Sentiment Analysis: A Review and Comparative Analysis of Web Services

Jesus Serrano-Guerrero*,[a], Jose A. Olivas[a], Francisco P. Romero[a], Enrique Herrera-Viedma[b,c]

[a] *Department of Information Technologies and Systems*
University of Castilla-La Mancha, 13071, Ciudad Real, Spain
[b] *Department of Computer Science and Artificial Intelligence*
University of Granada, 18071, Granada, Spain
[c] *Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Abstract

Sentiment Analysis, also called Opinion Mining, is currently one of the most studied research fields. It aims to analyze people's sentiments, opinions, attitudes, emotions, etc., towards elements such as topics, products, individuals, organizations, services, etc. Different techniques and software tools are being developed to carry out Sentiment Analysis. The goal of this work is to review and compare some free access web services, analyzing their capabilities to classify and score different pieces of text with respect to the sentiments contained therein. For that purpose, three well-known collections have been used to perform several experiments whose results are shown and commented upon, leading to some interesting conclusions about the capabilities of each analyzed tool.

Key words: Sentiment Analysis, Sentiment Classification, Rating Prediction, Web Services

1. Introduction

Sentiment Analysis, also called Opinion Mining, is one of the most recent research topics within the field of Information Processing. Textual informa-

*Corresponding author. Tel: +34 926 295300 ext. 6332. Fax: +34 926 295354
Email address: jesus.serrano@uclm.es (Jesus Serrano-Guerrero)

tion techniques are mainly focused on processing, searching or mining factual information. Facts have an objective component; however, there are other textual elements which express subjective characteristics. These elements are mainly opinions, sentiments, appraisals, attitudes, and emotions, which are the focus of Sentiment Analysis (Liu, 2010). All of them are closely related, however, they present slight differences. This fact involves the birth of many related tasks in this new research field, such as *opinion mining*, *subjectivity analysis*, *emotion detection* or *opinion spam detection*, among others.

Sentiment Analysis offers many opportunities to develop new applications, especially due to the huge growth of available information in sources such as blogs, social networks, etc. For example, recommendations of items proposed by any *recommender system* can be computed taking into account aspects such as positive or negative opinions about those items. Review- and opinion-aggregation websites could collect information from different sources in order to summary or compose an opinion about a candidate, product, etc., thus replacing systems which require explicitly opinions or summaries. Question answering systems represent another field where opinions play an important role. Detection of opinion-oriented questions and possible answers, and its treatment are essential to compute good answers. Detection of subjective information is really important in fields related to argumentation where objective sentences are usually more valuable. But certainly, one of the most important fields where Sentiment Analysis has a greater impact is in the industrial field. Small and big companies, as well as other organizations such as governments, desire to know what people say about their marques, products or members (McGlohon et al., 2010; Tumasjan et al., 2010; Chen et al., 2010; Mohammad, 2012; Bar-Haim et al., 2011; Groh and Hauffa, 2011; Castellanos et al., 2011; Bollen and Mao, 2011; Moreo et al., 2012).

Sentiment Analysis is a big concept that encompasses many tasks such as extraction of sentiments, sentiment classification, subjectivity classification, opinion summarization or opinion spam detection, among others. To perform any of these activities, Sentiment Analysis has to deal with many challenges. The first one is the definition of the elements involved in this area. Thus, it is necessary to define clearly concepts such as opinion, subjectivity or emotion, however, this task is not really easy. For example, in a simple way a user opinion could be considered as a positive or negative sentiment about an entity or an aspect of that entity. On the other hand, subjectivity does not imply necessarily a sentiment but it allows expressing feelings or beliefs, and specifically, our own feelings or beliefs and our *emotions*.

These definitions have to be represented by mathematical expressions that can be computed and used as inputs for the aforementioned activities. Accordingly, Sentiment Analysis success mainly depends on the ability to extract the necessary features of those definitions from texts to perform those tasks. Thus, Natural Language Processing (NLP) techniques are essential to achieve good results depending on the task that has to be carried out. This is another of the main challenges of this research field, along with all problems related to the adaptation of typical techniques for classifying or summarizing texts in this field, as well as the creation of new techniques and algorithms specialized on opinions.

Despite the complexity and difficulty of this problem, many companies and universities are developing new tools and web services which deal with several of the issues aforementioned. These services could be included, especially for research purposes, into other applications without the need of being expert in Sentiment Analysis, such as other platforms do.

Following this idea and due to the growing number of new services related to Sentiment Analysis, the aim of this work is twofold. On the one hand, to present a detailed description of a set of 15 well-known free access services focused on Sentiment Analysis. These tools might have been developed by private companies or universities, but all of them allow free access to the functionalities that will be analyzed in this work. For that reason, all of them may be especially interesting for research purposes, as it is not necessary to implement services which are already working and are free.

And on the other hand, this work will assess the main functionalities from these 15 services related to Sentiment Analysis and analyze the results obtained. For that purpose, three well-known data collections in the field of Sentiment Analysis will be used. This way, this work will allow the user/researcher to have enough information about the different capabilities provided by each tool, and consequently, the user/researcher can choose the most appropriate one to be included into his own platform.

In summary, this paper presents a comprehensive and in-depth critical assessment of 15 Sentiment Analysis web tools that has never been done before. To properly perform this assessment, a suite of evaluation criteria and well-known data collections from the field of Sentiment Analysis has been selected to allow the reader to look into the pros and cons of the use of theses tools regarding aspects such as discovery of sentiments within short and long texts, detection of irony or computation of polarity ratings, among others. Apart from these standard data collections, these tools have also been assessed by emulating a more real scenario, in which the effectiviness for recommending movies from real users' comments has been tested using information collected from the well-known website IMDb[1].

The remainder of the work is organized as follows: Section 2 presents the main concepts related to Sentiment Analysis discussed in several recent works. Section 3 shows the main characteristics of many Web services which allow computing sentiments. Section 4 presents several experiments that have been performed in order to compare the Web services commented on the previous section, as well as the results obtained. Finally, Section 5 points out several conclusions.

2. Background

The concepts Opinion Mining, Sentiment Analysis and Subjectivity Analysis are broadly used as synonyms; however, their origins are not exactly the same and some authors consider that each concept presents different connotations. Pang and Lee (Pang and Lee, 2008) present a more detailed review on the origins of these concepts and others closely related, e.g., Affective Analysis, Review Mining or Appraisal Extraction. Therefore, it is necessary to define some concepts to understand the issue dealt with in this work.

2.1. Definition of main concepts

An opinion could be simply defined as a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity (product, person, event, organization or topic) or an aspect of that entity from a user or group of users.

Following that definition, an opinion can be mathematically defined as a 5-tuple $(e_j, a_{jk}, so_{jkil}, h_i, t_l)$ where e_j represents a target entity and a_{jk} is the k-th aspect/feature of the entity e_j. so_{ijkl} is the sentiment value of the opinion from the opinion holder h_i on aspect a_{jk} of entity e_j at time t_l. That value can be positive, negative, or neutral, or even a more granular rating can be used. h_i is the opinion holder and t_l is the time when the opinion was expressed (Liu, 2010).

Opinions can be classified into different groups, for instance, they could be regular and comparative opinions. Most of opinions are regular, and they can be subdivided into direct or indirect opinions. Direct opinions

[1] http://www.imdb.com/

express an idea on an entity or an aspect of an entity, whereas indirect opinions express an opinion on an entity or an aspect of an entity based on the effects on other entities. On the other hand, comparative sentences express the resemblance between entities considering common aspects or features (Jindal and Liu, 2006a,b; Yang and Ko, 2011). Furthermore, opinions can be classified into explicit or implicit, depending on whether they express subjective or objective ideas (Liu, 2011).

Apart from sentiment and opinion, there are two important concepts close to them, subjectivity and emotion. Subjectivity allows us to express personal feelings, views, or beliefs; however, a subjective sentence does not imply necessarily any sentiment. According to Liu, the difference between objective and subjective sentences is "an objective sentence expresses some factual information about the world, while a subjective sentence expresses some personal feelings or beliefs". An example might be the sentence: "I think they are gone". Nevertheless subjectivity sometimes involves sentiments to some extent when is dealing with affect, judgment, appreciation, speculation, agreement, etc. (Liu, 2010)

On the other hand, an emotion can be seen as an expression of our own subjective feelings and thoughts. Emotions are really close to sentiments, indeed, the way of measuring the strength of an opinion is linked to the intensity of certain emotions, such as love, joy, surprise, anger, sadness, or fear. An example might be the sentence: "I love this car", in which the speaker expresses his objective love for his car.

It is also necessary to comment the concept of mood, which could be considered as "a mix of sentiments, emotions, feelings that move the author of a certain text to write that comment, observation, criticism, etc." (Loia and Senatore, 2014).

2.2. Tasks

Many tasks arise linked to Sentiment Analysis. Some of them are closely related and it is difficult to separate them clearly because they share many aspects. The most important are:

1. Sentiment classification: also called sentiment orientation, opinion orientation, semantic orientation or sentiment polarity (Yu et al., 2013). It is based on the idea that a document/text expresses an opinion on entity from a holder and tries to measure the sentiment of that holder towards the entity. Therefore, it mainly consists in classifying opinions

into three main categories: positive, negative or neutral. It seems a simple task; however, it is a really complex task, especially when opinions come from multiple domains or languages (Rushdi-Saleh et al., 2011; He et al., 2011; Boiy and Moens, 2008). This task is closely related to sentiment rating prediction, which consists in measuring the intensity of each sentiment. Different scales can be used to measure an opinion, for example, the range $[-1, 1]$ where -1 indicates the maximum negative degree and 1 the maximum positive degree, or a scale of five stars where an opinion holder can select zero stars to express maximum negativity or five stars otherwise (Nigam and Hurst, 2006; Li and Tsai, 2013; Martin-Valdivia et al., 2012; Zhou and Chaovalit, 2008).

2. Subjectivity classification. It mainly consists in detecting whether a given sentence is subjective or not. An objective sentence expresses factual information while a subjective sentence can express other types of personal information such as opinions, evaluations, emotions, beliefs, etc. Furthermore, subjective sentences can express positive or negative sentiment, but not all of them do. This task can be seen as a previous step before classifying sentiments. A good subjectivity classification can ensure a better sentiment classification. It is even considered as a process more difficult than distinguishing between positive, neutral or negative sentiments. (Barbosa and Feng, 2010; Raaijmakers and Kraaij, 2008; Sarvabhotla et al., 2011; Tang et al., 2009; Esuli and Sebastiani, 2006; Montoyo et al., 2012; Maks and Vossen, 2012).
3. Opinion summarization. It is especially focused on extracting the main features of an entity shared within one or several documents and the sentiments regarding them (Wang et al., 2013). Thus, two perspectives can be distinguished in this task: single-document and multi-document summarization. Single-document summarization consists in analyzing internal facts present within the analyzed document, for example, changes in the sentiment orientation throughout the document or links between the different entities/features found, and mainly showing those pieces of texts which better describe them. On the other hand, in multi-document summarization once features and entities have been detected, the system has to group and/or order the different sentences which express sentiments related to those entities or features. The final summary can be presented as a graphic or a text showing the main features/entities and quantifying the sentiment with regard to each

one in some way, for example, aggregating intensities of sentiments or counting the number of positive or negative sentences (Beineke et al., 2004; Pang and Lee, 2004; Park et al., 2012; Nishikawa et al., 2010b,a; Ganesan et al., 2010, 2012; Tata and Di Eugenio, 2010).

