A Knowledge-Based Approach for Polarity Classification in Twitter

Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López
Computer Science Department, University of Jaén, Campus Las Lagunillas, 23071, Jaén, Spain.
E-mail: {amontejo, emcamara, maite, laurena}@ujaen.es

Until now, most of the methods published for polarity classification in Twitter have used a supervised approach. The differences between them are only the features selected and the method used for weighting them. In this article, we present an unsupervised method for polarity classification in Twitter. The method is based on the expansion of the concepts expressed in the tweets through the application of PageRank to WordNet. In addition, we integrate SentiWordNet to compute the final value of polarity. The synsets values are weighted with the PageRank scores obtained in the previous random walk process over WordNet. The results obtained show that disambiguation and expansion are good strategies for improving overall performance.

Introduction

Today, the use of the Internet is broadly extended in society, and the ease of publishing new content has turned the web from a static container of information into a live environment in which any user can publish any type of information. Users utilize different platforms such as blogs, forums, wikis, or social networks to share their knowledge and to interact with other users.

However, in the current scenario centered around the concept of Web 2.0, users not only participate in content creation but also comment, review, or express their opinions on a wide range of topics such as politics, movies, commercial and financial products, literature, and so on. The knowledge implicit in the opinions of the users is important for companies, public administration officials, and political parties because it is a rich source of feedback. Therefore, the idea of processing these comments or reviews automatically has attracted researchers in the field of text mining, with a focus on extracting a general opinion about a particular item among the unstructured data available on the Internet. The task that focuses on the automatic treatment of opinions, sentiments, and subjectivity in text is called *sentiment analysis* (SA), which also is known as opinion mining (Pang & Lee, 2008). The SA task is a wide and interesting research field that covers an increasing number of applications. Although there are several topics related to SA, the following two subtasks stand out: subjectivity classification and polarity or sentiment classification. *Subjectivity classification* encompasses all the techniques related to the classification of sentences or documents as subjective or opinionated. Some of the most relevant works related to subjectivity classification can be found in Pang and Lee (2004), Wiebe and Riloff (2005), and Banea, Mihalcea, Wiebe, and Hassan (2008). *Polarity classification* refers to the classification of an opinionated document as expressing a positive or negative opinion. Polarity classification has been studied more in depth than has subjectivity classification, as demonstrated by the extensive list of published papers about this task (Blitzer, Dredze, & Pereira, 2007; Prabowo & Thelwall, 2009; Zirn, Niepert Stuckenschmidt, & Strube, 2011). A more complete state of the art about SA can be found in Liu (2012), Pang and Lee (2008), and Tsytsarau and Palpanas (2012).

The accuracy of polarity classification can be influenced by the context or domain of the terms that appear in the documents. Some words or phrases can express different sentiments in different domains. A simple example is shown in Turney (2002). The word "unpredictable" may have a positive semantic orientation in a movie review, but could have a negative meaning in an opinion about a car's steering in a specific model. Thus, it is important to consider the right meaning of each word for the determination of the polarity of a document. The task of computationally determining which "sense" of a word is activated by the use of the word in a particular context is called *word sense disambiguation* (WSD; Agirre & Edmonds, 2006). The problem of WSD can be addressed using different strategies, such as supervised (Pradhan, Loper, Dligach, & Palmer, 2007) or knowledge-based WSD (Lesk, 1986). Some years ago,

Received November 29, 2012; revised February 8, 2013; accepted April 2, 2013

 • Published online 26 November 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.22984

graph-based methods for knowledge-based WSD gained attention (Mihalcea, 2005). These used well-known, graph-based techniques to find and exploit the structural properties of the graph underlying a particular lexical knowledge base (LKB). As noted in Agirre and Soroa (2009) and Sinha and Mihalcea (2007), graph-based methods for WSD are suitable for the disambiguation of word sequences, and they manage to exploit the interrelations among the senses in a given context.

