Main Concepts, State of the Art and Future Research Questions in Sentiment Analysis

Orestes Appel, Francisco Chiclana, Jenny Carter

Centre for Computational Intelligence, Faculty of Technology, De Montfort University, The Gateway, Leicester, LE1 9BH, United Kingdom
orestes.appel@email.dmu.ac.uk , chiclana@dmu.ac.uk, jennyc@dmu.ac.uk

Abstract: This article has multiple objectives. First of all, the fundamental concepts and challenges of the research field known as Sentiment Analysis (SA) are presented. Secondly, a summary of a chronological account of the research performed in SA is provided as well as some bibliometric indicators that shed some light on the most frequently used techniques for addressing the central aspects of SA. The geographical locations of where the research took place are also given. In closing, it is argued that there is no hard evidence that fuzzy sets or hybrid approaches encompassing unsupervised learning, fuzzy sets and a solid psychological background of emotions could not be at least as effective as supervised learning techniques.

Keywords: sentiment analysis; opinion mining; social media; fuzzy sets; supervised learning; unsupervised learning

1 Introduction

Sentiment Analysis (SA) – or Opinion Mining (OM) – is a discipline that has seen a lot of activity since about 2000 [41]. The main reason for this, so it seems, is the proliferation of social media and its tools (e.g. Twitter, Facebook, LinkedIn, etc.), that has made the accessibility to information about *how people feel about things* more readily available to the masses. In addition, companies and other profit and non-profit organisations have accumulated a vast amount of data on how their employees or customers *feel* about the products and services they receive from the aforementioned organisations. Even Human Resources divisions are keen on understanding whether potential employees will be loyal and become a long-term member of the company or whether they would leave after receiving training and benefits.

In a way, a discipline that started as a research topic in Natural Language Processing (NLP) in Computer Science schools around the world, has now made a transition to other departments in academia and the industry, like those more related to business and management schools. The reason is very simple: everyone

wants to maximise their profits, and getting to understand what people think about oneself and one's company products could make a big difference business-wise.

According to some respected researchers [41, 44], there are many challenges lying ahead for SA as will be elaborated in the following sections. The reasons are many, but the fact that NLP has been around for a long time and has only focused on SA recently suggests the intrinsic difficulties with this discipline. Indeed, SA combines the application of NLP, Computational Linguistics and Text Analysis in its own way. A definition of SA attributed to Michelle de Haaff, in her article "Sentiment Analysis, Hard But Worth It" published in her CustomerThink (2010) *blog*, is as follows: ". . . *classifying the polarity of a given text at the document, sentence, or feature or aspect level, whether the expressed opinion in a document, a sentence or an entity feature or aspect is positive, negative, or neutral".* Michelle de Haaff goes ahead to loosely define Advanced SA, as the one that goes 'beyond polarity' sentiment classification, and it looks, for instance, at emotional states such as 'angry', 'sad', and 'happy'.

Understanding the emotions being conveyed by a given source – may it be a tweet, a document, a report, a blog, a segment of a politician's speech, etc. – has proven to be an important activity for humans. However, when volumes of *opinions* are very high, human processing becomes a challenge, hence the need for automated processes to extract sentiments from a variety of sources that keep growing in volume, complexity and diversity.

This article aims to present the fundamental concepts and challenges of the SA research field. To do this, an insight into the basics of SA and a summary of a chronological account of the research performed in SA will be provided. Additionally, some bibliometric indicators are included to shed some light on the techniques, and their geographical origins, most often used in addressing the central aspects of SA. Finally, the following research hypothesis is put forward: "there is no hard evidence to suggest that *fuzzy sets* or *hybrid approaches* encompassing unsupervised learning, fuzzy sets and a solid psychological background of emotions could not be as effective – or even better – as supervised learning methods."

2 Sentiment Analysis Basics

A sentiment (opinion) lexicon is defined as 'a list of positive and negative opinion words or sentiment words for English' [29]. It is assumed that such a lexicon could be built as well for any other language that one desires to use. According to Feldman [22], *the sentiment lexicon* "is the most important resource for most crucial analysis algorithms". Weichselbraun et al. [65] highlight the importance of context when producing sentiment lexicons. Indeed, they claim that "the limited ability of automated systems to resolve ambiguities and process context

information represents a major challenge". Thus, as the reader has probably guessed already, the importance of producing an accurate sentiment lexicon is that any polarity/sentiment evaluation to be performed will be based on the lexicon.

Opinions are easy to understand for human beings, but it is not that easy for a computer to have the same level of understanding. The notion of opinion as given by Liu [40] consists of the following items: (1) *Opinion targets*: entities and their features/aspects: (2) *Sentiments*: positive or negative; (3) *Opinion holders*: people who hold the opinions; (4) *Time*: when opinions are expressed.

Formally, an opinion is represented as a quintuple (e_j , a_{jk} , so_{ijkl} , h_i , t_l), where: e_j is a target entity, a_{jk} is an aspect/feature of the entity e_j; so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on feature a_{jk} of entity e_j at time t_l; so_{ijkl} is positive, negative or neutral, or more granular ratings; h_i is an opinion holder; and t_l is the time when the opinion is expressed. Liu provides a number of caveats to this definition [40], though: (i) although introduced using a product review, the definition is generic enough, in the sense that is applicable to other domains, e.g. politics, social events, services, topics, etc.; (ii) (e_j, a_{jk}) is also called the *opinion target* – opinion without knowing the target is of limited use; (iii) the five components in (e_j , a_{jk} , so_{ijkl} , h_i , t_l) must correspond to one another; (iv) the five components are essential – without any of them, it can be problematic in general.

Liu continues to describe the SA task as requiring to "structure the unstructured" [40] because Natural Language is regarded as unstructured data, and therefore the problem definition should provide *a structure* to the unstructured problem on the following three areas: (1) *Key tasks*: identify key tasks and their interrelationships; (2) *Common framework*: provide a common framework to unify different research directions; (3) *Understanding*: help us understand the problem better.