4. Opinion retrieval. It tries to retrieve documents which express an opinion about a given query. In this kind of systems, two scores are required to be computed for each document, the relevance score against the query and the opinion score about the query, and both are usually used to rank the documents (Lee et al., 2011; Guo and Wan, 2012).
5. Sarcasm and irony. It is focused on detecting statements which contain ironic and sarcastic content. This is one of the most complicated tasks in this field, especially, because of the absence of agreement among researchers on how irony or sarcasm can be formally defined (Filatova, 2012; Reyes and Rosso, 2012; Reyes et al., 2012).
6. Others. Apart from the previously mentioned activities, other tasks related to Sentiment Analysis can be highlighted, e.g., *genre or authorship detection* tries to determine the genre or the person who has written a text/opinion (Finn and Kushmerick, 2006; Savoy, 2012; Montesi and Navarrete, 2008; Seki et al., 2009) or *opinion spam detection* tries to detect opinions or reviews which contain untrusted contents published to distort public opinion towards people, companies or products (Jindal and Liu, 2007; Ott et al., 2012; Mukherjee et al., 2012; Xie et al., 2012; Wang et al., 2012).

2.3. Techniques

Several works try to show the different techniques applied to Sentiment Analysis. Most of them group the works from the point of view of the different applications/challenges that can be found in SA as in (Pang and Lee, 2008) and (Liu and Zhang, 2012). Other works like (Tsytsarau and Palpanas, 2011) or (Feldman, 2013) are focused on the main topics of SA.

Thus, Feldman groups all works under five main groups: document-level sentiment analysis, sentence-level sentiment analysis, aspect-based sentiment analysis, comparative sentiment analysis and, sentiment lexicon acquisition (Feldman, 2013). And on the other hand, Tsytsarau and Palpanas mainly focus on opinion aggregation, opinion spam and contradictions analysis, especially applied to Web services, for example, microblogs or streaming data, among others (Tsytsarau and Palpanas, 2011). They present four different

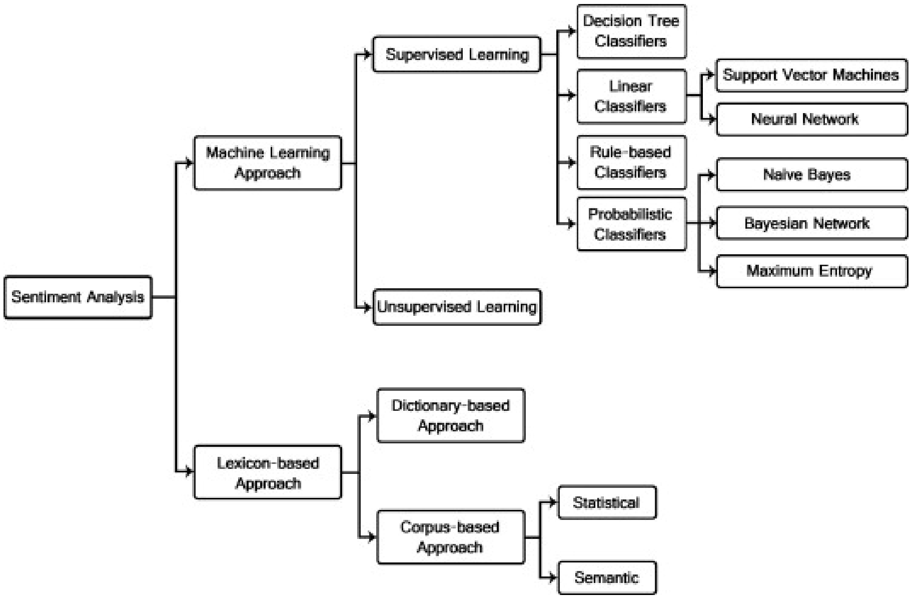


Figure 1: Sentiment classification techniques

points with respect to previous works to classify Sentiment Analysis techniques: machine learning, dictionary-based, statistical and semantic.

Possibly, the most interesting work from the point of view of the SA techniques is (Medhat et al., 2014), which presents a refined categorization to well-known SA techniques (see Fig. 1) including new trends such as Emotion Detection (Rao et al., 2014), Building Resources and Transfer Learning.

2.3.1. Machine learning approaches

They can be grouped in two main categories: supervised and unsupervised techniques. The success of both is mainly based on the selection and extraction of the appropriate set of features to model the classifier. In this task Natural Language Processing techniques play a very important role because some of the most important features used are for example: (1) terms (words or n-grams) and their frequency; (2) part of speech information, adjectives play an important role but nouns can be significant; (3) negations can change the meaning of any sentence; (4) syntactic dependencies (tree parsing) can determine the meaning of sentence; among others (Liu and Zhang, 2012; Chenlo and Losada, 2014).

With respect to supervised techniques, support vector machines (SVM), Naive Bayes, Maximum Entropy are some of the most common techniques used (Ye et al., 2009; Rushdi-Saleh et al., 2011; Montejo-Raez et al., 2014). Whereas semi-supervised and unsupervised techniques are proposed when it is not possible to have an initial set of labeled documents/opinions to classify the rest of items. In this case, other approaches such as statistical and semantic methods have to be performed to overcome that handicap (He and Zhou, 2011; Xianghua et al., 2013).

Besides, hybrid approaches, combining supervised and unsupervised techniques, or even semi-supervised techniques, can be used to classify sentiments (Kim and Lee, 2014; König and Brill, 2006).

2.3.2. Lexicon-based approaches

Lexicon-based approaches mainly rely on a sentiment lexicon, i.e., a collection of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon (Wilson et al., 2005a); but, even more complex structures like ontologies (Kontopoulos et al., 2013), and score-based methods (Turney, 2002; Taboada et al., 2011) can be used for this purpose.

Two subclassifications can be found here: Dictionary-based and Corpus-based approaches. The former is based on the use of corpora that is usually annotated in a manual way such as WordNet (Miller, 1995) to develop a thesaurus called SentiWordNet (Baccianella et al., 2010). The main drawback of this kind of approaches is the incapability to deal with domain and context specific orientations; even so, it might be an interesting solution depending on the problem (Martin-Valdivia et al., 2012; Montejo-Raez et al., 2014).

The corpus-based techniques arise with the objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use either statistical techniques or semantic techniques. Statistical methods such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990) or simply the frequency of apparition of the words within a collection of documents can be used (Cao et al., 2011). And on other hand, semantic methods such as the use of synonyms and antonyms or relationships from thesaurus like WordNet may also represent an interesting solution (Zhang et al., 2012).

2.4. Natural Language Processing and Information Retrieval in Sentiment Analysis

According to Cambria, Sentiment Analysis can be considered as a very restricted NLP problem, where it is only necessary to understand the positive or negative sentiments concerning each sentence and/or the target entities or topics (Cambria et al., 2013). However, in spite of being a restricted problem, all works in this field, as well as all works in Information Retrieval, always struggle with NLPs unresolved problems (negation handling, named-entity recognition, word-sense disambiguation, ...) which are essential to detect keys of language such as irony or sarcasm (Reyes and Rosso, 2012; Reyes et al., 2012), and consequently, to find and rate sentiments.

One of the main aspects that NLP has to deal with is the different levels of analysis. Depending on whether the target of study is a whole text or document, one or several linked sentences, or one or several entities or aspects of those entities, different NLP and Sentiment Analysis tasks can be performed. Hence, it is necessary to distinguish three levels of analysis that will clearly determine the different tasks of Sentiment Analysis: (i) document level, (ii) sentence level and (iii) entity/aspect level.

Document level considers that a document is an opinion on an entity or aspect of it. This level is associated with the task called *document-level sentiment classification* (Zhang et al., 2009; Moraes et al., 2013; Duric and Song, 2012; He and Zhou, 2011; Zhou et al., 2010; Yessenalina et al., 2010; Paltoglou and Thelwall, 2010; Li and Liu, 2012). However, if a document presents several sentences dealing with different aspects or entities, then the sentence level is more suitable. Sentence level is related to the task *subjectivity classification*; it considers each sentence as a positive, negative, or neutral opinion(Wilson et al., 2005b, 2009; Agarwal et al., 2009; Remus and Hänig, 2011). And finally, when more precise information is necessary, then the entity/aspect level arises. It is the finest-grained level, it considers a target on which the opinion holder expresses a positive or negative opinion. This last level is possibly the most complex because it is necessary to extract with high precision many features such as dates or time spans, the different features/aspects and entities to be opinionated, along with the relations between them, the opinion holders and their characteristics, etc. It is closely related to tasks like Opinion Mining and Opinion Summarization (Thet et al., 2010; Ojokoh and Kayode, 2012; Vechtomova, 2010).

Many of these papers follow the same general strategies as other Information Retrieval works did before, but replacing several statistical or semantic variables for aspects related to sentiments. For example, (Vechtomova and Karamuftuoglu, 2008) propose the use of lexical cohesion, i.e., the "physical" distance between collocations to rank documents whereas in sentiment ranking Vechtomova proposes a similar method, but measuring the distance between subjective words (Vechtomova, 2010). Thus, the main difference between these works is the feature selection process.

Another example might be the work presented in (Moraes et al., 2013), that applies well-known supervised methods to Sentiment Classification, Artificial Neural Networks and Support Vector Machines, which have been used thousands of times in Information Retrieval. In this case again, the difference with other works on Information Retrieval (Zhang et al., 2008) is the feature selection. The experiments were carried out following the Abassi's ideas, who proposes that feature selection methods should be tailored to sentiment analysis by combining syntactic properties of text features with sentiment-related semantic information extracted, for example, from sources like SentiWordNet (Abbasi et al., 2011; Abbasi, 2010).

As can be seen, the use of lexicons is many times necessary to support several of these NLP activities (Gerani et al., 2012; Thet et al., 2010; Loia and Senatore, 2014). And furthermore, the problem of analyzing the different syntactic levels can be even more complex working with texts written in different languages (Abbasi et al., 2008; Kim et al., 2010; Banea et al., 2010).

3. Web Services

This section summarizes the main features of several Web services which incorporate functionalities related to Sentiment Analysis. Apart from these tools, many others could be found or are arising just now, for instance, SocialMention[2], TweetFeel[3], SenticNet[4] or Luminoso[5]. However, we had free access only to these tools. In addition, they offer simple web access points to program the experiments shown in the following Section in order to compare their capabilities regarding Sentiment Analysis.

[2] http://socialmention.com

[3] http://www.tweetfeel.com

[4] http://sentic.net

[5] http://luminoso.com

3.1. AlchemyAPI

AlchmeyAPI[6] is a Software as a Service (SaaS) platform which enriches textual content through automated tagging, linguistic analysis, categorization, and semantic mining. It is a tool based on natural language processing and machine learning which offers functionalities such as Named Entity Extraction, Concept Tagging, Keyword / Term Extraction, Sentiment Analysis, Relation Extraction, Author Extraction, Automatic Language Identification, Text Extraction/Web Page Cleaning, Structured Data Extraction/Content Scraping, Microformats Parsing/Extraction, and RSS/ATOM Feed Detection. These functionalities are accessible via a HTTP REST interface and different Software Development Kits developed in languages like Java, C# or Perl.

Regarding Sentiment Analysis, it classifies sentiments within the analyzed text into three categories: positive, negative and neutral, and measures the sentiment degree in range $[-1, 1]$. Text should be written in English or German. It is able to detect directional sentiment (positive/negative statements) for Subject-Action-Object relations, user-targeted sentiments and also performs Sentiment Analysis at different levels: document, entity or keyword level.