On the other hand, social networks are a rich source of data for SA, which allow millions of users to publish information about any topic in a simple way and to share it with their networks of contacts or "friends." Moreover, these social networks have evolved and become a continuous flow of information. A clear example is the microblogging platform Twitter. Twitter users regularly post their comments on a particular news item, a recently purchased product or service, and even on everything that happens around them. The changes in communication patterns of Internet users introduced by Twitter began to be studied early on. In 2007, the first published research on Twitter was more closely related to the field of sociology than to computer science. Java, Song, Finin, and Tseng (2007) and Krishnamurthy, Gill, and Arlitt (2008) were among the first to perform a sociological study of Twitter. Both studies showed that the number of updates and the number of followers of each user exhibited a power-law distribution. They also highlighted the fact that much of the information flowing through Twitter is generated by a small group of users, and therefore a large number of users are mere consumers of information. The relentless growth of the number of Twitter users and the amount of information that flows through it also has gained the interest of the Natural Language Processing (NLP) community, which has begun to study the texts published in Twitter, and more specifically, those related to SA challenges. The first publication in this field was by Go, Bhayani, and Huang (2009), which is described later.

Tweets often express opinions or emotions (Diakopoulos & Shamma, 2010; Jansen, Zhang, Sobel, & Chowdury, 2009), but they have special characteristics and are different from reviews and comments published in forums, websites, or blogs. The purpose of reviews is to summarize the authors' thoughts, whereas tweets are more casual and limited to 140 characters. Furthermore, tweets have to be analyzed using phrase-level or sentence-level sentiment classification, whereas reviews are normally processed at the document level. Moreover, the text of tweets is sometimes informal and badly constructed, with poor grammar structure and misspellings. In addition, due to the length limit of 140 characters, usage of abbreviations is common. This represents a substantial challenge because the correct operation of the subsequent processes depends on the quality and correctness of the texts (Go et al., 2009).

So far, the most outstanding articles on SA in Twitter have followed a supervised approach and have focused on the question of which features are best for polarity classification. On one hand, most of them have built their own corpus following different heuristics to automatically assign the sentiment tag to each tweet (Go et al., 2009), considering those tweets with positive emoticons as positive and those with negative emoticons as negative. On the other hand, Bollen, Pepe, and Mao (2011) worked only with those tweets containing expressions such as "I'm . . ." or "I feel . . .." Although the result of those processes is a labeled corpus, the quality of the tags is not the same as the tags obtained in a manual process. These kinds of labels are known as *noise labels.* The methods presented in these articles followed this method to build a representative corpus because the manual building of a labeled corpus for SA in Twitter is complicated, time-consuming, and tedious. Therefore, one of the difficulties facing research into the task of SA in Twitter is the lack of quality-labeled corpora that can be used to address experiments. Thus, it is necessary to develop semisupervised or unsupervised methods without the precondition of a labeled corpus for a training process.

The main contribution of this article is an unsupervised and context-dependent approach for SA in Twitter. The polarity classification problem is solved by weighting SentiWordNet values with the score of a random walk analysis (PageRank) of the concepts found in the text over the WordNet graph. To validate our approach, we have used the well-known Stanford Twitter Sentiment[1] (STS)[2] corpus (Go et al., 2009). The choice of this corpus allows us to perform several experiments to analyze major issues in our method and to compare it with the rest of the approaches applied to the same corpus.

The rest of the article is organized as follows: First, we outline existing work on SA, with a focus on Twitter SA. Next, we describe the architecture of the nonsupervised approach that is proposed. Then, the large set of experiments we carried out is specified and analyzed, and characteristics of the corpus used are revealed. Subsequently, a comparison with all the methods that have been applied to the STS corpus is covered. Finally, we conclude our work and present future directions for research.