2.1. Key Tasks to Perform in SA

In general terms the problem of SA has two different abstraction aspects [40]: (1) *Opinion definition*, which has been addressed above, and (2) *Opinion summarisation*: opinions are subjective, and are needed from a significant number of people. Hence, some kind of summarisation will be required. The main components of the process of extracting sentiment from a given source, as taken from Kumar & Sebastian [36], are:

1. **Subjectivity Classification:** A document is a collection of sentences that may, or may not, express the author(s) opinion, which are called *subjective*. The sentences that are factual in nature are called *objective*. Usually, both types are present in a document. Subjectivity classification is "the task of classifying sentences as opinionated or not opinionated" [36].
2. **Sentiment Classification:** After finishing the task of identifying whether a text is opinionated, the polarity of the opinion must be found. Usually, classifying an opinion as either positive or negative is enough, i.e. *Values* = [*positive*,

negative]. However, sometime a multi-class classification might be used, with a range of values possible as exemplified with *Values* = [*extremely negative, negative, neutral, positive, extremely positive*].

3. **Complimentary Tasks: (a) Opinion Holder Extraction:** Depending on the type of application of SA, it would be necessary to identify the *opinion holder*. In some types of documents, there could be multiple opinion holders expressing their opinions about different subjects, hence the need to identify in those cases who is the opinion holder in every case. **(b) Object/Feature Extraction:** A task that may be necessary to execute – or not depending on the type of document being processed – is the identification of the target entity about which opinions are being issued. For instance, in social media it is not uncommon that a number of issues may be addressed (e.g. in blogs), so it is key to get to know about which object/feature opinions are being expressed.

2.2. Level of Analysis Issues

SA can be performed at many levels and at different complexity standards. According to Liu [41] there are commonly three different levels of analysis: Document level; Sentence level; and Entity Feature or Aspect level. Notice that these three levels of analysis are contained in Michelle de Haaff's definition of SA given before. Using a slightly different approach, Kumar & Sebastian [35] differentiate between the following levels of SA:

- **Document Level:** The whole document being analysed is the basic unit whose sentiment orientation will be determined.
- **Sentence level:** At this level, research focuses on detection of subjective sentences in a document containing a mixture of objective and subjective sentences.
- **Word Level:** At this level, usually the focus is to look for adjectives. However, verbs, adverbs and nouns could just as well convey a sense of subjectivity and carry opinions.
- **Feature based:** The common example to illustrate this level of SA is to consider a *review* containing positives and negatives of a product reflecting the reviewer liking and disliking of some of its features.

2.3. NLP Issues

We must always keep in mind that SA is a NLP problem and, consequently, many of the issues in NLP are also problems that must be addressed when dealing with SA problems. In [8], Bird, Loper and Klein address the need for a NLP toolkit that could be used efficiently in education, research and industry, and they provide the so-called Natural Language Toolkit (NLTK) as a platform for building *Python* programs to work with human language data. According to Liu [41], some of the

sub-problems that still are the object of further research attention by the NLP community are: (1) Coreference resolution; (2) Negation handling; (3) Word sense disambiguation; (4) Meaning extraction; and (5) Optimised parsing. Obviously, the proper resolution or any added improvements to any of these challenges above will have a positive effect in advancing the understanding of the SA problem. Dale [19] considers the following stages of analysis in processing natural language (in order of execution, from 1 to 5). The input to the process is *text* and the output is the *speaker's intended meaning*:

1. **Tokenization:** Converting a string of characters into words, symbols, sentences or other items conveying some sort of meaning, called *tokens.*

2. **Lexical analysis:** Usually deals with generating a *lexicon* and with applying *tagging* to the tokens already generated in the previous step. Most often, the tagging process is called *Parts of Speech (PoS) tagging.*

3. **Syntactic Analysis:** Provides a structure for every single sentence in a given text, including *parsing.*

4. **Semantic Analysis:** Aims to find the *literal meaning* of sentences or text.

5. **Pragmatic Analysis:** Attempts to determine the *meaning in context* of the sentence.

As *Tokenization* is rather mature, any affirmative contribution on the remaining 4 steps above would translate into improvements in the SA process.

2.4. Present Key Challenges in SA

The following list summarises the sub-topics that are considered to be key challenges for the SA discipline according to [12, 22, 29, 38, 39, 41, 51]: (a) Named Entity Recognition, (b) Anaphora Resolution, (c) Parsing, (d) Sarcasm & irony identification, (e) Subjectivity classification, (f) Polarity and graduality of opinions, (g) Use of abbreviations, poor spelling, punctuation or grammar, etc., (h) Sentiment (Opinion) Lexicon acquisition, (i) Negation handling, (j) Aspect-based & Comparative Sentiment Analysis, (k) Effective Classification of multiple opinions (aggregation).

3 SA Research Approaches

It is clear that the challenges present in the SA discipline are many. Liu [39, 41], Feldman [22], Pang & Lee [51] and Manning et al. [44], among others, consider that the future of the research on this area lies on exploring as many options as possible among the many challenges available and explore many sub-domains: customer reviews, politicians' blogs, marketing sites, company's opinion boards, etc. According to Cambria et al. [12] "mining opinions and sentiments from

natural language is challenging because it requires a deep understanding of the explicit and implicit, regular and irregular, and syntactical and semantic language rules. Sentiment analysis researchers struggle with NLP's unresolved problems: co-reference and anaphora resolution, negation handling, named-entity recognition, and word-sense disambiguation. Opinion mining is a very restricted NLP problem because the system only needs to understand the positive or negative sentiments of each sentence and the target entities or topics. Therefore, sentiment analysis is an opportunity for NLP researchers to make tangible progress on all fronts of NLP, and potentially have a huge practical impact."

Liu [39] claims that the main technical challenges for the multi-faceted problem that SA represents can be found among the following topics: (1) Object identification; (2) Feature extraction and synonym grouping; (3) Opinion orientation classification: and (4) Integration. The paragraph that follows will be utilised to define the topics of discussion.

(1) Yesterday, I bought a Nokia phone and my girlfriend bought a moto. (2) We called each other when we got home. (3) The voice on my phone was not clear. (4) The camera was good. (5) My girlfriend said that the sound on her phone was clear. (6) I wanted a phone with good voice quality. (7) So I was satisfied and returned the phone to BestBuy yesterday.