The user can provide a main phrase regarding the analyzed text and the system is able to extract sentiment targeted toward that phrase.

Directional sentiment consists in detecting relations subject and object and computing the sentiment "directed at" the relation Object via the Subject-Action. For instance, the sentence "Ugly Bob attacked beautiful Susan", the subject "Ugly Bob" has a negative sense, the object "beautiful Susan", a positive sense, however, Susan has a negative directional sentiment from the relation Subject-Object (Bob → attacked → Susan).

3.2. Lymbix

Lymbix[7] is a company focused on social media analysis through NLP with adaptive learning gathered from human-powered techniques. It offers several services able to compute Sentiment Analysis, tone analysis, entity extraction or topic discovery through several interfaces developed in PHP, Java, .NET, Ruby or Python.

[6] http://www.alchemyapi.com/

[7] http://www.lymbix.com/

Lymbix goes further than a simple sentiment classification (positive, negative or neutral categories), it measures the emotive context in social conversations through different concepts grouped into different positive or negative categories. The positive categories are affection/friendliness, enjoyment/elation, amusement/excitement, contentment/gratitude, and on the other hand, sadness/grief, anger/loathing, fear/uneasiness, humiliation/shame are the negative categories (see Figure 2). Both concepts, sentiments and emotions, are not only classified but are also measured in the range $[-10, 10]$.

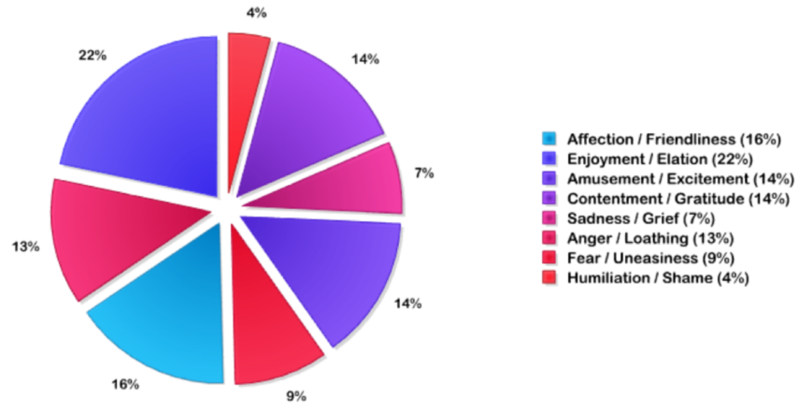
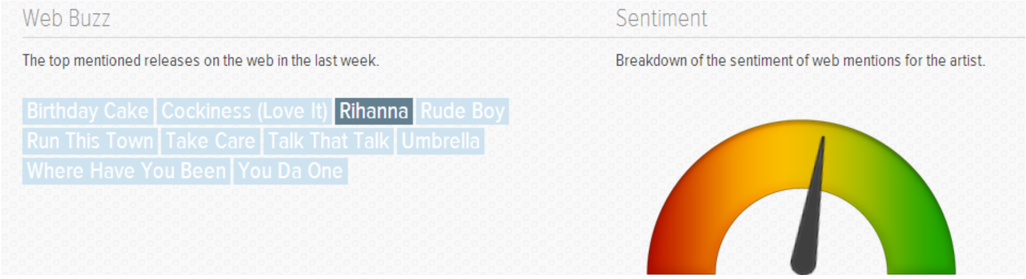


Figure 2: Summary of categories decteted by Lymbix from a document

3.3. Musicmetric

Musicmetric[8] is a software company specialized in tracking the real-time activity of users of social networks and blogs mentioning over 600.000 artists and 10 million individual releases. It performs several statistical tasks such as counting the number of mentions, fans or performances of a specific artist from several well-known sites: Youtube, MySpace, LastFM, etc. This platform also computes sentiments with respect to a concrete artist as can be seen in Figure 3, however, currently the API only classifies sentiments into five categories (1-very negative, 2-negative, 3-neutral, 4-positive, 5-very positive) without weighting them. For that purpose, the system includes matching learning models that can be trained with different corpora (tweets, blogs comments, etc.) in order to deal with different contexts.

[8] http://www.musicmetric.com/



Web Buzz

The top mentioned releases on the web in the last week.

Sentiment

Breakdown of the sentiment of web mentions for the artist.

Figure 3: Sentiment for the query: *Rihanna*

3.4. Openamplify

Openamplify[9] is a company specialized in Social Media Marketing which provides a text analytics platform able to process thousands of texts extracting features related to semantic aspects or sentiments, among others.

Texts, structured or unstructured, without the need for training or special vocabularies are analyzed using natural language processing techniques which are able to detect many features such as topics and entities, domains, categories and classifications, actions, intentions, decisiveness, emotions[10] and sentiments at both the topic and whole text level.

Polarity is considered as the attitude expressed towards a topic found in the text. It is computed from a combination of linguistic features and takes account of negations and even multiple negations, on a scale of -1.0, very negative, to very positive, $+1.0$. The mean polarity is computed as the mean value of the single polarity degrees for each instance of a given topic.

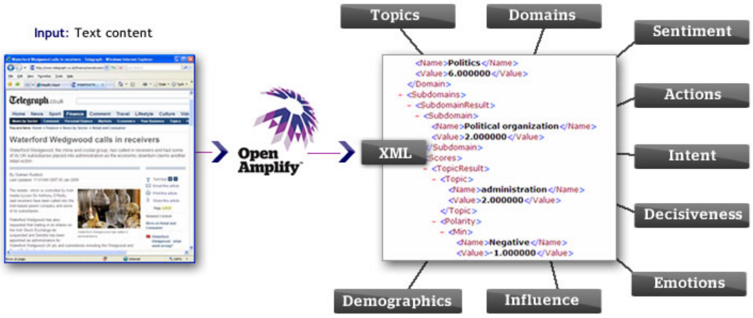
3.5. Opinion Crawl

Semantic Engines LLC is a private company that develops products in the fields of Information Search and Retrieval, Text Mining, Semantic Analysis, Sentiment Analysis, and Contextual Advertising. An example is SenseBot[11], a semantic search engine which allows searching in several languages (English, French, German and Spanish), generating a text summary of the retrieved results on the topic of the user query. For that purpose, Text Mining and

[9] http://www.openamplify.com/

[10] In the provided documentation this concept is not clearly related to emotions such as anger, fear, etc.

[11] http://www.sensebot.net/



```
    <Name>Politics</Name>
    <Value>6.000000</Value>
  </Domain>
- <Subdomains>
  - <SubdomainResult>
    - <Subdomain>
        <Name>Political organization</Name>
        <Value>2.000000</Value>
      </Subdomain>
      Scores>
    - <TopicResult>
      - <Topic>
          <Name>administration</Name>
          <Value>2.000000</Value>
        </Topic>
      - <Polarity>
        - <Min>
            <Name>Negative</Name>
            <Value>-1.000000</Value>
```

Figure 4: Openamplify scheme

Multidocument Summarization techniques are used to extract senses from retrieved results and present them to the user.

This company developed another platform, Opinion Crawl[12], which is a search engine able to assess sentiments on a subject in Internet platforms such as Twitter or news services. This search engine shows charts expressing real-time sentiments, the latest news headlines, and a few recent images along with a tag cloud of key semantic concepts related to the searched topic. Apart from the web page, there exists a Sentiment Analysis API which supports SOAP (Simple Object Access Protocol) and REST (Representational State Transfer) protocols and allows assessing sentiments within pieces of text or web pages as a whole, or the user can provide a key subject and the sentiment-targeted toward that subject is computed. The main parameters retrieved by this service are the polarity sense (positive, negative or neutral) of the text, the number of positive, negative and neutral expressions found as well as the ratio of positive to negative expressions.

3.6. Opendover

Opendover[13] is a Java-based web service whose objective is the extraction of semantic features within texts from different sources like blogs, content management systems or web sites. It is an ontology-based service specialized in different domains such as education, law, politics, health, economy, ecology, etc. It consists of a knowledge base containing thousands of opinion words,

[12] http://www.opinioncrawl.com/

[13] http://opendover.nl/

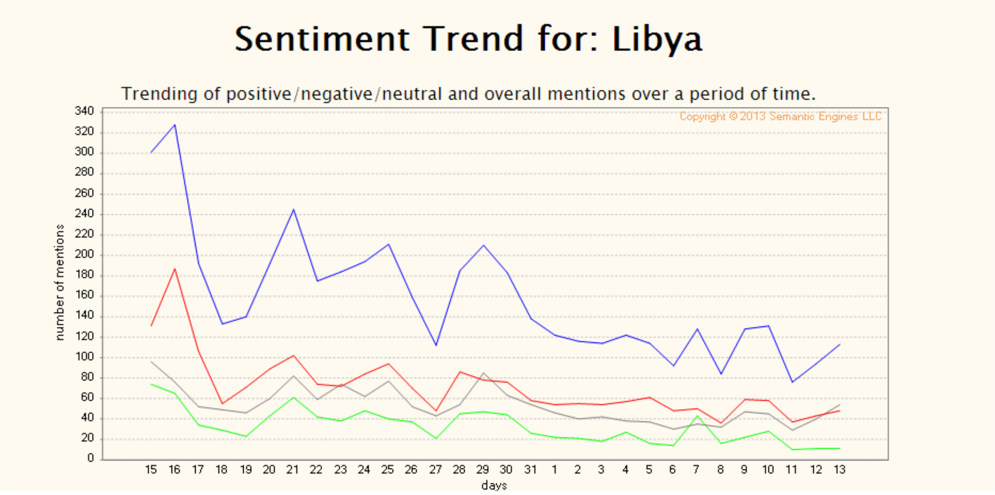


Figure 5: Opinion Crawl diagram

domain-related words and relations. This structure allows offering several functionalities like automatic disambiguation of opinion and domain-related words, recognition of context dependent or independent opinion words, recognition and categorization of texts, among others.

Regarding sentiments, it classifies sentiments into three categories: judgment, appreciation and emotional state. Sentiments are labeled as positive, neutral or negative, along with a value scoring the sentiment strength in the range [-9,9].

SentimentSearch[14] is an example of web site that uses this service to explore sentiments from different domains in Twitter or blogs.

3.7. Repustate

Repustate[15] is a multilingual sentiment engine which deals with several languages such as English, French, German, Spanish, or Arabic. It allows processing texts from many sources such as Facebook pages, Twitter or simply pieces of texts. Texts can be processed as a whole document or can be chunked through Natural Language Processing techniques in order to find those elements on which sentiments are addressed. In addition, all sentiments are scored in the range $[-1, 1]$.

[14] http://sentimentsearch.nl/

[15] https://www.repustate.com/

3.8. Semantria

Semantria[16] is a multilingual sentiment engine which deals with several languages such as English, French, German, Spanish, or Portuguese. Like Repustate, it allows the processing of information from many sources such as Facebook, Wordpress, Twitter, as well as simple pieces of texts.

Sentiments detection can be carried out at paragraph, sentence or entity level. For that purpose it is necessary to identify, through POS tagging, structural portions of a document, paragraph or sentence including nouns, adjectives, verbs and adverbs. It also allows the processing of whole collections up to 100 documents at the same time.