Related Work

Usually, two approaches are proposed to tackle polarity classification in SA. The first approach is based on the use of lexical resources such as lexicons (also known as the *unsupervised* or *semantic orientation approach*) (Turney, 2002), and the other is based on the use of machine learning techniques (also known as the *supervised approach*) (Pang, Lee, & Vaithyanathan, 2002). Those based on lexical resources often obtain low recall values because they depend on the presence of the words comprising the lexicon in the text being analyzed to determine the orientation of opinion. Meanwhile, methods based on machine learning depend on the availability of labeled data sets. Regarding SA in Twitter,

[1]http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip
[2]http://www.sentiment140.com/

DOI: 10.1002/asi

the first strategy has the problem of the varied and changing nature of the language used on Twitter, and the second one has the difficulty of obtaining a large corpus of labeled tweets. To solve these problems, some studies have proposed a hybrid method for the classification of polarity on Twitter. For example, Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) proposed a hybrid system for the analysis of sentence-level opinions on a corpus of English tweets. First, they performed a preprocessing to clean up the documents. Then a lexicon-based method for classifying tweets according to their polarity was applied. To solve the problem of recall, the authors attempted to identify a greater number of words indicative of subjective content. Thus, the authors applied the χ^2 test, with the idea that if a term is more likely to appear in a positive or negative judgment, it is more likely to be a subjective content identifier. They then attempted to identify a greater number of words indicative of subjective content. In this way, they could increase the number of labeled tweets. Finally, the SVM algorithm was applied as a machine learning method for classification of the tweets.

Although most studies on polarity classification on Twitter are based on the machine learning approach, until now the lack of resources on Twitter, mainly due to the novelty of the task, makes it difficult to obtain results for real scenarios. Experiments are usually run on a small data set; therefore, it is difficult to extrapolate. In this article, we apply an unsupervised approach using a semantic orientation. One of the first studies on the classification of polarity based on semantic orientation was performed by Kim, Gilbert, Edwards, and Graeff (2009). This study presented an analysis of the mood of the users who posted tweets about the death of Michael Jackson. The authors did not use a machine learning algorithm but, rather, discriminated the expressions of sadness in tweets based on the score that the lexicon Affective Norms for English Words[3] (ANEW) assigned to each term appearing in the tweets. ANEW (Bradley & Lang, 1999) provides a set of 1,034 English words, which are scored on a scale of 1 (*displeasure/calm/weakness*) to 9 (*pleasure/excitement/strength*) concerning three different moods: valence (pleasure/displeasure), arousal (excitement/calm), and dominance (strength/weakness).

Another interesting study was presented by Thelwall, Buckley, and Paltoglou (2011), who analyzed the possible relation between events and changes in the opinions expressed in social networks. They tried to determine the intensity of opinion, both negative and positive, of the tweets using an algorithm they developed. The SentiStrength[4] algorithm (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) is designed for the analysis of short messages and is trained to work with abbreviations and jargon specific to the data source. The algorithm assigns each message a score of positivity and negativity on a scale of 1 (*no sentiment*) to 5 (*very strong positive/negative sentiment*). Although SentiStrength was initially designed for the analysis of texts published on MySpace, the authors have recently extended the algorithm to manage any kind of document, including messages posted on Twitter (Thelwall, Buckley, & Paltoglou, 2012).

Hernández and Sallis (2011) also proposed an unsupervised method of reducing features for SA. Their method is based on the Latent Dirichlet Allocation (LDA) method, which is summarized in their article. The method was evaluated with a corpus of 10,000 tweets in English on the iPad tablet, which were downloaded during the months of March and April 2011. After cleaning the corpus, the tweets were represented following the vector space model and using the tf–idf metric to weight the terms. Once all the tweets were represented, the authors applied their proposal for the reduction of features. The authors did not carry out a polarity classification process to compare the performance of the complete data set and the reduced data set. They based their conclusion on the idea that the reduced model is better because its entropy value is lower than the entropy value of the complete model.

Finally, our previous study (Montejo-Ráez, Martínez-Cámara, Martín-Valdivia, & Ureña-López, 2012) presented a novel unsupervised approach that combines Wordnet, SentiWordNet, and the PageRank algorithm. The method was tested over a corpus of tweets we generated, and the results were promising. For this reason, in the current article, we wanted to test a similar, but evolved, strategy using the corpus of tweets STS because it is one of the few available Twitter corpora that includes a set of manually labeled tests.