- **Object identification:** Discovering what the object is, about which an opinion has been provided. In the paragraph used as an example the objects are *Motorola*, abbreviated as 'moto' and *Nokia*. The noun 'BestBuy' corresponds to the name of the store; hence, it is *not* part of the comparison processed that the reviewer is providing and is not an object in terms of the products' comparison.
- **Feature extraction and synonym grouping:** The features commented on in the example are 'voice', 'sound' and 'camera'. According to Liu [39] "although there were attempts to solve this problem, it remains to be a major challenge". In addition, a feature can be referred to in different ways (i.e. 'voice' and 'sound' refer to the same feature in our example above.
- **Opinion orientation classification:** The objective of this task is to find out whether there is an opinion on a feature in a given sentence. If there is one, is it positive, negative or neutral? Here again, Liu [39] claims that the "existing approaches are based on supervised and unsupervised methods. One of the key issues is to identify opinion words and phrases, which are instrumental in sentiment analysis. The problem is that there are seemingly an unlimited number of expressions that people use to express opinions, and in a different domains they can be significantly different. Even in the same domain, the same word may indicate different opinions in different contexts".
- **Integration:** As the main objective of SA is to discover all quintuples (e_j, f_{jk}, so_{ijkl}, h_i, t_l), given as an input to an opinionated document, there is a need to integrate the above tasks, which is complex because the five pieces of information in the quintuple are to be matched. The quote just presented

corresponds to the definition of a direct opinion. In addition, Liu [38, 39] mentions that the fundamental problem here is that NLP techniques that still need improvement must be applied to resolve challenges like parsing, word-sense disambiguation, and co-reference resolution. In regard to the example provided, we observe the following issues: (i) understanding, depending on the context, what is meant by 'my phone' and 'her phone' in sentences (3) and (5); (ii) to which phone does the camera belong to?; (iii) in (4), "The camera was good", we do not have a pronoun and neither the sentence mentions a specific phone. According to Liu [38, 39] these are classical examples of **co-reference resolution**, the latter being a problem that despite the fact that the NLP community has studied it for a long time, it is still not accurately resolved.

In the following section, we will briefly describe the main approaches applied in the SA discipline.

3.1. Machine Learning

Machine Learning (ML) has played a fundamental role in NLP, and therefore it is extensively applied in the field of SA. Kumar & Sebastian [51] claim that "most researchers have defined the SA problem as essentially a *text classification problem* and machine learning techniques have proved their dexterity in resolving the sentiment analysis tasks". The main two learning approaches in ML are:

- *Supervised Learning:* "Machine Learning classification relies on the training set used, the available literature reports detail classifiers with high accuracy, but they are often tested on only one kind of sentiment source, mostly movie review, thus limiting the performance indication in more general cases" [36].
- *Unsupervised Learning:* "Use sentiment driven pattern to obtain labels for words and phrases" [36].

3.2. Fuzzy Sets/Logic Contribution to NLP and SA

In [20] Dzogang et al. go beyond the most common motivation for SA – to automate the classification of social media opinions, books reviews, film rankings, etc. – and attempt to "review methods taking account of *intrinsic psychological models components of graduality* as well as extrinsic components issued from computational intelligence approaches. In particular, beyond psychological models of sentiments that define affective states as multidimensional vectors in affective continuous spaces, we identify three components of graduality, namely composition or blending, intensity and inheritance". Basically, these authors perform a deeper analysis of the origins of emotions and sentiments and investigate among the technical tools at hand, which are closer to the nature of the problem being analysed. They highlight the nature of emotions and their graduality and fuzziness, and claim that "...it must be underlined that some

appraisal based approaches make use of graduality through fuzzy inference and fuzzy aggregation for processing affective mechanisms ambiguity and imprecision...". The caveat they make, though, is that these so-called appraisal-based methods are *not* great at *sentiment discrimination*. Nevertheless, these arguments make us think that even if not great for deep psychological and physiological analysis, the fact that fuzzy sets can be used successfully to model the ambiguity and imprecision of *affective states*, will make them an acceptable tool for modelling sentiments.

There have been some successful applications of Fuzzy Sets/Logic theory to both NLP and SA. In the literature, the use of Fuzzy Logic is found in *Anaphora Resolution* – given an expression S_i, its interpretation depends upon another expression S_j in context – in the work of Witte & Bergler [71]. Analysis of affect in text using fuzzy sets by means of the concept of *fuzzy semantic typing* can be found in [57]. Named Entity recognition has been addressed as well using fuzzy techniques in [32]. In summary, there seems to be evidence that a fuzzy approach could be applied in a number of sub-topics of sentiment analysis, and that some researchers are considering this research avenue worthy of further exploration.

3.3. Sentic Computing

As defined in Cambria & Hussein [11], "sentic computing is a multi-disciplinary approach to sentiment analysis that exploits both computers and social sciences to better recognise, interpret, and process opinions and sentiments over the Web. The approach specifically brings together lessons from both affective computing and common sense computing because in the field of opinion mining, not only common sense knowledge, but also emotional knowledge is important to grasp both the cognitive and affective information (termed semantics and sentics) associated with natural language opinions and sentiments". In a way, this approach is fairly new, and at this point it is uncertain whether this research avenue will be appealing enough to gain the attention of researchers in the near future.

4 State of the Art and Research time-line in SA

As SA sits at the confluence of several sub-disciplines – fundamentally NLP/ Computational Linguistics, Text Data Mining and AI – its origins cannot be tracked down to a specific date, but rather to a collection of moments in time that defines progress in the sub-areas mentioned above. Most of the important work in syntax and formal languages is attributed to Chomsky [15, 16] and his revolutionary work that occurred between the late 1950s and the late 1960s. Chomsky laid down the bases for modern languages & grammar theory, syntax theory and also for the concept of transformational grammar. In turn, these advances led to improvement in the automatic processing of syntax and grammars

by using productions and recursive calls. Parsing and Compiling theory, that today is taken for granted by many, was positively influenced by the work of Chomsky and others that followed. By 1872, Charles Darwin had already published his work '*The Expression of the Emotions in Man and Animals*', where he mainly addressed aspects of behaviours that are genetically determined. This is probably the first work related to determining the origin and characteristics of emotions. Many other authors in the Psychology camp have since then augmented the knowledge we have today about emotions as a fundamental human trait. This section aims to show how the SA discipline has evolved chronologically since 1970.