From those portions, sentiment-bearing phrases are identified and scored considering the rate of appearance of the phrase near a set of known good or bad words, because the system consists of a dictionary of phrases and their comparative scores. For example, given a phrase X, the system submits the queries "X near (good wonderful, spectacular)" and "X near (bad, horrible, awful)', and depending on the retrieved results, the sentiment score of that phrase is calculated in the range $[-1, 1]$. This score is computed for a concrete document, but not for a concrete sentence. Sentences are divided into chunks referred to an entity or facet which has an associated sentiment, and then this sentiment is scored as well. Therefore sentiments are computed at entity level rather than sentence level. Besides, the system is flexible and allows the user to insert his own dictionary with the associated weights for each word included.

3.9. Sentiment140

Sentiment140[17] is a project at Standford University (Go et al., 2009). The objective of the project is to classify sentiments in Tweets into three categories (negative, neutral, positive), however, it does not assign a score to each sentiment. Tweets are short texts with special characteristics such as emoticons, links or usernames, which have to be detected before classifying them. After detecting and normalizing the main features of each Tweet, the classification process can execute several machine learning algorithms such as Naive Bayes, Maximum Entropy, or Support Vector Machines (SVM), which use unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags as feature extractors.

[16] http://semantria.com/

[17] http://www.sentiment140.com/

Web pages like Twitty City[18] use this technology to measure sentiments; in this particular case, the general sentiment of different English cities is computed through the Tweets written.

3.10. SentimentAnalyzer

SentimentAnalyzer[19] is a simple web service which computes sentiment for English, German or French texts. It is able to classify the polarity (positive, negative or neutral) of a whole text and score it in the range $[-1, 1]$. Unfortunately, there is not much information available about this site.

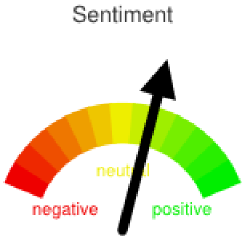


Figure 6: Example of SentimentAnalyzer interface

3.11. SentiRate

SentiRate[20] is a sentiment processing engine that analyzes pieces of text in order to detect emotions or attitudes. It is able to analyze text at two levels, whole text and sentence by sentence. Polarity is grouped into 11 categories: very_positive, quite_positive, positive, fairly_positive, a_little_positive, neutral, a_little_negative, fairly_negative, negative, quite_negative, very_negative. It also scores each sentence or the whole text by number with two decimals.

3.12. Sentimetrix

Sentimetrix[21] is a commercial tool whose objective is to analyze what consumers are expressing and share opinions about brands, products or services of a company. It performs real time analysis of the blogosphere in order to

[18] http://www.twittycity.co.uk/

[19] http://sentimentanalyzer.appspot.com/

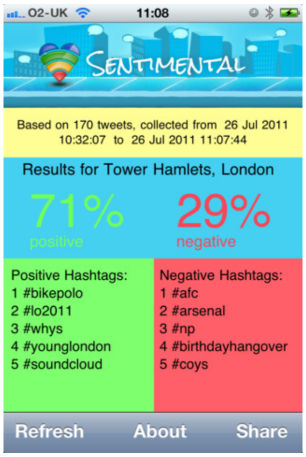
[20] http://sentirate.com

[21] http://www.sentimetrix.com

provide precise sentiment scoring, and attributes level opinion analysis. All results are rated on a continuous scale in the range $[-1, 1]$.

API developers criticize the use of lists of words to detect and weight sentiments, and propose the use of statistical learning as the best solution. It allows learning types of words people use to express emotions, for example, emoticons, slang, hashtags, etc. However, it also needs a huge database to train sentiments detection models. For that reason, over 110 million tweets have been used to train the Sentimetrix models. It supports nine languages, especially, the most spoken (English, Chinese, Italian, Spanish, French or Russian). Gate[22], the well-known Java-based library, has been used to perform many of the most important NLP tasks of this tool.

Thanks to the API of this tool, applications like Sentimental[23] have been developed. It collects tweets from two miles around your iPhone's current location and analyzes sentiments within them. In Figure 7 an example of the graphic interface can be seen.



Based on 170 tweets, collected from 26 Jul 2011 10:32:07 to 26 Jul 2011 11:07:44

Results for Tower Hamlets, London

71% positive

29% negative

Positive Hashtags:	Negative Hashtags:
1 #bikepolo	1 #afc
2 #lo2011	2 #arsenal
3 #whys	3 #np
4 #younglondon	4 #birthdayhangover
5 #soundcloud	5 #coys

Figure 7: Sentimetal interface

[22] http://gate.ac.uk/

[23] https://itunes.apple.com/us/app/sentimental/id452701697?ls=1&mt=8

3.13. Uclassify

Uclassify[24] is a free web service which performs many functionalities such as language detection, text gender and age recognition, spam filter, Sentiment Analysis, document tagging, mood (happy or upset), among others. These services are carried out thanks to classifiers that can be created, used and shared freely. As example, the web site Genderanalyzer[25] uses the Uclassify service to detect whether a web page has been written by a man or a woman, or the Wordpress plugin called Trollguard[26] detects spam comments from blogs.

Regarding Sentiment Analysis, a sentiment classifier, which uses 40.000 Amazon reviews[27] from 25 different product genres, has been used. It reveals if a document is positive or negative, and how positive or negative is a web page or a piece of text by means of the probability of being classified as positive or negative.

3.14. ViralHeat

ViralHeat[28] is a web service based on a supervised classifier able to analyze and capitalize on the explosion of individual opinions expressed on online services such as Facebook, LinkedIn, Twitter, or Google+. This tool allows monitoring real-time stream of mentions for brands, products or topics of interest to each customer.

Apart from that web interface which allows inserting several social network accounts and monitoring data contained in those accounts, it provides an API which allows collecting statistical details such as the number of Facebook likes, tweets, retweets, etc. And, in addition to the functionalities focused on social network accounts, it offers a service which processes texts classifying them into two categories, positive or negative, and computing the probability with which the system thinks that the proposed text has the output sentiment.

[24] http://uclassify.com/

[25] http://www.genderanalyzer.com/

[26] http://www.trollguard.com/

[27] http://www.cs.jhu.edu/˜mdredze/datasets/sentiment/

[28] https://www.viralheat.com/

3.15. Wingify

Wingify[29] is a company specialized in optimizing web sites in order to increase sales and decrease costs. It offers a product called Visual Website Optimizer which permits creating different versions of a web site and optimizing all of them looking for different goals. This tool measures these optimizations by different multivariable tests and issues reports about the performance of them.

Winfigy also offers a Restful API which provides functionalities such as tracking visitors or real-time information on Visitors, and in addition to these functions, this tool is able to extract tags, concepts, categories, sentiments and related links for a piece of text or a given URL. It does not classify specifically, only presents a value in the range $[0, 1]$ (negativeness to positiveness), and even, if it is not able to analyze a text then the value -1 is returned.

3.16. Comparative analysis

Next, a summary of the main features of each tool can be seen in Table 1. This table presents those characteristics that can be explicitly read from the provided documentations or through the use of these web services, e.g., those features related to Sentiment Analysis that these tools provide, the syntactic level at which sentiments can be detected, or whether they are able to rate the intensity of each sentiment.

4. Experiments

4.1. Data collections

As aforementioned, Sentiment Analysis tools can perform several tasks. Two of the most important are *sentiment classification* and *sentiment rating prediction*. We are trying to analyze the capabilities of these tools with respect to these two tasks; the ability to classify text as positive, negative or neutral, and the capability for measuring the intensity of each sentiment detected.

In order to evaluate these characteristics, three well-known collections have been chosen. Each one provides different features that make them interesting for these experiments:

[29] http://wingify.com/

	Docum. Level	Senten. Level	Entity Level	Polarity rating	emotions	length constraints	Others
AlchemyAPI	X	X	X	X	-	-	author extraction
Lymbix	X	X	-	X	-	-	-
Musicmetric	X	-	-	-	-	-	-
Openamplify	-	-	X	X	X	-	gender, age of author, and disagreement -
Opinion Crawl	-	X	-	-	-	-	-
Opendover	-	-	X	X	-	-	subjectivy
Repustate	X	-	X	X	-	-	-
Semantria	X	-	X	X	-	-	-
Sentiment140	X	-	-	-	-	-	-
SentimentAnalyzer	X	-	X	-	-	-	-
SentiRate	X	X	-	X	-	-	-
Sentimetrix	X	-	-	X	X	X	-
Uclassify	X	-	-	X	-	X	gender recognition, age recognition spam filter, mood
ViralHeat	X	-	-	X	X	X	-
Wingify	X	-	-	X	-	-	-

Table 1: Summary of features for each Web service

- Large movie review dataset[30] is a collection of movie reviews for binary sentiment classification collected by Andrew Maas from Stanford University (Maas et al., 2011). It contains $25,000$ highly polar movie reviews for training and $25,000$ for testing, along with additional unlabeled data. Each review has a value using a star rating on a $[1, 10]$ scale and a category, positive or negative; neutral reviews are not available.

- Twitter dataset[31] is a collection of tweets collected from the time period between April 6, 2009 to June 25, 2009. It is used for sentiment classification as well and the sentiment polarity of each tweet is not scored. It contains $800,000$ positive tweets and $800,000$ negative tweets. A more detailed explanation about this dataset can be found in (Go et al., 2009).

- Amazon product review dataset[32] was designed to identify sarcasm on two levels: a review or a text utterance (where a text utterance can be as short as a sentence and as long as a whole review). It contains 437

[30] http://ai.stanford.edu/ãmaas/data/sentiment/

[31] http://help.sentiment140.com/for-students

[32] http://storm.cis.fordham.edu/f̃ilatova/SarcasmCorpus.html

product reviews collected from Amazon[33] and scored using stars in the range [1,5]. Each review has been manually selected and classified as ironic or sarcastic by means of the Mechanical Turk service[34]. But apart from detecting the ironic or sarcastic reviews, the concrete sentences, which express sarcasm or irony within each review, were extracted. A more detailed explanation on how all reviews and text utterances were manually collected and classified is available in (Filatova, 2012).

These datasets are really large because they are designed to train learning algorithms; however, to assess the performance of the provided APIs it is not necessary to use so many documents. In addition, most of the tested APIs have constraints regarding the time between two consecutive calls or the number of calls per day, i.e., the free service only allows submitting a specific number of calls each day, if the user wants to submit an undetermined number of calls, he/she has to pay for a subscription service. For these reasons, to test the first collection (Large movie review dataset) $1,000$ documents have been selected from the testing set, the first 500 positive reviews and the first 500 negative reviews.

With respect to the second collection (Twitter dataset), 498 manually analyzed data are provided by the Standford University as is explained in (Go et al., 2009), which have been used to assess these APIs. In this case, unlike the Large movie review dataset, this dataset contains neutral opinions in addition to positive and negative opinions. The distribution of the tweets by sentiment is: 177 negative, 182 positives and 139 neutrals.

And finally, from the Amazon collection, all reviews (437) which present ironic or sarcastic content were used (Filatova, 2012). In this case the distribution of the tweets by sentiment is: 289 positives, 20 neutrals and 128 negatives.

Regarding the textual characteristics of the different collections, Table 2 shows a summary of the different characteristics: characters, words, sentences, words per opinion and sentences per opinion.