The vast amount of information published on Twitter and the wide variety of topics on which users write make the construction and manual tagging of a corpus for classification of polarity difficult and expensive. Thus, the authors of the STS corpus use the emoticons that usually appear in tweets to differentiate between positive and negative tweets. Through Twitter Search application programming interfaces (APIs), authors generate a corpus of positive tweets, with positive emoticons ":)" and tweets with negative emoticons ":(". The validity of this technique was demonstrated by Read (2005). There are several papers using this corpus but most of them applied a machine learning approach (Bifet & Frank, 2010; Saif, He, & Alani, 2012a; Speriosu, Sudan, Upadhyay, & Baldridge, 2011). On the contrary, we have developed an unsupervised strategy based on a semantic orientation over the STS corpus. We think that polarity classification in Twitter needs unsupervised approaches due to the high variety of domains, so strategies far from domain dependency usually related to supervised approaches must be further explored. Much research is being carried out on this subject, known also as the *Sentiment Transfer problem*[5] (He, Lin, & Alani, 2011; Wu & Tan, 2011).

[3]http://csea.phhp.ufl.edu/media/anewmessage.html

[4]http://sentistrength.wlv.ac.uk/

[5]The Sentiment Transfer problem or cross-domain sentiment classification is the field that focuses on the study of techniques to classify the polarity of documents in a target domain, with training data from a source (different) domain, so the goal is to develop domain-independent SA techniques.

DOI: 10.1002/asi

A more extended description of the state of the art in SA in Twitter can be found in Martínez-Cámara, Martín-Valdivia, Ureña-López, and Montejo-Ráez (2012).

Random Walk Algorithms in SA

A random walk is a mathematical formalization of a trajectory that consists of taking successive random steps. The use of random walk algorithms in SA is based on the idea that if the process starts at a given word, it is more likely to hit another word with the same semantic orientation before hitting a word with a different polarity orientation. This idea was explored by Hassan and Radev (2010), who applied a Markov random walk model to WordNet (Fellbaum, 1998), producing a polarity estimate for any given word. Google PageRank (Page, Brin, Motwani, & Winograd, 1999) is another random walk algorithm that could be used to measure the positive or negative grade of words. Esuli and Sebastiani (2007) wanted to obtain a sentiment ranked version of WordNet, so they used a modified version of the original formula of PageRank to bias the random walk to positive and negative terms. They ran the algorithm twice, once for positive terms and again for negative terms, resulting in the ranking of WordNet synsets[6] in terms of how strongly they possess a given semantic property, but the scores assigned per each synset are context-independent. In SA, the context is important because the polarity could change; for example, the adjective "funny" may have a positive semantic orientation when used in the review of a comedy film, but may have a negative polarity when used in an action movie. Wijaya and Bressan (2008) proposed a context-dependent ranking procedure that can rank items directly from the opinions in their reviews using PageRank as a ranking algorithm.

SentiWordNet

SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010) is a lexical resource based on the well-known WordNet, which provides additional information on synsets related to sentiment orientation. SentiWordNet returns from every synset a set of three scores representing the level of positivity, negativity, and objectivity (computed from the two previous scores). This resource has been used by the SA community because it provides a domain-independent resource for obtaining certain information about the degree of sentiment charge of its concepts (Denecke, 2008; Ogawa, Ma, & Yoshikawa, 2011; Perea-Ortega, Martín-Valdivia, Ureña-López, & Martínez-Cámara, in press).

SentiWordNet also has been used in Twitter SA. Chamlertwat, Bhattarakosol, Rungkasiri, and Haruechaiyasak (2012) described a system (Micro-blog Sentiment Analysis System) for discovering consumer opinions. First, the system performs a subjective classification to distinguish between subjective and objective tweets. The second step is to measure the polarity of the tweets, which is obtained by an aggregation process of the polarity values of each term of the tweet. To obtain the semantic value, the authors use SentiWordNet; if the tweet has more positive words, the tweet is considered positive, and if it has more negative words, the tweet is considered negative.