4.1. 1970 through 1979

The 1970s witnessed a lot of progress related to the refining of syntactic techniques and the generation of more advanced parsing and compiling ideas, together these resulted in more efficient algorithms. Making sure the proper parsing tree is generated is a fundamental step before more complex tasks can be started. In this arena, the work by Hopcroft [28] and Aho & Ullman [3, 4] is decisive, despite the fact that it concentrates in programming languages instead of natural languages, this is because it either brought rigour and formality to the parsing techniques or presented the works of others in a digestible way.

4.2. 1980 through 1989

It is possible to argue that no remarkable work applicable to SA was done until the 1980s. The work of Banfield [7] seems to have been instrumental in proposing the use of subjective and objective sentences as indicators, as well as in searching the text by means of simple queries. In 1983, Winograd [70] published work on language as a cognitive process that started a wave of further research into the cognitive aspects of emotions. In 1987, Ortony et al. published a landmark article [49] along with his 1988 book [50] that are a common reference to building an affective lexicon and have become important parts of the puzzle in SA.

4.3. 1990 through 1999

In 1990, Miller et al. gave to the research community *WordNet*, "a useful tool for computational linguistics and natural language processing" [46]. In 1991, Miller and Charles [47] discussed what they called the 'contextual correlates of semantic similarity', advancing the field even more when they researched the basis of semantic similarities in a given context, which would prove to be instrumental in the progress of the SA discipline. In 1994, the concept of Part-of-Speech was pushed forward and new ideas were put on the table in order to improve methods for part-of-speech tagging [10]. This technique would become key to properly identifying the different parts and component of sentences in order to build algorithms that would focus on extracting meaning and orientation out of

sentences. The use of statistical methods in Natural Language Parsing, Linguistics and more generically into NLP as a whole was brought by [1, 13, 45]. 'Text-Based Intelligent System' that focuses on the concept of directionality (e.g. is the agent in favour of, neutral, or opposed to a given event?) was presented in [27]. The authors claim that with their method, sentence meaning is mapped to a metaphorical model that is self-contained as no external references are required to find the directionality of a given sentence or paragraph. In [66, 67] the concept of extraction of subjective words is articulated properly, to the point that the author even proposed a method to determine the beliefs of the characters in the narrative, once the subjective terms have been identified. The possibility of predicting the semantic orientation of adjectives, which carry an important weight on determining the semantic orientation of phrases, was addressed in a landmark paper published in 1997 [25]. Towards the end of the 1990s some researchers started looking at the use of fuzzy reasoning in SA [35].

4.4. 2000 through 2014

From the start of 2000, the SA discipline starts an accelerated development process. Up to 2008, Pang & Lee's book [51] was considered the most complete work in the area. In [39], Liu addressed the complexities and multiple faces that this discipline can show, with the most updated version of the discipline later presented and published in 2012 [41]. We have mentioned before the importance of WordNet. In 2006, Esuli & Sebastiani published a lexical resource specific to SA [21], SentiWordNet, that assigns to each synset – sets of synonyms for groups of English words – of WordNet three sentiment scores: positivity, negativity, or objectivity. In 2013, Feldman attempted to bring SA to the front-page with their work [22], which has created a lot or additional attention in the research community. New techniques have been showing up steadily, with Cambria and collaborators [11, 12] proclaiming that new techniques, like the so-called Sentic Computing, could offer some new lights into the SA problem.

Most of the tools utilised in SA to resolve subjectivity identification and polarity extraction are based on some sort of Machine Learning technique. Most of the literature and bench marking established is based in Supervised Learning [36, 52]. However, some Unsupervised Learning techniques have been very successful as well, as it is unsupervised technique based on the PMI-IR algorithm that is used to estimate the semantic orientation of a phrase by measuring the similarity of pairs of words or phrases [60]. Alternative methods have been proposed, like the *Bootstrapping* Method for Building *Subjectivity Lexicons* for Languages with Scarce Resources [6], the techniques for generating a quality lexicon [59], the recognition of contextual polarity in [69] and the gradability of subjective sentences based on adjective orientation [26]. In all these cases, the focus of the research is at the sentence/phrase level. When one attempts to establish the orientation of the sentiment in a document, one is faced with the need for summarising somehow all the contents. Techniques to effectively summarise

opinions are addressed in [29]. On the same topic, Suanmali et al. [56] proposed a fuzzy logic based method for improving the summarisation of text, while Liu in [39] stressed one more time the important aspect of SA/OM of determining subjectivity by differentiating between *objective* and *subjective* sentences. Somehow less common, some researchers have looked into the possibility of applying semi-supervised learning methods like the one used for opinion summarization and classification for online product reviews [17]. Some hybrid methods started to flourish in the late 2000s. In [42], Liu & Tsou focus on using Supervised Machine Learning methods jointly with a qualified sentiment lexicon, while on the same token, Wiebe & Riloff [68] proposed a method to do simultaneously subjectivity analysis and information extraction based on their claim that by doing one enables the other one, somehow. We tend to believe that the future of research in SA is probably bound to lean towards hybrid methods.

In 2001, Subasic & Huettner [57] released the most important contribution we are aware of to the use of Fuzzy Sets/Logic principles in SA. Since then, others have followed their footsteps, but certainly, Fuzzy Sets Theory has been so far a bit of an outsider in the research field for SA. In 2010, Dzogang et al. published their influential article [20] in which, to be successful in SA and related disciplines, they depict how it is necessary, to understand two key factors: (a) the inclusion of the use of some fundamental emotion structure coming from the world of Psychology and (b) the further looking into the potential fitness of fuzzy sets to model *graduality* in a proper way. In 2011, the interest in fuzzy methods flares up again, with van der Heide et al. [61] addressing the topic of modelling affect through applying fuzzy logic; and by Kar & Mandalof's [33] study on using fuzzy logic for determining the strength of pre-established opinions in web reviews. In 2013 we see a cluster of interest in applying *fuzzy techniques* to SA [31], either to assist in resolving ambiguity in text as in [46], in using fuzzy sets to model sentiment classification at the sentence-level (for the Chinese languages) in [23], to drive the semantic understanding of general linguistic items [34], or a *fuzzy linguistic hedges*-based method for opinion mining in online user product reviews [18]. At the moment of writing this paper, this one seems to be most recent contribution of fuzzy methods to SA/OM.