4.2. Evaluation measures

Regarding the way of assessing this work, and following recent works (Ye et al., 2009; Moraes et al., 2013), the accuracy, recall, precision and the mean

[33] www.amazon.com

[34] https://www.mturk.com

	characters	words	sentences	words/op.	senten./op.
Twitter dataset	28, 323	5, 683	902	11.4	1.8
Movie dataset	1, 094, 159	234, 791	12, 389	234.8	12.4
Amazon whole reviews	23, 498, 116	444, 844	5, 710	224.5	13
Amazon utterances	139, 510	30, 716	1, 845	70.2	4.2

Table 2: Summary of characteristics of the data collections

square error have been selected as measures to assess the different capabilities of each tool.

The accuracy has been defined as follows:

$$accuracy = \frac{\#hits}{\#total\_reviews} \tag{1}$$

where $\#total\_reviews$ is the total number of reviews used for this experiment and $\#hits$ represents the number of reviews (positive, negative or neutral) that have been correctly classified by each tool.

The precision can be computed taking into account only one kind (positive, negative or neutral) of review that has been correctly classified. Thus, for example, the positive precision could be defined as:

$$positive\_precision = \frac{\#positive\_well\_classified}{\#positive\_well\_classified + \#positive\_bad\_classified} \tag{2}$$

where $\#positive\_bad\_classified$ represents the number of negative or neutral reviews that were incorrectly classified as positive, and $\#positive\_well\_classified$ is the number of original positive reviews that were submitted to a specific tool and labelled as positive. The negative and neutral precisions are computed in the same manner.

The recall can be computed taking into account only one kind (positive, negative or neutral) of review that has been correctly classified. Thus, for example, the positive recall could be defined as:

$$positive\_recall = \frac{\#positive\_well\_classified}{\#total\_positive\_reviews} \tag{3}$$

where $\#total\_positive\_reviews$ represents the number of reviews originally considered as positive within the data collection, and $\#positive\_well\_classified$ is the number of original positive reviews that were submitted to a specific tool and labelled as positive. The negative and neutral recall are computed in the same manner.

Besides, in order to compare the capability of these tools to weight the intensity of each sentiment, the mean square error (mse) has been chosen as comparison measure:

$$mse = n^{-1} \sum_{i=1}^{n} (x_i - y_i)^2, \tag{4}$$

where the vectors x, y represent the polarity ratings for all reviews, the original ones and the computed by each tool respectively, and n is the length of both vectors.

4.3. Previous considerations

Before commenting the results, it is necessary to say that all texts/documents from each collection have been submitted to each tool using the configuration by default because these tools barely allow configuring one parameter or none, most of them none. Besides, for several reasons some of the submitted reviews could not be analyzed. Some of these reasons are, for example, the length of each text, some services are not able to process texts larger than a certain amount of bytes or characters, or they are not able to process sentences containing special characters such as '#' or '@'. These aspects are especially present in the first data collection (column #1), whose texts are larger than the texts of the rest of collections used and specially contain this kind of particular characters ('#', '@'). Table 3 shows the number of movie reviews, tweets, whole ironic reviews or only ironic sentences that each tool was not able to process correctly.

The most remarkable data are the 269 failures found by Wingify in the Twitter dataset. This fact is due to the fact that the API was not able to classify the document and returned a value "Can't say", possibly because it needs context to classify each tweet and they are too short. In the other dataset, where texts are larger, this problem does not occur. These data have not been considered for statistics.

	MoviesReviews	Tweets	WholeReviews	OnlySentences
AlchemyAPI	1	0	1	4
Lymbix	0	0	0	0
Musicmetric	51	0	6	10
Openamplify	68	17	0	8
Opinion Crawl	6	0	0	0
Opendover	1	0	0	0
Repustate	44	0	6	7
Semantria	1	5	1	3
Sentiment140	0	0	0	4
SentimentAnalyzer	23	0	6	6
SentiRate	58	0	9	5
Sentimetrix	0	0	0	4
Uclassify	0	0	0	0
ViralHeat	61	0	0	0
Wingify	12	269	8	46
# Total	326	291	37	94

Table 3: Number of reviews that could not be analyzed

It is difficult to compare all tools because each one presents different characteristics, for that reason, several decisions have been made in order to be able to compare them.

Most of them process at document level, for that reason, the level used to compare them has been this one. Despite this, a few tools only process at sentence level. For these tools, all sentences of each document/text have been processed and the whole document has been labeled as positive or negative, depending on whether the number of positive sentences is greater than the negatives or not. Furthermore, in order to compute the final polarity rating of a whole document/text, all sentences from each text have been extracted and submitted to each tool, and the scores obtained for each sentence have been aggregated by using an arithmetic mean, obtaining a final single score.

The outputs of each tool are really simple, the classification results are returned as textual strings (positive, negative or neutral) and the polarity ratings as real numbers. However, the polarity ratings may be in the range $[0, 1]$ or in the range $[0, 10]$, thus, all polarity ratings have been normalized into the range $[-1, 1]$.

Besides, some tools do not classify texts as positive or negative in an explicit way but they compute polarity degrees. In these cases, the range of polarity previously normalized is divided into two, and those texts whose polarity is in the bottom half, $[-1, 0)$, are considered as negative, and positive whether they are in the top half, $(0, 1]$; if the polarity degree of a text is 0 then it is classified as a neutral text.

4.4. Results

Using these datasets, all tools have been tested obtaining the results shown in this subsection. Next, the results obtained, from the point of view of classification and polarity rating, are commented collection by collection:

4.4.1. Large movie review dataset

a) Classification

Starting with the Large movie review dataset, Table 4 shows the percentage of correctly classified reviews according to the measures mentioned in the subsection 4.2.

	Accuracy	Pos.Rec.	Neg.Rec.	Pos.Prec.	Neg.Prec.	Neutral
AlchemyAPI	73.6%	72.5%	74.8%	74.2%	73.1%	0
Lymbix	51.2%	65%	37.4%	65.1%	74.2%	249
Musicmetric	71.2%	66%	76.4%	**85.5**%	83.2%	147
Openamplify	66%	73.6%	58.2%	63.9%	69.6%	5
Opinion Crawl	61.3%	56%	66.7%	79.4%	72.2%	185
Opendover	66.1%	**90.5**%	41.8%	63.3%	**88.1**%	49
Repustate	60%	70.1%	49.8%	63.1%	71.2%	90
Semantria	74.3%	75.1%	73.6%	73.9%	74.9%	1
Sentiment140	64.6%	69%	60.2%	76.1%	77.1%	157
SentimentAnalyzer	**80.4**%	**80.3**%	**80.5**%	80.5%	80.3%	0
SentiRate	30.7%	47.8%	13.9%	78.3%	85.7%	579
Sentimetrix	58.9%	80.4%	42.9%	51%	74.8%	0
Uclassify	75.8%	74.6%	77%	76.4%	75.2%	0
ViralHeat	66%	57.8%	74.1%	69%	64.2%	0
Wingify	51.4%	69.2%	33.6%	61%	69.4%	189

Table 4: Summary of classification results for movie review dataset

Observing the results of Table 4, SentimentaAnalyzer is shown as the best tool to classify movies reviews (accuracy 80.4%). Considering only positive opinions, Opendover might seem the best one classifying positive texts (recall 90.5%); however, this data is due to the fact that most of results have been labeled as positive (precision 63.3%), and for that reason, the percentage of negative documents well-classified is low (recall 41.8%).

On the contrary, SentimentAnalyzer presents a more stable behavior, for both, positive and negative texts (positive recall 80.3%, negative recall 80.5%, positive precision 80.5%, negative precision 80.3% and accuracy 80.4%).

It is also necessary to remark the stable behavior of other tools such as AlchemyAPI, Semantria, Uclassify and Musicmetric, especially the last one, because it presents similar results (positive 85.5%, negative 83.2%) in spite of detecting 147 neutral values.

It is necessary to remind the reader that this collection has no neutral texts, for that reason, in Table 4 there are no results for neutral values in terms of precision and recall.

b) Polarity rating

Once the ability to classify documents/opinions has been assessed, the ability to score sentiments has been tested and assessed using equation (4). In addition to computing the total mse, the other two data, positive and negative mse, have been computed, considering only those reviews which have been classified in the original dataset as positive and negative. Remember that Sentiment140, Uclassify and ViralHeat do not compute polarity scores.

	Total mse	Positive mse	Negative mse
AlchemyAPI	**0.07**	**0.01**	0.14
Lymbix	0.11	0.06	0.18
Musicmetric	0.11	0.14	**0.07**
Openamplify	0.1	0.05	0.15
Opinion Crawl	0.1	0.1	0.11
Opendover	0.13	0.08	0.19
Repustate	0.16	0.12	0.2
Semantria	**0.07**	0.03	0.12
Sentiment140	-	-	-
SentimentAnalyzer	0.13	0.17	0.08
SentiRate	0.1	0.02	0.16
Sentimetrix	0.1	0.04	0.17
Uclassify	-	-	-
ViralHeat	-	-	-
Wingify	0.17	0.1	0.23

Table 5: MSE for movie review dataset

In this case, the reader can observe from Table 5 that AlchemyAPI and Semantria present the lowest errors (0.07). AlchemyAPI also presents the best results with respect to positive texts (0.05), whereas Musicmetric computes the lowest error for negative texts. On the other hand, the worst results were computed by Repustate (0.16) and Wingify (0.17). Opinion Crawl is a curious case because the error calculated is low (0.1) and the score for each opinion has been computed as the mean of the different scores of each sentence detected.

4.4.2. Tweets dataset

Regarding the other dataset, in Table 6 the percentages of well-classified tweets can be seen. In this case, the dataset has neutral texts and the effectiveness of the APIs with respect to these texts is computed as well. However, the Twitter dataset has not got polarity ratings, and as a consequence, it is not possible to compare the computed polarity scores.

	Accur.	Pos.Rec.	Neg.Rec.	Neu.Rec.	Pos.Pre.	Neg.Pre.	Neu.Pre.
AlchemyAPI	**62.5%**	68.1%	61%	57.2%	63.9%	72.4%	**51.2%**
Lymbix	61.2%	62%	45.7%	80%	66.8%	87%	47%
Musicmetric	49%	75.2%	48.5%	15.1%	50%	52.7%	33.3%
Openamplify	60.5%	54%	53.5%	77.7%	68.3%	77.7%	46.6%
Opinion Crawl	52.2%	37%	35%	**94%**	**80%**	**89.8**%	37.7%
Opendover	47.7%	44.5%	16.9%	91.3%	72.9%	81%	36.2%
Repustate	56.4%	61.5%	41%	69.7%	60.2%	72%	45.7%
Semantria	59.4%	56.1%	46.6%	78%	68.4%	79.6%	45.4%
Sentiment140	52%	41.2%	36.1%	86.3%	82.4%	84.2%	36.2%
SentimentAna.	52%	77.4%	62.7%	0.05%	50.5%	55.5%	36.8%
SentiRate	57.8%	63.4%	41.8%	70.5%	62.7%	75.5%	45.5%
Sentimetrix	55.6%	82.4%	**71.7%**	0%	50.1%	63.8%	0%
Uclassify	47.3%	66.5%	65%	0%	46.9%	47.9%	0%
ViralHeat	53%	75.8%	71.1%	0%	46.6%	62.3%	0%
Wingify	58.5%	**88.6%**	51%	10.2%	54.1%	77.6%	22.2%

Table 6: Summary of classification results for Twitter dataset

Table 6 demonstrates that AlchemyAPI presents the quite good results again (accuracy 62.5%). Other tools which show a stable behavior are Lymbix and Openamplify, which apparently do not present such good results in Table 4 due to the fact that movies dataset does not have neutral documents.

