day Greece. *Elliniki Epitheorissi Politikis Epistimis*, 5, pp. 81–100.

Voulgaris, Y. (2006). Globalization and national identity: Monitoring Greek culture today. *Portugese Journal of Social Sciences*, 5(2), pp. 141–153.

Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J. and Tounsi, L. (2014). DCU: Aspect-based Polarity Classification for SemEval Task 4. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, 23-24 August, 2014, pp. 223–229.

Wahl, K. (2002). Development of xenophobia and aggression. *International Journal and Applied Criminal Justice*, (26)2, pp. 247–256.

Walrave, M. and Heirman, W. (2010). Cyberbullying: Predicting Victimisation and Perpetration. *Children & Society*, 25(1), pp. 59–72.

Waltman, M. S. and Haas, J. W. (2010). *The Communication of Hate*. New York: Peter Lang Publishers.

Wang, Y., Huang, M., Zhu, X. and Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, Austin, Texas, 1-4 November, 2016, pp. 606–615.

Wang, S., Mazumder, S., Liu, B., Zhou, M. and Chang, Y. (2018). Target-Sensitive Memory Networks for Aspect Sentiment Classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia, 15-20 July, 2018, pp. 957–967.

Warner, W and Hirschberg, J. (2012). Detecting hate speech on the World Wide Web. In: *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, Montreal, Canada, 7 June, 2012, pp. 19–26.

Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In: *Proceedings of the NAACL 2016 Student Research Workshop*, San Diego, California, 13-15 June, 2016, pp. 88–93.

Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In: *Proceedings of the 1st Workshop on Abusive Language Online*, Vancouver, Canada, 4 August, 2017, pp. 78–84.

Weisband, S. P. (1992). Group discussion and first advocacy effects in computer-mediated and face-to-face decision making groups. *Organizational Behavior and Human Decision Processes*, 53(3), pp. 352–380.

Willard, N. (2005). *Educator's Guide to Cyberbullying Addressing the Harm Caused by Outline Social Cruelty*. [Online] Available at: http://www.asdk12.org/MiddleLink/AVB/bully_topics/EducatorsGuide_Cyberbullying.pdf.

Xenos, D., Theodorakakos, P., Pavlopoulos, J., Malakasiotis, P. and Androutsopoulos, I. (2016). AUEB-ABSA at Semeval-2016 task 5: Ensembles of classifiers and embeddings for Aspect Based Sentiment Analysis. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, San Diego, California, 16-17 June, 2016, pp. 312–317.

Xiang, G., Fan, B., Wang, L., Hong, J. and Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, Maui, Hawaii, USA, 29 October - 2 November, 2012, pp. 1980–1984.

Xiang, Y., He, H. and Zheng, J. (2018). Aspect Term Extraction Based on MFE-CRF. *Information* 2018, 9, pp. 198–213.

Xu, A. and Zhu, S. (2010). Filtering offensive language in online communities using grammatical relations. In: *Proceedings of the 7th Annual Collaboration, Electronic Messaging, AntiAbuse and Spam Conference*, Redmond, Washington, 13-14 July, 2010, pp. 1–10.

Xu, J-M., Jun, K-S., Zhu, X. and Bellmore, A. (2012). Learning from bullying traces in social media. In: *Proceedings* of the *2012 Conference* of the *North American Chapter* of the *Association for Computational Linguistics*: *Human Language Technologies (NAACL-HLT 2012)*, Montréal, Canada, 3-8 June, 2012, pp. 656–666.

Yaratan, H. and Uludag, O. (2012). The Impact of Verbal Aggression on Burnout: An Empirical Study on University Students. *Procedia - Social and Behavioral Sciences*, 46, pp. 41–46.

Yarowsky, D. (1994). Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 1994)*, Las Cruces, New Mexico, 27-30 June, 1994, pp. 88–95.

Yakushko, O. (2009). Xenophobia: Understanding the Roots and Consequences of Negative Attitudes toward Immigrants. *Educational Psychology Papers and Publications.* [Online] Available at: http://digitalcommons.unl.edu/edpsychpapers/90.

Ybarra, M. L. and Mitchell, K. J. (2004). Online aggressor/targets, aggressors, and targets: A comparison of associated youth characteristics. *Journal of Child Psychology and Psychiatry,* 45(7), pp. 1308–1316.

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M. and Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. In: *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16),* New York City, NY, 9-15 Jul, 2016, pp. 2979–2985.

Zhang, L., Liu, B., Lim, S. H. and O'Brien-Strain, E. (2010). Extracting and Ranking Product Features in Opinion Documents. In: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010),* Beijing, China, 23-27 August, 2010, pp. 1462–1470.

Zhang, L. and Liu, B. (2014). Aspect and Entity Extraction for Opinion Mining. *Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenges, and Opportunities,* pp. 1–40.

Zhu, P. and Qian, T. (2018). Enhanced Aspect Level Sentiment Classification with Auxiliary Memory. In: *Proceedings of the 27th International Conference on Computational Linguistics,* Santa Fe, New Mexico, 20-26 August, 2018, pp. 1077–1087.

Zhu, Y., Gao, X., Zhang, W., Liu, S. and Zhang, Y. (2018). A Bi-Directional LSTM-CNN Model with Attention for Aspect-Level Text Classification. *Future Internet,* 10(12), pp. 116–127.

Zhuang, L., Jing, F., yan Zhu, X. and Zhang, L. (2006). Movie review mining and summarization. In: *Proceedings of 15th ACM International Conference on Information and Knowledge Management (CIKM-2006),* Arlington, Virginia, 6-11 November, 2006, pp. 43–50.