For some time, the focus of identifying subjectivity by analysing adjectives was the standard. However, verbs and nouns are capable as well of conveying emotions and sentiments. The analysis of verbs to create a *verb lexicon* that would aid with establishing sentiments in opinion mining applications is explored in [43], while a combination of adjective, verbs and adverbs in an effort to improve subjectivity classification is presented in [42]. Using all of these three components of part-of-speech simultaneously may contribute to a more reliable subjectivity classification process. In 2012, Nguyen et al. [72] combine the new features with conventional ones obtained from already established research lines of work, and as such their method can somehow be considered as a *hybrid* approach. A system combining together concept-level sentiment analysis and opinion mining lexicon-

based and learning-based approaches is proposed in [48]. Dealing with metaphoric language is hard, and some researchers have spent some time suggesting how to address the problem [9, 53]. Recognising irony and sarcasm are tough topics as well, and some work has been done in this area [62]. These topics are very important when dealing with opinions, particularly when the text being analysed has politics content. A departure from most of the methods we have presented above, and an alternative possibility that some researchers are currently considering in SA is that of addressing the space of entity-related opinion detection and sentiment ontology trees [64].

Lotfi Zadeh put forward the concept of Precisiated Natural Language (PNL) for describing perceptions [73]. What Zadeh does mean by precisiated? As per [73], "...precisiated in the sense of making it possible to treat propositions drawn from a natural language as objects of computation?". Despite the fact that the idea is very tempting, not too much additional research has been conducted in PNL, as far as SA goes, although we believe that it puts forwards a concept that could possibly blossom in the SA arena.

4.5. Bibliometrics and References Distribution

This section attempts to 'see' the portion of work in SA that has been carried out using supervised machine learning methods versus those based on fuzzy sets theory. The years of publishing as well as the country where the work has been conducted are presented, for which a simple search based on the keywords *machine learning* or *fuzzy sets* in the larger context of SA was used. Although it is true that this search will exclude articles written before the term SA saw the light and became fully accepted – later on in the middle 2000s – it will, however, provide us with good indicators of the numbers of publications on the topic since the middle 2000s as well as the country where the research initiative was conducted and published. Likewise, articles indexed by other sub-topics of SA, like subjectivity classification or identification, polarity extraction, etc. could have been partially excluded, too. Nevertheless, we believe that based on the review of the literature carried out, the potential exclusion of those articles will not create a deviation in the results already obtained. For consistency, we are including only articles written in the English language. The sources used for the search is **Scopus,** a large abstract and citation database of peer-reviewed literature by Elsevier B.V., and the **Web of Knowledge** (WoK), an academic citation indexing and search service by Thomson Reuters. However, as the results obtained using both databases are equivalent, only the results obtained using Scopus are included below:

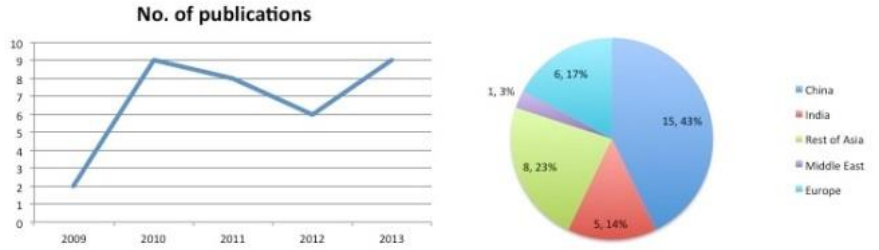

Figure 1
Research using keywords *Fuzzy Sets* and *Sentiment Analysis*

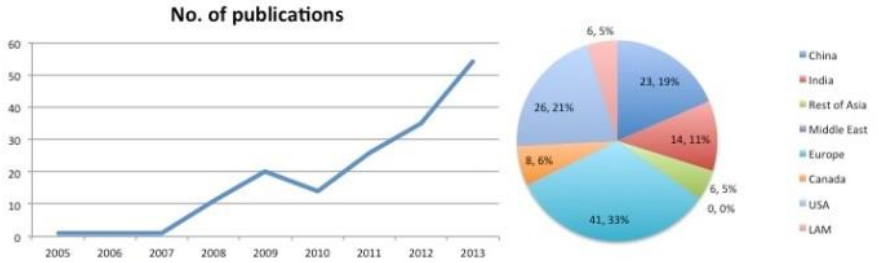

Figure 2
Research using keywords *Machine Learning* and *Sentiment Analysis*

As we take a closer look to the graphics provided above, we immediately note two characteristics:

1. Research in SA using Machine Learning (ML) techniques depicts a curve that shows primarily a clear trend towards sustained growth, whilst the one representing the utilisation of Fuzzy Sets (FS) shows no material growth and a clear lesser number of publications.

2. When we look at the countries where the articles have been published we notice that the utilisation of ML techniques – more specifically Supervised Learning – is high across the board, showing the USA, China, India and Europe as clear leaders. If we observe the utilisation of FS to model the SA problem, the first thing that we realise is that it is mainly an affair mostly pursued in China, India and Europe. In addition, research using ML techniques started earlier as well.

Why is this? One dares to venture that it is perhaps because a significant number of the researchers with a Computer Science/Mathematics educational background involved nowadays with researching SA/OM come from the Text Data Mining/Processing and NLP fields where the use of statistical methods have been a well-established tool for some time now. Hence, it would be natural to export the same knowledge, skills and techniques and apply them to a new domain that, nevertheless, is somehow related. Or is it perhaps that the utilisation of FS

techniques have been proven to be not successful? If so, how has *success* been measured? It is interesting to see that in the cases of China and India, and to a certain extent Europe as well, research efforts are present in both camps (ML and FS). However, for the USA, at least for the period of time chosen and the search keywords and data sources utilised, the focus is clearly on the ML camp, despite the fact that one of the most influential papers supporting the use of FS was written in 2001 in the USA [57]. One may think as well that two of the most reputed researches in SA/OM, Bing Liu and Bo Pang, have made ML their fundamental tool. For instance, Liu's early research was in data mining, Web mining and *machine learning*, fields in which he published abundantly (as appears in his Biography in [41]). At this point, we can only draw some conjectures, but we believe that it will not be completely nonsensical to think that the primary research interest of some authors may have migrated to this newer field of SA/OM. Moreover, the use of statistical techniques in NLP/Computational Linguistics are common and have been aptly utilised since at least 1996 [13, 14].