Despite the fact that other services present better results for a precise polarity (positive, negative or neutral), e.g., Wingify computes the best results for positive recall (88.6%), Sentimetrix the best result for negative recall

(88.6%) and Opinion Crawl for neutral recall (94%), these tools present irregular behaviors with different polarities because they classify many texts under the same sign but at the same time, they are misclassifying many other texts as their low precisions demonstrate. Thus, for example, Sentimetrix shows very good results for positive and negative values considering only the recall, 82.4% and 71.7% respectively, due to the fact that it does not compute neutral values, and consequently the precision values descrease to a great extent (positive precision 50.1%, negative precision 63.8%). And, despite the fact that Winfigy computes neutral texts, the percentage of well-classified documents is really low (neutral recall 10.2%).

In this case, the tools which show worse performance are Opendover, Musicmetric, Opinion Crawl, Sentiment140 and Sentimentanalyzer. Even, in spite of the fact that some of them like Opinion Crawl and Sentiment140 present very good results from the point of view of precision for some polarities (positive, negative or neutral), however, their results from the point of view of the recall demonstrate that they do not present a stable behaviour because they tend to classify well a specific sign, but not the rest of them. The case of Sentiment140 and SentiRate is especially curious, because they use learning models trained with tweets, and their results are not really the best. Moreover, it is necessary to remark that it is difficult to compare Sentimetrix, Uclassify and Viralheat with the rest of tools because their APIs do not detect neutral texts.

4.4.3. Amazon dataset (complete reviews)

This subsection includes the results of submitting the whole Amazon reviews.

a) Classification

Considering the whole reviews from the Amazon dataset, Table 7 summarizes all collected results for each tool.

As can be seen, SentimentAnalyzer seems to present the best results (accuracy 71.5%, negative recall 78.8%); however, this fact may be misleading because its capability of classifying neutral text is lacking. Despite this, it presents the best balance with respect to positive reviews (recall 66.4% and precision 57.4%). A similar problem occurs with SentiRate, it presents very good results with respect to neutral text (neutral recall 73.7%), because it tends to classify every text as neutral (neutral precision 5.1%), misclassifying the rest of types as a consequence. The best

	Acc.	Pos.Rec.	Neg.Rec.	Neu.Rec.	Pos.Pre.	Neg.Pre	Neu.Pre.
AlchemyAPI	60%	60.1%	63.5%	4.59%	41.1%	75.3%	16.6%
Lymbix	32%	41.4%	28.7%	0.3%	33.9%	77.5%	3.3%
Musicmetric	68.4%	56.2%	77.7%	15%	54.5%	78.2%	16.6%
Openamplify	49%	59.3%	47%	10%	32.9%	73.5%	9.5%
Opinion Crawl	53%	42.2%	58.8%	40%	45.7%	77.6%	8%
Opendover	52.3%	49.5%	52.3%	15%	37.2%	70.2%	3%
Repustate	48%	66.4%	42.7%	5%	35.4%	72.8%	4%
Semantria	57.8%	64%	58.3%	10%	40%	75.3%	25%
Sentiment140	45%	44.5%	47.4%	15%	39.8%	72.1%	2.8%
SentimentAn.	**71.5%**	66.4%	**78.8%**	0%	**57.4**%	78.7%	0%
SentiRate	22.6%	41.7%	10.6%	**73.7%**	44.1%	**85.7**%	5.1%
Sentimetrix	49.4%	**68.7%**	44.2%	0%	33.8%	72.3%	0%
Uclassify	64.5%	67.1%	67.8%	0%	45.5%	79%	0%
ViralHeat	58.5%	30%	75.5%	0%	33.3%	67.4%	0%
Wingify	42.6%	64.8%	34.7%	15.8%	20.1%	75%	**37.9**%

Table 7: Summary of classification results for Amazon dataset using whole reviews

score regarding positive hits (68.7%) is achieved by Sentimetrix, however, this tool does not achieve quite good results with respect to the rest of parameters.

It is worth noting the stable behaviour of Alchemy and Semantria, but in this case, it is especially necessary to highlight Musicmetric, which seems to deal with text containing irony in an effective manner.

Considering all columns, the worst results are achieved by Lymbix and SentiRate by far, and observing in particular the last column of the table, it is apparent that very few tools deal with neutral texts in a correct way. It is also remarkable the fact that most of tools have more problems dealing with positive documents than negatives.

b) Polarity rating

It can be seen from the data in Table 8 that the behaviour of the web services is very similar to the one shown in Table 7. SentimentAnalyzer obtains the best results with respect to all reviews (0.16) and the negatives (0.13), and also, SentiRate regarding neutral texts (0.02), along with AlchemyAPI. In this case, Lymbix presents the best result for positive reviews (0.05).

In Table 8 the most significant data are provided by the neutral texts. Reading the data from Table 7, many tools do not classify correctly neutral texts because they only work with two values, positive or negative, but even

	Total mse	Positive mse	Negative mse	Neutral mse
AlchemyAPI	0.2	0.21	0.21	**0.01**
Lymbix	0.2	**0.05**	0.29	0.02
Musicmetric	0.23	0.35	0.17	0.16
Openamplify	0.23	0.2	0.25	0.02
Opinion Crawl	0.21	0.29	0.18	0.04
Opendover	0.18	0.21	0.28	0.19
Repustate	0.26	0.2	0.3	0.05
Semantria	0.2	0.18	0.23	0.02
Sentiment140	-	-	-	-
SentimentAn.	**0.16**	0.24	**0.13**	0.12
SentiRate	0.22	0.18	0.26	**0.01**
Sentimetrix	0.23	0.18	0.26	0.03
Uclassify	-	-	-	-
ViralHeat	-	-	-	-
Wingify	0.27	0.2	0.31	0.04

Table 8: MSE for Amazon dataset using the whole reviews

so, the ratings demonstrate that these tools are able to score them as neutral texts. It is also necessary to remember that the number of neutral texts (20) available in this collection is not excessive.

From a broader point of view, rather than seeing individual data, SentimentAnalyzer seems the most stable service, because in spite of not classifying explicitly the texts as positive, neutral or negative; it rates well.

4.4.4. Amazon dataset (ironic sentences)

The results presented here have been collected by submitting only the ironic sentences extract from the whole Amazon reviews. Observing the results from Table 9 and 10, they are very similar to Table 7 and 8, but the results from the whole texts are slightly better, generally. This fact may be because of the effect of context, large texts provide more information than short texts.

a) Classification

b) Polarity rating

	Acc.	Pos.Rec.	Neg.Rec.	Neu.Rec.	Pos.Pre	Neg.Pre	Neu.Pre
AlchemyAPI	53.3%	53.5%	56.3%	10%	36.1%	71.2%	**10.5%**
Lymbix	27.7%	32%	24.5%	0.45%	33.6%	70.2%	4.2%
Musicmetric	64.8%	63.2 %	69.9%	5%	**57.6%**	**79.1%**	2.4%
Openamplify	42.9%	51.6%	41%	15%	34.5%	72.9%	3.6%
Opinion Crawl	40%	39%	40.1%	40%	45.8%	73.4%	4.7%
Opendover	55%	60.2%	50.2%	20%	37.4%	38.7%	2%
Repustate	42.5%	62.6 %	35.9%	10%	34.3%	69.3%	3.7%
Semantria	48.6%	58.6%	46.5%	15%	36.2%	70.3%	7.8%
Sentiment140	31.1%	33%	29%	50%	36.5%	66.4%	5%
SentimentAn.	**71.5%**	64.2%	**79.6%**	0%	56.6%	78.8%	0%
SentiRate	23.6%	37%	14.3%	**70%**	40%	87.2%	5%
Sentimetrix	52.6%	**65.3%**	50.7%	0%	35.4%	72.8%	0%
Uclassify	59.2%	62.5 %	61.9%	0%	40%	75.5%	0%
ViralHeat	55.9%	36.7 %	68.1%	0%	33%	67.7%	0%
Wingify	40.4%	64.2%	31.9%	15.8%	32.4%	70.9%	5.7%

Table 9: Summary of classification results for Amazon dataset using only ironic sentences

4.5. Final remarks

Summarizing, AlchemyAPI and Semantria seem the most regular tools classifying both, short and large texts, and predicting their corresponding polarity ratings. Working specifically with larger texts, SentimentAnalyzer presents the better results classifying, especially when there are no neutral documents. The main weakness of these tools is clearly to deal with texts expressing neutral information, most of them do not even detect neutral information and, for that reason, it is better to analyze the polarity ratings in order to know if the content tends to be neutral or not.

Besides, despite the third collection containing ironic information, which could be considered as more complicated to detect, it is remarkable the fact that most of tools showed a similar behaviour to that observed in the other two collections.

On the other hand, it is difficult to select those tools whose performance is more irregular; Wingify or Viralheat seem to be the two services which do not offer really interesting results for both activities, classification and rating prediction.

And finally, it is necessary to point out that tools like Musicmetric or Uclassify present good results, especially working with large texts/opinions on different topics. This fact seems more remarkable because they are based on matching learning algorithms trained with collections specialized in spe-

	Total mse	Positive mse	Negative mse	Neutral mse
AlchemyAPI	0.21	0.22	0.22	**0.01**
Lymbix	0.2	**0.06**	0.28	0.02
Musicmetric	0.22	0.28	0.19	0.19
Openamplify	0.23	0.2	0.25	0.02
Opinion Crawl	0.28	0.32	0.27	0.11
Opendover	0.28	0.23	0.3	0.16
Repustate	0.27	0.2	0.31	0.08
Semantria	0.24	0.22	0.26	0.03
Sentiment140	-	-	-	-
SentimentAna.	**0.16**	0.23	**0.13**	0.15
SentiRate	0.22	0.2	0.24	**0.01**
Sentimetrix	0.23	0.2	0.2	0.03
Uclassify	-	-	-	-
ViralHeat	-	-	-	-
Wingify	0.32	0.21	0.37	0.06

Table 10: MSE for Amazon dataset using only ironic sentences

cific domains such as music or Amazon products reviews, which are not necessarily the domains of the data collections used to test them.

A summary including the tools which present better behaviour according to different aspects are presented in table 11.

Aspect	**Tools**
Short texts	AlchemyAPI, Lymbix, OpenAmplify, Semantria, Winfigy
Large texts	SentimentAnalyzer, AlchemyAPI, Semantria, Uclassify, Musicmetric
Irony	SentimentAnalyzer, Musicmetric, Uclassify, ViralHeat, AlchemyAPI
Polarity	SentimentAnalyzer, AlchemyAPI, Lymbix, Semantria, Uclassify
Rating	SentimentAnalyzer, AlchemyAPI, Lymbix, SentiRate, Sentimetrix

Table 11: Best tools according to different aspects

4.6. An experiment applied to a real scenario

The previous subsections have shown the effectiveness of the different tools using standard collections; nonetheless, in other applications, tools for evaluating sentiments have been used to address real problems, for example, estimating votes in elections through Tweets (Tumasjan et al., 2010), recommending products according to customers' opinions (Garcia Esparza et al.,

2012) or analyzing stock markets (Li et al., 2014; Smailović et al., 2014). Thus, following that idea, and with the aim of analyzing the usefulness of these tools to recommend products in a real scenario as many other works propose (Porcel et al., 2012; Tejeda-Lorente et al., 2014; Serrano-Guerrero et al., 2011, 2013), an experiment working with real comments collected from an important website is presented. This experiment will allow the reader to see the capability of these tools to recommend movies from real users' comments.