[6]Sets of cognitive synonyms that are the minimum unit of information of the English lexical database WordNet.

System Architecture

Our system proposal consists of different components and technologies. Two of the main features of the system architecture are its modular nature and the fact that it allows use of the most suitable tool or linguistic resource in each module. We can summarize the processing of each tweet to obtain a final polarity classification in four basic steps:

1. Clean the tweet of nontextual information.
2. Disambiguate the terms in the tweet to obtain the senses.
3. Use these senses to perform a random walk over WordNet graph to expand the original list of synsets with additional ones.
4. Combine the rank of the synsets and their polarities from SentiWordNet to compute a final polarity for the tweet.

The process outlined is shown in Figure 1. In this diagram, further steps are introduced along with the other tools and libraries.

A detailed explanation of all the elements follows, with the sample tweet "@cwong08 I have a Kindle2 (& Sony PRS-500). Physical device feels good. Font is nice. Pg turns are snappy enuf. UI a little klunky.":

1. *Cleaning:* Due to the nature of the analysis that is applied to the text, removing the noninformation that usually contains the tweet is recommended. The cleaning process has involved the following:
   (a) Remove the URLs that appear in the tweets.
   (b) Although tweets rarely contain system paths, we also remove these expressions.
   (c) Remove all the emoticons.
   (d) Remove HTML tags and HTML entities.
   (e) Remove mentions. A mention in Twitter is the reference of a user to another user. Mentions have the pattern @username.
   (f) Remove any expressions of laughter. An example of laughter expression is "hahaha."
   (g) Remove nonalphanumeric characters. The punctuation symbols are not erased because they are necessary for subsequent analysis.

The result for the sample tweet is:

I have a Kindle2 Sony PRS-500. Physical device feels good. Font is nice. Pg turns are sappy enuf. UI a little klunky.

2. *Tokenization:* To process the text in the tweet, sentence splitting and word tokenization have to be performed. We have two different implementations of this process: using

DOI: 10.1002/asi

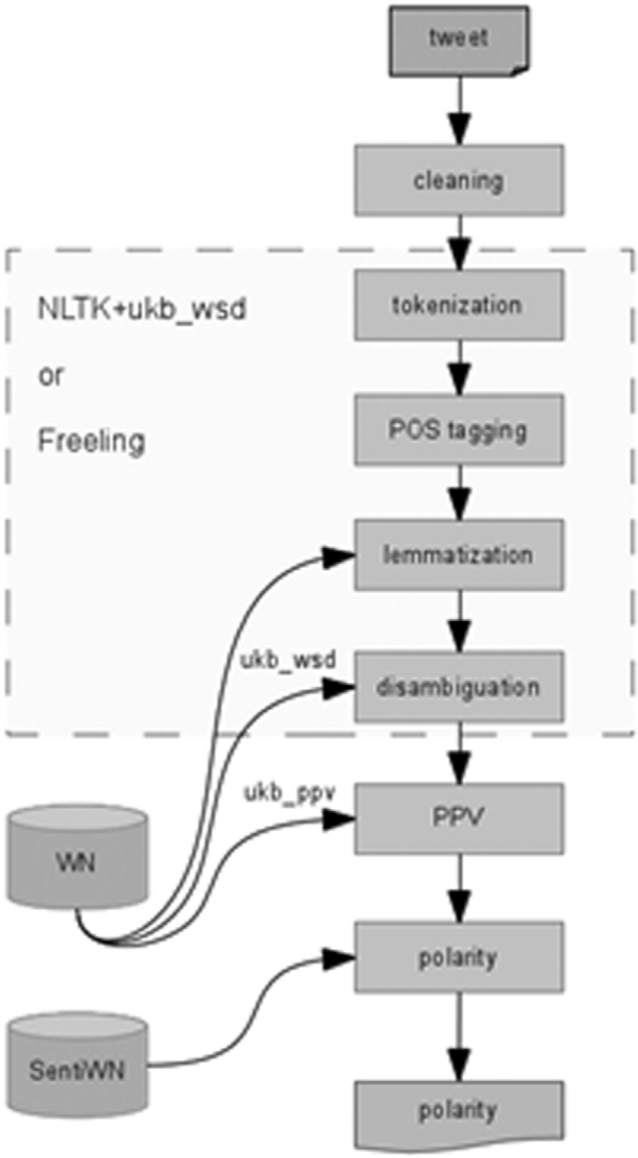

FIG. 1. Architecture of the polarity classification system.