Fuzzy Sets have been used extensively to model uncertainty and ambiguity, traits that are undoubtedly inherent to Natural Languages and as a consequence part of the challenges inherited by SA/OM. Somehow, Fuzzy Sets may be seen as alien to the community of Linguistics, with the exception perhaps of the utilisation of Fuzzy Grammars [4]. We conclude then that there are a number of potential reasons that could explain why the use of Supervised Machine Learning techniques has been favoured. However, so far, we have not been able to find hard evidence that the utilisation of Fuzzy Sets, perhaps in combination with some other syntactic techniques and even Unsupervised Learning tools, could not yield favourable results. Bing Liu, one of the most world-wide recognised experts in the area of SA/OM and one the researchers that has attempted to push the limits in the field of SA, has mentioned that "we probably relied too much on Machine Learning" [38, 39, 41], when referring to how limited our understanding is about the SA problem, despite the recent progress that has been achieved.

As a result of the discussion presented in this section and other arguments to be presented in the next sections, we do believe there is merit in investigating further the potential use of *Fuzzy Sets* in the *Sentiment Analysis problem*; especially in the research sub-areas of *subjectivity, polarity and graduality* [20, 32, 61, 73, 74].

5 Potential Future Research Path in SA

Traditionally in ML we think of unsupervised, semi-supervised or supervised learning. Supervised learning, as we well know, relies heavily on training, which implies counting with the adequate data sets. To avoid, if possible, having to count on prior data for training purposes would also somehow, disqualify as well until certain extent the use of semi-supervised methods. In the context of SA, an

unsupervised strategy would rather "measure how far a word is inclined towards positive and negative" [63]. Somehow, this makes out of an unsupervised method a *semantic orientation approach* or a *lexicon-based method*.

Ultimately, the problem of SA/OM is basically a NLU/NLP problem with emphasis in finding when a sentence reveals an opinion – as opposed to a fact – and extracting the polarity of the opinion (Negative, Positive, Neutral, etc.). Being successful at determining if a sentence is objective or subjective will predetermine by far how accurate the establishing of polarity on subjectivity will be. Banea [6] claims that "the problem of distinguishing subjective versus objective instances has often proven to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification". This translates into the idea that getting the differentiation between objective and subjective sentences correctly would guide one through the right path.

Kanaga [32], in discussing ideas presented by Lotfi Zadeh in [74], says: "The semantics of natural languages and information analysis is best handled by the epistemic facet of Fuzzy Logic. In the epistemic facet, natural language is viewed as a system for describing perceptions and an important branch of the same is possibility theory and computational theory of perceptions". Hence, is it worthy to take a new look to Fuzzy Sets/Logic as a potential effective tool in SA/OM? Would it be helpful as well to stick to a strong psychological foundation of emotions and feelings to assist us in modelling the problem at hand? The recipe that we would like to pursue will include a solid foundation of emotions theory, unsupervised learning (semantic approach) and *fuzzy sets/logic* as fundamental components of a *hybrid approach* towards SA/OM.

5.1. Where shall We Go from Here?

Based on the information, references and discussions shown above, possible research directions that deviate from the current most followed path – Supervised Machine Learning – are suggested:

1. In SA the most utilised approach is text classification relying heavily on ML techniques; especially Support Vector Machine (SVM) and Naïve Bayes
2. Fuzzy Sets and Fuzzy Logic have been used to a lesser extent and the literature about it is rather reduced when compared to (1) above.
3. If determining subjectivity properly contributes to a more accurate polarity identification, then it is worth it to spend additional time on the topic, before attempting to conclude on polarity.
4. One of the main objections of the use of Fuzzy Logic/Sets in SA is [5] as follows: "...we can show that while the fuzzy models of emotion perform well for a series of cases that fit the described patterns, they remain weak at the time

of acquiring, combining and using new information". However, we believe that some of these shortfalls can be minimised by combining together fuzzy methods and some semantic and linguistic techniques. See, for example, the progress reported on acquiring new information in a given lexicon in Kruse et al. [35] (using neuro-fuzzy modelling) and Hüllermeier [30] (applying learning fuzzy rules).

5. Hatzivassiloglou et al. [25, 26] proposed a methodology to predict the semantic orientation of adjectives. This strategy – so it seems – could be extended to nouns, adverbs, and verbs, as discussed in [58]. As such, predicting the semantic orientation of certain parts of speech can greatly help on suggesting the semantic orientation of sentences and documents.
6. Grammatical dependencies may play a significant role in a proper understanding of a sentence. As quoted from [55]: "In any sentence, words are arranged in a proper sequence to communicate information. The complete meaning of a sentence is not only determined by the meaning of words, but also by the pattern in which words are arranged".
7. Supervised Learning (SL) has proven to be a strong classification technique. However, SL will depend enormously on the availability of training data. In a way, we would like to move towards a system that is less dependent on *pre-existing annotated data*. To rely more on the richness of fuzzy sets as a modelling apparatus – perhaps using *hedges* as well – and in syntactic analysis techniques.

5.2. Future Research Questions to be Addressed

The fundamental research question we are posing is whether a hybrid approach, combining together the *psychological foundations of emotions, linguistics tools, unsupervised learning and fuzzy sets/logic,* is well equipped to model subjectivity and polarity determination in SA/OM. By well equipped we mean for it to be capable of delivering the same or better results than the most commonly used techniques whilst remaining faithful to the original sources of emotions and to modelling tools that are akin to the inherent ambiguity present in natural languages. For simplicity, we will split this question into four sub-questions:

1. Is *Unsupervised Learning* capable of delivering similar accuracy to the one provided by *Supervised Learning* techniques in the determination of subjectivity in Sentiment Analysis?
2. Is *Fuzzy Reasoning* adequate to support subjectivity determination and to model polarity in SA by introducing gradualness (graduality)?
3. Can we represent with more accuracy sentiments expressed in natural languages by using as a bedrock concepts of emotions that originate in psychology [50]? Can we get closer to the heart of the matter by using this foundation and looking into the *cognitive model of emotions* or is doing so futile?

4. Is our model flexible enough to attempt to accommodate afterwards the recognition and understanding of metaphors? Can this be achieved without the use of supervised or semi-supervised machine learning approaches?