Thus, in order to present a more real scenario, the well-known page IMDb[35] has been chosen. The idea is to assess the ability of these web services to recommend movies through the comments collected from real users, and compare whether or not these recommendations match the recommendations made by IMBd's regular voters and to what extent.

IMDb is one of the world's most popular sources for movies, TV and celebrity content. It mainly provides information related to the most famous series, movies, TV programs and video games. In this case, the movies section is the most interesting because it allows users to interact with this page by rating and inserting comments about the different movies. The opinions are free texts without any restrictions in terms of length, and the ratings are expressed through a scale of 10 starts, where 10 starts indicate the most positive opinion and zero the worst. Both are not strictly necessary, i.e., a user can rate a movie without including any opinions, and vice versa, comment a movie without rating it.

From those ratings, the webpage presents a list containing the top rated movies; although, as indicated at the IMDb page, only votes from regular IMDb voters, who are unknown, are used to rank movies. Therefore, thousands of opinions, not necessarily rated, about movies are available and a listing of the top rated films.

Thus, starting from that information, this experiment consists in assessing the capability of the different web services to rank the 100 top rated movies according to IMDb[36], only working with the users' opinions. And for that purpose, the $1,000$ most recent opinions from the 100 first movies of the IMDb ranking have been collected, but the ratings have been omitted. It is necessary to remark that these opinions not necessarily belong to any of the

[35] http://www.imdb.com

[36] http://www.imdb.com/chart/top?ref_=nv_ch_250_4

regular IMDb voters, and most of them do not even include any ratings. The statistics for those opinions are: $2,708,630$ words, $149,017$ sentences, 270.86 words/opinion and 14.9 sentences/opinion.

Once those data were collected, every opinion was submitted to a particular web service in order to compute the corresponding rating. And once all ratings for a specific movie were computed, the average of them was calculated. The process was repeated with the 100 movies, and finally a list of movies ordered by the average rating was made. This list is supposed to contain the 100 most recommendable movies according to that web service.

To assess the quality of this new list of recommendable movies, it is necessary to compare it with the list provided by IMDb, which is considered in this case as the ideal list. For that purpose, the evaluation measure chosen is the Spearman correlation coefficient Kendall and Gibbons (1990), broadly used in domains like Information Retrieval to compare rankings of documents. It is defined as:

$$R = 1 - \frac{6 * \sum d_i^2}{n^3 - n}$$

where d_i is the rank difference of the common document i, in this case the movie i, and n stands for the number of documents. Two rankings are identical when the coeficient is 1, and in reverse order when the coefficient is -1.

Hence, taking into account only those web services which are able to rate opinions (Sentiment140, Uclassify and ViralHeat do not compute ratings), Table 12 shows the results obtained after submitting all opinions to the different tools.

As can be seen in this table, the tools that present the most similar recommendations with respect to the IMDb's list are SentimentAnalyzer, Lymbix, AlchemyAPI and Sentimetrix. To interpret these results it is necessary to analysis Table 11, where the best tools were proposed according to different criteria. As the aim of this experiment is to rate large opinions, it is necessary to pay special attention to the second (*large texts*) and fifth (*rating*) rows, where SentimentAnalyzer and Alchemy are expected to be two of the most suitable tools in this case; and as can be seen in Table 12, they are, appearing in the first and third positions, respectively.

Apart from these tools, Sentimetrix and Lymbix appear within the first positions of Table 12. This fact is easily explainable because of its great capability to deal with ratings, as the last row of Table 11 corroborates.

Tool	**Spearman correlation**
SentimentAnalyzer	0.5
Lymbix	0.42
AlchemyAPI	0.36
Sentimetrix	0.35
Repustate	0.33
Semantria	0.28
Musicmetric	0.25
Wingify	0.22
SentiRate	0.18
Opendover	0.18
Openamplify	0.16
Opinion Crawl	0.15

Table 12: Spearman correlation for each web service

Therefore, this experiment seems useful to confirm the analysis performed from the results obtained in the previous subsections; and it may confirm the usefulness of this work as an interesting guide for those researchers who need a tool for carrying out his/her experiments on this research field.

5. Conclusions

This work presents a detailed review of 15 web services which include functionalities related to Sentiment Analysis. Some of these services belong to private companies, but even so, they allow restricted free access to their functionalities, and the others are totally free services. This fact is interesting to those users/researchers who desire to include Sentiment Analysis capabilities within their own platforms without having to develop their own algorithms; hence, these tools are especially interesting for researching purposes and rapid prototyping. Besides, due to the fact that the selected services can work as web services, the inclusion of them into any platform is really easy.

Moreover, in order to facilitate the task of selecting the most appropriate service depending on the user needs, the capabilities of these services related to Sentiment Analysis have been assessed under different circumstances. For this purpose three different collections have been chosen containing information of a different nature: large and short texts; positive, negative, neutral information; and even ironic contents. Such collections have been used to assess the classification and polarity rating capabilities of the proposed tools.

From the results obtained, services such as Alchemy and Semantria could be taken into account for any kind of text. SentimentAnalysis may be really interesting to the user if the analyzed texts are quite large and you want to classify them as positive or negative. Musicmetric and Uclassify are other tools that could be considered. All these tools could also be considered a good option if the texts contain ironic sentences.

On the other hand, the findings of this work may discard tools like Wingify or Viralheat because of the obtained results. It is also necessary to comment that the main disadvantage of all these tools is that they are unable to obtain satisfactory results working with neutral texts. And furthermore, these tools still have to deal with several challenges such as the explicit detection of subjectivity within texts.

To corroborate these statements, a real scenario has been proposed to check whether the results and conclusions extracted from the experiments using standard collections. This new experiment has demonstrated that those tools that were supposed to be the best with respect to the prior experiments, were really the most interesting tools for the proposed scenario.

Therefore, from this work any user/researcher has enough information about the services offered and the possible results expected from them, to decide the most appropriate one for his/her interests.

As a final concluding remark, it is necessary to comment that these tools have a lot of challenges ahead. They suffer from problems like the excessive simplicity while classifying, generally, only positive, negative or neutral categories are used; or the incapability to aggregate ratings from different sentences or paragraphs, in order to get a general rate about a complete opinion. These two examples might be mitigated to a certain extent through the use of techniques like *fuzzy logic*, which enables systems to classify items into more precise different categories or aggregate information in a more comprehensive manner by using the ordered weighted averaging operators, for example.

Acknowledgments

This work has been supported by the national projects with the financing FEDER funds, TIN2010-17876 and TIN2013-40658-P, and also with the Andalusian Excellence Projects TIC-05299 and TIC-5991.

References

Abbasi, A., 2010. Intelligent Feature Selection for Opinion Classification. IEEE Intelligent Systems 25, 75–79.

Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages. ACM Transactions on Information Systems 26, 1–34.

Abbasi, A., France, S., Zhang, Z., Chen, H., 2011. Selecting Attributes for Sentiment Classification Using Feature Relation Networks. IEEE Transactions on Knowledge and Data Engineering 23, 447–462.

Agarwal, A., Biadsy, F., Mckeown, K.R., 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic N-grams, in: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), pp. 24–32.

Baccianella, S., Esuli, A., Sebastiani, F., 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, in: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (Eds.), Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), Valletta, Malta. pp. 2200–2204.

Banea, C., Mihalcea, R., Wiebe, J., 2010. Multilingual subjectivity: are more languages better?, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 28–36.

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., Goldstein, G., 2011. Identifying and following expert investors in stock microblogs, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), pp. 1310–1319.

Barbosa, L., Feng, J., 2010. Robust sentiment detection on Twitter from biased and noisy data, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 36–44.

Beineke, P., Hastie, T., Manning, C., Vaithyanathan, S., 2004. Exploring Sentiment Summarization, in: Shanahan, J.G., Wiebe, J., Qu, Y. (Eds.), Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications, AAAI Press. pp. 1–4.

Boiy, E., Moens, M.F., 2008. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval 12, 526–558.

Bollen, J., Mao, H., 2011. Twitter Mood as a Stock Market Predictor. Journal of Computational Science 44, 91–94.

Cambria, E., Schuller, B., Xia, Y., Havasi, C., 2013. New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems 28, 15–21.

Cao, Q., Duan, W., Gan, Q., 2011. Exploring determinants of voting for the helpfulness of online user reviews: A text mining approach. Decision Support Systems 50, 511–521.

Castellanos, M., Dayal, U., Hsu, M., Ghosh, R., Dekhil, M., Lu, Y., Zhang, L., Schreiman, M., 2011. LCI: a social channel analysis platform for live customer intelligence, in: Proceedings of the 2011 international conference on Management of data - SIGMOD '11, ACM Press, New York, New York, USA. pp. 1049–1058.

Chen, B., Zhu, L., Kifer, D., Lee, D., 2010. What is an Opinion About? Exploring Political Standpoints using Opinion Scoring Model, in: Proceeedings of AAAI Conference on Artificial Intelligence (AAAI-2010), pp. 1007–1012.

Chenlo, J.M., Losada, D.E., 2014. An empirical study of sentence features for subjectivity and polarity classification. Information Sciences 280, 275–288.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A., 1990. Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41, 391–407.

Duric, A., Song, F., 2012. Feature selection for sentiment analysis based on content and syntax models. Decision Support Systems 53, 704–711.

Esuli, A., Sebastiani, F., 2006. Determining Term Subjectivity and Term Orientation for Opinion Mining, in: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL06), pp. 193–200.

Feldman, R., 2013. Techniques and applications for sentiment analysis. Communications of the ACM 56, 82–89.

Filatova, E., 2012. Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey. pp. 392–398.

Finn, A., Kushmerick, N., 2006. Learning to classify documents according to genre: Special Topic Section on Computational Analysis of Style. Journal of the American Society for Information Science and Technology 57, 1506–1518.

Ganesan, K., Zhai, C., Han, J., 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 340–348.

Ganesan, K., Zhai, C., Viegas, E., 2012. Micropinion generation: An Unsupervised Approach to Generating Ultra-Concise Summaries of Opinions, in: Proceedings of the 21st international conference on World Wide Web - WWW '12, ACM Press. pp. 869 – 878.

Garcia Esparza, S., OMahony, M., Smyth, B., 2012. Mining the real-time web: A novel approach to product recommendation. Knowledge-Based Systems 29, 3–11.

Gerani, S., Carman, M., Crestani, F., 2012. Aggregation Methods for Proximity-Based Opinion Retrieval. ACM Transactions on Information Systems 30, 1–36.

Go, A., Bhayani, R., Huang, L., 2009. Twitter sentiment classification using distant supervision. Technical Report. Standford University.

Groh, G., Hauffa, J., 2011. Characterizing Social Relations Via NLP-Based Sentiment Analysis, in: Adamic, L.A., Baeza-Yates, R.A., Counts, S. (Eds.), ICWSM, The AAAI Press. pp. 502–505.

Guo, L., Wan, X., 2012. Exploiting syntactic and semantic relationships between terms for opinion retrieval. Journal of the American Society for Information Science and Technology 63, 2269–2282.

He, Y., Lin, C., Alani, H., 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification, in: Proceedings of the 49th

Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), pp. 123–131.

He, Y., Zhou, D., 2011. Self-training from labeled features for sentiment analysis. Information Processing and Management 47, 606–616.

Jindal, N., Liu, B., 2006a. Identifying comparative sentences in text documents, in: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06, ACM Press, New York, New York, USA. pp. 244 – 251.