FreeLing or the Natural Language Toolkit (NLTK) Python library (Loper & Bird, 2002). FreeLing is an open-source language-analysis toolkit[7] (Padró & Stanilovsky, 2012) that is freely available. This tool is used for tokenization, part-of-speech (POS) tagging, lemmatization, and disambiguation. The implementation with the NLTK[8] (a Python library for natural language processing) uses this library for the same purpose as does FreeLing, but the disambiguation process is performed directly with the UKB software (described later). This is the reason those steps within the box of dashes are labeled with these two implementation possibilities. During this step, stop words are removed. Terms not present in WordNet also are removed from the tweet because any possible synset can be associated. Thus, the sample tweet would become (sentences in square brackets, words in parentheses) the following:

[(Physical) (device) (feels) (good)] [(Font) (nice)][(turns) (snappy)]

3. *POS tagging:* Both lemmatization and the disambiguation need proper labeling of the terms within four possible categories: verb, adverb, noun, or adjective. Thus, a POS tagging process is performed before continuing with the next steps. The following would be the result for the sample tweet:

[(physical, ADJ) (device, NN) (feels, VRB) (good, ADJ)] [(font, NN) (nice, ADJ)] [(turns, NN) (snappy, ADJ)]

4. *Lemmatization:* After this step, terms are converted to a canonical form. This implies singular forms for plural ones in nouns, infinitive forms for verbs, and so on. The *lemma* is needed to search for associated possible senses in the following step. The result with the sample would be the following:

[(physical, ADJ) (device, NN) (feel, VRB) (good, ADJ)] [(font, NN) (nice, ADJ)] [(turn, NN) (snappy, ADJ)]

5. *Disambiguation:* Here, a WSD is performed, assigning one unique synset from WordNet to each term. WordNet[9] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations (Miller, 1995). This graph nature of the WordNet structure is the basis for the UKB disambiguation method proposed by Agirre and Soroa (2009). The UKB disambiguation algorithm performs a random walk process over the graph starting from term nodes, where each term node points to all its possible senses (i.e., synsets). After the iteration process finishes, a ranked list of synsets is obtained, choosing those with the highest rank as candidate sense. The weight of the final selected synset is used in the next step and therefore stored. Again, two possible implementations have been coded: one using the integrated UKB disambiguation of FreeLing and another where the *ukb_wsd* program is called from Python code. This program is included in the UKB software package.[10] Table 1 shows a sample output from this step.
6. *Personalized PageRank vectors (PPVs):* Once the synsets for the tweet are computed, this step performs a second run of the random walk process described by Agirre and Soroa (2009). Using the *Personalized Page Rank*, a vector of ranked synsets (PPVs) is obtained. This vector is a list of synsets with their ranked values. The key to our approach resides in taking from this vector additional synsets not related directly to the synsets that represent the senses of the terms in the tweet. In this way, the

[7]Available under GPL license from http://nlp.lsi.upc.edu/freeling/
[8]Available from http://nltk.org/

[9]See more information at http://wordnet.princeton.edu/. We used the 3.0 release.
[10]Available under the GPL licenses from http://ixa2.si.ehu.es/ukb/. We used the 2.0 version.

DOI: 10.1002/asi

TABLE 1. UKB disambiguation results.

Lemma	Synset	Disambiguation rank
Sentence 1		
physical	01778212-a	0.002469
device	03183080-n	0.003064
feel	01771535-v	0.001470
good	01586752-a	0.000944
Sentence 2		
font	06825399-n	0.007905
nice	01586342-a	0.003513
Sentence 3		
turns	00350030-n	0.001661
snappy	00971933-a	0.003695

TABLE 2. Synset expansion.