Is there going to be synergy among all these elements? Currently, most of research performed has been conducted using Supervised Methods in Machine Learning (mostly SVM, Naïve Bayes and others). Hence, our comparison base will be defined by the results already obtained using the latter methods. In a way, we must try to determine whether good results in the sub-questions will have an aggregated *positive* effect when all of them get combined together. The key performance indicators that will be chosen for the comparison will be decisive in understanding how successful the research journey has been.

Conclusions

In this article we have attempted to cover a few fundamental aspects related to Sentiment Analysis/Opinion Mining. Firstly, we wanted to address the basics of the topic and its main challenges. Secondly, we have provided a chronological account of the research that has been conducted to date as well as some bibliometric aspects showing a distribution of articles published based either on *machine learning* or *fuzzy sets* as the main tools to model the SA/OM sub-problem. Finally, we have ventured to suggest that there is not enough evidence to justify abandoning other potential research paths that may rely on hybrid mechanisms combining a number of foundations, strategies and techniques.

References

[1] S. Abney. *Statistical Methods and Linguistics*. In The Balancing Act: Combining Symbolic and Statistical Approaches to Language. The MIT Press, 1996

[2] A. V. Aho and J. D. Ullman. *The Theory of Parsing, Translation and Compiling, Vol. I: Parsing*. Prentice Hall, 1972

[3] A. V. Aho and J. D. Ullman. *The Theory of Parsing, Translation and Compiling, Vol. II: Compiling*. Prentice Hall, 1973

[4] O. Appel. *Fuzzy Grammars: What They Are and What Their Potential Applications Could Be*. Unpublished; Final Assignment for the course Applied Computational Intelligence at De Montfort University, UK, 2012

[5] A. Balahur. *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types*. PhD thesis, Department of Software and Computing Systems, University of Alicante (Spain), 2011

[6] C. Banea, R. Mihalcea, and J. Wiebe. *A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources*. In LREC 2008, 2764-2767, 2008

[7] A. Banfield. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge and Kegan Paul, 1982

[8] S. Bird, E. Loper, and E. Klein. *Natural Language Processing with Python*. O'Reilly Media Inc, 2009

[9] D. Bollegala and E. Shutova. *Metaphor Interpretation Using Paraphrases Extracted from the Web*. PLoS ONE e74304, 8(9):1614-1617, 2013

[10] E. Brill. *Some Advances in Transformation-based Part of Speech Tagging*. In AAAI '94, 722-727, 1994

[11] E. Cambria and A. Hussain. *Sentic Computing: Techniques, Tools and Applications*. Springer Briefs in Cognitive Computation, 2012

[12] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. *New Avenues in Opinion Mining and Sentiment Analysis*. IEEE Intelligent Systems, 28(2):15-21, 2013

[13] E. Charniak. *Statistical Language Learning*. The MIT Press, 1996

[14] E. Charniak. *Statistical Techniques for Natural Language Parsing*. AI Magazine, 18(4):33-44, 1997

[15] N. Chomsky. *Syntactic Structures*. Mouton de Gruyter (formerly Mouton, The Hague), 2nd revised (2002) edition, 1957 (1st edition)

[16] N. Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, 1969

[17] M. K. Dalal and M. A. Zaveri. *Semisupervised Learning-based Opinion Summarization and Classification for Online Product Reviews*. Appl. Comp. Intell. Soft Comp., 2013(Article ID 910706), 2013

[18] M. K. Dalal and M. A. Zaveri. *Opinion Mining from Online User Reviews Using Fuzzy Linguistic Hedges.* Appl. Comp. Intell. Soft Comp., 2014:1-9, 2014

[19] R. Dale. *Classical Approaches to Natural Language Processing*. In Handbook of Natural Language Processing, Chapter I, pp. 3-7, 2010

[20] F. Dzogang, M.-J. Lesot, M. Rifqi, and B. Bouchon-Meunier. *Expressions of Graduality for Sentiments Analysis - A Survey*. In FUZZ-IEEE2010, 1-7, 2010

[21] A. Esuli and F. Sebastiani. *SentiWordNet: a High-Coverage Lexical Resource for Opinion Mining.* Technical Report Institute of Information Science and Technologies of the Italian National Research Council, 2006

[22] R. Feldman. *Techniques and Applications for Sentiment Analysis*. Communications of the ACM, 56(4):82-89, 2013

[23] G. Fu and X. Wang. *Chinese Sentence-Level Sentiment Classification Based on Fuzzy Sets*. In COLING2010, 312-319, 2010

[24] T. Galli, F. Chiclana, J. Carter and Helge Janicke: *Modelling Execution Tracing Quality by Means of Type-1 Fuzzy Logic.* Acta Polytechnica Hungarica, Volume 10, Issue 8, pp. 49-67, 2013

[25] V. Hatzivassiloglou and K. R. McKeown. *Predicting the Semantic Orientation of Adjectives*. In ACL1997, 174-181, 1997

[26] V. Hatzivassiloglou and J. M. Wiebe. *Effects of Adjective Orientation and Gradability on Sentence Subjectivity.* In ACL2000

[27] M. A. Hearst. *Direction-based Text Interpretation as an Information Access Refinement*. Text-Based Intelligent Systems, 1-13, 1992

[28] J. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, 1979

[29] M. Hu and B. Liu. *Mining and Summarizing Customer Reviews*. In ACM SIGKDD, 22-25, 2004

[30] E. Hüllermeier. *Fuzzy Methods in Machine Learning and Data Mining – Status and Prospects*. Fuzzy Sets and Systems 156(3):387-406, 2005

[31] S. Jusoh and H. M. Alfawareh. *Applying Fuzzy Sets for Opinion Mining*. In (ICCAT2013, 1-5, 2013

[32] V. R. Kanagavalli and K. Raja. *Detecting and Resolving Spatial Ambiguity in Text Using Named Entity Extraction and Self Learning Fuzzy Logic Techniques*. CoRR, abs/1303.0445, 2013

[33] A. Kar and D. P. Mandal. *Finding Opinion Strength Using Fuzzy Logic on Web Reviews*. International Journal of Engineering and Industries, 2, 2011

[34] R. Khoury, F. Karray, Yu Sun, M. Kamel, and O. Basir. *Semantic Understanding of General Linguistic Items by Means of Fuzzy Set Theory*. IEEE Transactions on Fuzzy Systems, 15(5):757-771, 2007