Jindal, N., Liu, B., 2006b. Mining comparative sentences and relations, in: Proceedings of the 21st national conference on Artificial intelligence (AAAI'06), pp. 1331–1336.

Jindal, N., Liu, B., 2007. Review spam detection, in: Proceedings of the 16th international conference on World Wide Web (WWW '07), ACM Press. pp. 1189–1190.

Kendall, M., Gibbons, J., 1990. M. Kendall, J.D. Gibbons. Oxford University Press.

Kim, J., Li, J.J., Lee, J.H., 2010. Evaluating multilanguage-comparability of subjectivity analysis systems, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), pp. 595–603.

Kim, K., Lee, J., 2014. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction. Pattern Recognition 47, 758–768.

König, A.C., Brill, E., 2006. Reducing the human overhead in text categorization, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06, ACM Press. pp. 598–603.

Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N., 2013. Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications 40, 4065–4074.

Lee, S.W., Song, Y.I., Lee, J.T., Han, K.S., Rim, H.C., 2011. A new generative opinion retrieval model integrating multiple ranking factors. Journal of Intelligent Information Systems 38, 487–505.

Li, G., Liu, F., 2012. Application of a clustering method on sentiment analysis. Journal of Information Science 38, 127–139.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., Chen, Y., 2014. The effect of news and public mood on stock movements. Information Sciences 278, 826–840.

Li, S.T., Tsai, F.C., 2013. A fuzzy conceptualization model for text mining with application in opinion polarity classification. Knowledge-Based Systems 39, 23–33.

Liu, B., 2010. Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing 5, 1–38.

Liu, B., 2011. Exploring Hyperlinks, Contents, and Usage Data. Springer.

Liu, B., Zhang, L., 2012. A Suvery of opinion mining and sentiment analysis, in: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data. Springer US, Boston, MA. chapter 13, pp. 415–464.

Loia, V., Senatore, S., 2014. A fuzzy-oriented sentic analysis to capture the human emotion in Web-based content. Knowledge-Based Systems 58, 75–85.

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C., 2011. Learning Word Vectors for Sentiment Analysis, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA. pp. 142–150.

Maks, I., Vossen, P., 2012. A lexicon model for deep sentiment analysis and opinion mining applications. Decision Support Systems 53, 680–688.

Martin-Valdivia, M.T., Martinez-Cámara, E., Perea-Ortega, J.M., Ureña Lopez, L.A., 2012. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications 40, 3934–3942.

McGlohon, M., Glance, N., Reiter, Z., 2010. Star Quality: Aggregating Reviews to Rank Products and Merchants, in: Proceedings of Fourth International Conference on Weblogs and Social Media (ICWSM), pp. 114–121.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal .

Miller, G.A., 1995. WordNet: a lexical database for English. Communications of the ACM 38, 39–41.

Mohammad, S.M., 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems 53, 730–741.

Montejo-Raez, A., Martinez-Camara, E., Martin-Valdivia, M.T., Ureña Lopez, L.A., 2014. Ranked WordNet graph for Sentiment Polarity Classification in Twitter. Computer Speech & Language 28, 93–107.

Montesi, M., Navarrete, T., 2008. Classifying web genres in context: A case study documenting the web genres used by a software engineer. Information Processing and Management 44, 1410–1430.

Montoyo, A., Martínez-Barco, P., Balahur, A., 2012. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. Decision Support Systems 53, 675–679.

Moraes, R., Valiati, J, F., Gavião Neto, W.P., 2013. Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications 40, 621–633.

Moreo, A., Romero, M., Castro, J., Zurita, J., 2012. Lexicon-based Comments-oriented News Sentiment Analyzer system. Expert Systems with Applications 39, 9166–9180.

Mukherjee, A., Liu, B., Glance, N., 2012. Spotting fake reviewer groups in consumer reviews, in: Proceedings of the 21st international conference on World Wide Web (WWW '12), ACM Press. pp. 191–200.

Nigam, K., Hurst, M., 2006. Towards a Robust Metric of Polarity, in: Shanahan, J.G., Qu, Y., Wiebe, J. (Eds.), Computing Attitude and Affect in Text: Theory and Applications. Springer-Verlag, Berlin/Heidelberg. volume 20 of *The Information Retrieval Series*, pp. 265–279.

Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G., 2010a. Opinion summarization with integer linear programming formulation for sentence extraction and ordering, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 910–918.

Nishikawa, H., Hasegawa, T., Matsuo, Y., Kikui, G., 2010b. Optimizing informativeness and readability for sentiment summarization, in: Proceedings of the Association for Computational Linguistics (ACL'10), pp. 325–330.

Ojokoh, B.A., Kayode, O., 2012. A feature-opinion extraction approach to opinion mining. Journal of Web Engineering 11, 51–63.

Ott, M., Cardie, C., Hancock, J., 2012. Estimating the prevalence of deception in online review communities, in: Proceedings of the 21st international conference on World Wide Web (WWW '12), ACM Press, New York, New York, USA. pp. 201–210.

Paltoglou, G., Thelwall, M., 2010. A study of information retrieval weighting schemes for sentiment analysis, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), pp. 1386–1395.

Pang, B., Lee, L., 2004. A sentimental education, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04), Association for Computational Linguistics, Morristown, NJ, USA. pp. 271–es.

Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. volume 2. Foundations and Trends in Information Retrieval.

Park, K.M., Park, H., Kim, H.G., Ko, H., 2012. Review summarization based on linguistic knowledge, in: Yu, H., Yu, G., Hsu, W., Moon, Y.S., Unland, R., Yoo, J. (Eds.), Proceedings of the 17th international conference on Database Systems for Advanced Applications (DASFAA'12), Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 105–114.

Porcel, C., Tejeda-Lorente, A., Martínez, M., Herrera-Viedma, E., 2012. A hybrid recommender system for the selective dissemination of research resources in a Technology Transfer Office. Information Sciences 184, 1–19.

Raaijmakers, S., Kraaij, W., 2008. A Shallow Approach to Subjectivity Classification, in: Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM '08), pp. 216–217.

Rao, Y., Li, Q., Mao, X., Wenyin, L., 2014. Sentiment topic models for social emotion mining. Information Sciences 266, 90–100.

Remus, R., Hänig, C., 2011. Towards well-grounded phrase-level polarity analysis, in: Proceedings of the 12th international conference on Computational linguistics and intelligent text processing (CICLing'11), pp. 380–392.

Reyes, A., Rosso, P., 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. Decision Support Systems 53, 754–760.

Reyes, A., Rosso, P., Buscaldi, D., 2012. From humor recognition to irony detection: The figurative language of social media. Data and Knowledge Engineering 74, 1–12.

Rushdi-Saleh, M., Martín-Valdivia, M., Montejo-Ráez, A., Ureña López, L., 2011. Experiments with SVM to classify opinions in different domains. Expert Systems with Applications 38, 14799–14804.

Sarvabhotla, K., Pingali, P., Varma, V., 2011. Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents. Information Retrieval 14, 337–353.

Savoy, J., 2012. Authorship Attribution Based on Specific Vocabulary. ACM Transactions on Information Systems 30, 1–30.

Seki, Y., Kando, N., Aono, M., 2009. Multilingual opinion holder identification using author and authority viewpoints. Information Processing and Management 45, 189–199.

Serrano-Guerrero, J., Herrera-Viedma, E., Olivas, J.A., Cerezo, A., Romero, F.P., 2011. A Google Wave-based Fuzzy Recommender System to disseminate Information in University Digital Libraries 2.0. Information Sciences 181, 1503 – 1516.

Serrano-Guerrero, J., Romero, F.P., Olivas, J.A., 2013. Hiperion: A fuzzy approach for recommending educational activities based on the acquisition of competences. Information Sciences 248, 114–129.

Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M., 2014. Stream-based active learning for sentiment analysis in the financial domain. Information Sciences 285, 181–203.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37, 267–307.

Tang, H., Tan, S., Cheng, X., 2009. A survey on sentiment detection of reviews. Expert Systems with Applications 36, 10760–10773.

Tata, S., Di Eugenio, B., 2010. Generating Fine-Grained Reviews of Songs from Album Reviews, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden. pp. 1376–1385.

Tejeda-Lorente, A., Porcel, C., Peis, E., Sanz, R., Herrera-Viedma, E., 2014. A quality based recommender system to disseminate information in a university digital library. Information Sciences 261, 52–69.

Thet, T.T., Na, J.C., Khoo, C.S.G., 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. Journal of Information Science 36, 823–848.

Tsytsarau, M., Palpanas, T., 2011. Survey on mining subjective data on the web. Data Mining and Knowledge Discovery 24, 478–514.

Tumasjan, A., Sprenger, T.O., Sandner, P.G., Isabell Welpe, M., 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment, in: Proceeedings of International Conference on Weblogs and Social Media (ICWSM-2010), pp. 178–185.

Turney, P.D., 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, pp. 417–424.

Vechtomova, O., 2010. Facet-based opinion retrieval from blogs. Information Processing and Management 46, 71–88.

Vechtomova, O., Karamuftuoglu, M., 2008. Lexical cohesion and term proximity in document ranking. Information Processing and Management 44, 1485–1502.

Wang, D., Zhu, S., Li, T., 2013. SumView: A Web-based engine for summarizing product reviews and customer opinions. Expert Systems with Applications 40, 27–33.

Wang, G., Xie, S., Liu, B., Yu, P.S., 2012. Identify Online Store Review Spammers via Social Review Graph. ACM Transactions on Intelligent Systems and Technology 3, 1–21.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S., 2005a. OpinionFinder, in: Proceedings of HLT/EMNLP on Interactive Demonstrations -, Association for Computational Linguistics. pp. 34–35.

Wilson, T., Wiebe, J., Hoffmann, P., 2005b. Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), Association for Computational Linguistics, Morristown, NJ, USA. pp. 347–354.

Wilson, T., Wiebe, J., Hoffmann, P., 2009. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. Computational Linguistics 35, 399–433.

Xianghua, F., Guo, L., Yanyan, G., Zhiqiang, W., 2013. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon. Knowledge-Based Systems 37, 186–195.

Xie, S., Wang, G., Lin, S., Yu, P.S., 2012. Review spam detection via temporal pattern discovery, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, New York, USA. pp. 823–831.

Yang, S., Ko, Y., 2011. Finding relevant features for Korean comparative sentence extraction. Pattern Recognition Letters 32, 293–296.

Ye, Q., Zhang, Z., Law, R., 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Systems with Applications 36, 6527–6535.

Yessenalina, A., Yue, Y., Cardie, C., 2010. Multi-level structured models for document-level sentiment classification, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10), pp. 1046–1056.

Yu, L.C., Wu, J.L., Chang, P.C., Chu, H.S., 2013. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowledge-Based Systems 41, 89–97.

Zhang, C., Zeng, D., Li, J., Wang, F.Y., Zuo, W., 2009. Sentiment analysis of Chinese documents: From sentence to document level. Journal of the American Society for Information Science and Technology 60, 2474–2487.

Zhang, W., Xu, H., Wan, W., 2012. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. Expert Systems with Applications 39, 10283–10291.

Zhang, W., Yoshida, T., Tang, X., 2008. Text classification based on multi-word with support vector machine. Knowledge-Based Systems 21, 879–886.

Zhou, L., Chaovalit, P., 2008. Ontology-supported polarity mining. Journal of the American Society for Information Science and Technology 59, 98–110.

Zhou, S., Chen, Q., Wang, X., 2010. Active deep networks for semi-supervised sentiment classification, in: Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10), pp. 1515–1523.