[35] R. Kruse, D. Nauck, and C. Borgelt. *Data Mining with Fuzzy Methods - Status and Perspectives*. In EUFIT99, 1999

[36] A. Kumar and T. M. Sebastian. *Sentiment Analysis: A Perspective on Its Past, Present and Future*. International Journal of Intelligent Systems and Applications, 4(10):1-14, 2012

[37] A. Kumar and T. M. Sebastian. *Machine Learning assisted Sentiment Analysis*. In International Conference on Computer Science and Engineering, 123-130, 2012

[38] B. Liu. *Sentiment Analysis and Subjectivity*. In Handbook of Natural Language Processing, Chapter 26, pp. 627-666, Chapman & Hall CRC, 2010

[39] B. Liu. *Sentiment Analysis: A Multifaceted Problem*. IEEE Intelligent Systems, 25(3):76-80, 2010

[40] B. Liu. *Sentiment Analysis Tutorial, given at AAAI-2011*

[41] B. Liu. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers: Synthesis Lectures on Human Language Technologies, 2012

[42] B. Lu and B. K. Tsou. *Combining a Large Sentiment Lexicon and Machine Learning for Subjectivity Classification*. In Ninth IEEE International Conference on Machine Learning and Cybernetics, 3311-3316, 2010

[43] I. Maks and P. Vossen. *A Verb Lexicon Model for Deep Sentiment Analysis and Opinion Mining Applications*. In 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 10-18, 2011

[44] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008

[45] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999

[46] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. *Introduction to WordNet: an On-Line Lexical Database*. International Journal of Lexicography, 3(4):235-244, 1990

[47] G. Miller and W. Charles. *Contextual Correlates of Semantic Similarity*. Language and Cognitive Processes, 6(1):1-28, 1991

[48] A. Mudinas, D. Zhang, and M. Levene. *Combining Lexicon and Learning-based Approaches for Concept-Level Sentiment Analysis*. In 2012 International Workshop on Issues of Sentiment Discovery and Opinion Mining, 51-58

[49] A. Ortony, G. L. Clore, and M. A. Foss. *The Psychological Foundations of the Affective Lexicon*. Journal of Personality and Social Psychology, 53:751-766, 1987

[50] A. Ortony, G. L. Clore, and A. Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, 1988

[51] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval 2 (1-2), 1-135, 2008

[52] B. Pang, L. Lee, and S. Vaithyanathan. *Thumbs up? Sentiment Classification using Machine Learning Techniques.* In Conference on Empirical Methods in Natural Language Processing, 10:79-86, 2002

[53] V. Rentoumi, G. A. Vouros, V. Karkaletsis, and A. Moser. *Investigating Metaphorical Language in Sentiment Analysis: A Sense-to-Sentiment Perspective*. ACM Trans. Speech Lang. Process., 9(3):6:1-6:31, 2012

[54] E. Shutova. *Models of Metaphor in NLP*. In 48th Annual Meeting of the Association for Computational Linguistics, 688-697, 2010

[55] R. Srivastaval, MPS Bhatia, H. K. Srivastava, and C. P. Sahu. *Effects of Adjective Orientation and Gradability on Sentence Subjectivity*. In IEEE International Conference on Computer & Communication Technology, 768-775, 2010

[56] L. Suanmali, N. Salim, and M. S. Binwahlan. *Fuzzy Logic-based Method*

for Improving Text Summarization. International Journal of Computer Science and Information Security, 2(1), 2009

[57] P. Subasic and A. Huettner. *Affect Analysis of Text Using Fuzzy Semantic Typing*. IEEE Transactions on Fuzzy Systems, 9(4):483-496, 2001

[58] V. S. Subrahmanian and D. Reforgiato Recupero. *AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis*. IEEE Intelligent Systems, 23(4):43-50, 2008

[59] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. *Lexicon-based Methods for Sentiment Analysis*. Computational Linguistics, 37(2):267-307, 2011

[60] P. D. Turney. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification Reviews*. In 40th Annual Meeting of the Association for Computational Linguistics, 417-424, 2002

[61] A. van der Heide, D. Sánchez, and G. Triviño. *Computational Models of Affect and Fuzzy Logic*. In EUSFLAT 2011, 620-627, 2011

[62] A. A. Vanin, L. A. de Freitas, R. Vieira, and M. N. Bochernitsan. *Some Clues on Irony Detection in Tweets*. In 22nd International Conference on World Wide Web Companion, 635-636, 2013

[63] G. Vinodhini and RM. Chandrasekaran. *Sentiment Analysis and Opinion Mining: A Survey*. International Journal of Advanced Research in Computer Science and Software Engineering, 2(6):282-292, 2012

[64] W. Wei. *Analyzing Text Data for Opinion Mining*. In 16th International Conference on Natural Language Processing and Information Systems, NLDB'11, 330-335, 2011

[65] A. Weichselbraun, S. Gindl, and A. Scharl. *Extracting and Grounding Contextualized Sentiment Lexicons*. IEEE Intelligent Systems, 28(2):39-46, 2013

[66] J. Wiebe. *Identifying Subjective Characters in Narrative*. In International Conference on Computational Linguistics, COLING '90, 1990

[67] J. Wiebe. *Tracking Point of View in Narrative*. Comput. Linguist., 20(2):233-287, 1994

[68] J. Wiebe and E. Riloff. *Finding Mutual Benefit between Subjectivity Analysis and Information Extraction*. IEEE Transactions on Affective Computing, 2(4):175-191, 2011

[69] T. Wilson, J. Wiebe, and P. Hoffmann. *Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis*. Comput. Linguist., 35(3):399-433, 2009

[70] T. Winograd. *Language as a Cognitive Process, Volume I: Syntax*. Addison-Wesley, 1983

[71] W. and S. Bergler. *Fuzzy Coreference Resolution for Summarization*. In 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS), 43-50, 2003

[72] H. Nguyen, T. Xuan, A. Cuong Le, and L. M. Nguyen. *Linguistic features for subjectivity classification*. In International Conference on Asian Language Processing (IALP), 17-20, 2012

[73] L. A. Zadeh. *Precisiated Natural Language (PNL).* AI Magazine, 25(3):74-91, 2004

[74] L. A. Zadeh. *Is there a Need for Fuzzy Logic?* Information Sciences, 178:2751-2779, 2008