Synset	Score	Terms related
01586752-a	0.000944	good#6
01807219-a	0.007099	pleasing#1
00089051-a	0.006863	agreeable#1
01586342-a	0.006291	nice#1
05216365-n	0.005195	physical_structure#1

tweet is *expanded* with additional, but semantically related, concepts. The result is a longer list of pairs <*synset, weight*> where the weight is the rank value obtained by the propagation of weights of original synsets (those corresponding to the sense of the terms actually in the tweet) across the WordNet graph. Our first sample sentence could obtain, with this expansion, the five additional synsets shown in Table 2 with their corresponding scores, the result of the random walk iteration.

7. *Polarity:* The main resource used here is SentiWordNet,[11] a linguistic resource that maps every synset to polarity negativity and positivity scores (Baccianella et al., 2010). According to these values, in this last step, the final polarity value is calculated according to the following equation:

$$p(t)=\frac{1}{|t|}\sum_{s\in t}\frac{1}{|s|}\sum_{i\in s}\left(p_i^{+}-p_i^{-}\right)w_i \tag{1}$$

where

- $p(t)$ is the polarity of tweet t, t being a set of sentences.
- $|t|$ is the number of sentences in tweet t.
- s is a sentence in t, being itself a set of synsets.
- $|s|$ is the number of synsets in sentence s.
- i is a synset in s.
- p_i^{+} is the positive polarity of synset i.
- p_i^{-} is the negative polarity of synset i.
- w_i is the weight of synset i.

The tweet is positive whether the polarity value [$p(t)$] is greater than zero, and negative if the polarity value is less than zero.

[11]Available at http://sentiwordnet.isti.cnr.it/

The first sentence of our sample would obtain polarity as follows:

Pol(01778212-a) = (0.000000 – 0.000000) * 0.002469	*physical#1*
Pol(03183080-n) = (0.000000 – 0.000000) * 0.003064	*device#1*
Pol(01771535-v) = (0.125000 – 0.250000) * 0.001470	*feel#1, experience#4*
Pol(01586752-a) = (1.000000 – 0.000000) * 0.000944	*good#6*
Pol(01807219-a) = (0.500000 – 0.125000) * 0.007099	*pleasing#1*

Thus, the final polarity of the sentence is equal to 0.0123585, resulting in a positive sentence. This is averaged with the polarity of the remaining sentences to score the polarity of the whole tweet.

There are certain variations and configuration parameters throughout this process, such as the way we clean the original tweet text, the number of synsets added as expansion, or which WordNet relations are taken into account as edges in the graph for disambiguation and expansion. Because of this, we have performed several experiments to explore how they influence the final performance of the system.

Experiments and Results

Our goal is to explore the validity of an unsupervised polarity classification approach by means of concept expansion with a personalized PageRank algorithm. We have indicated in the description of each module that different tools (NLTK and FreeLing) can be used, but the experiments (discussed later) have been carried out using the NLTK library. As described earlier, the two main choices have to be determined: how the tweet is preprocessed (cleaned) and how the graph is constructed.

To process the tweet, we analyzed it considering several sentences or all sentences together. This is due to the fact that tweets are short pieces of text; therefore, the consideration of several sentences may affect the final polarity calculation because averaging sentences polarities may introduce "noise" in tweets with short sentences.

For graph construction, we have tried all the standard WordNet relations (hyperonimy, hyponomy, synonymy, meronymy, holonymy, antonymy . . .), without antonymy, or adding glosses relations (when related terms appear in other synset glosses).

Following these considerations, and to better understand other parameters (e.g., number of new terms to add in the expansion), three banks of experiments were carried out:

- *Baseline:* WSD using most frequent sense (MF approach; i.e., the first sense in WordNet), without weighting SentiWordNet scores and without performing expansion with additional synsets.
- *Weighted WSD:* WSD with UKB algorithm and weighting SentiWordNet polarity scores with the weights resulting from the random walk process in the previous disambiguation step.
- *Applying expansion:* The same as before, but with a second random walk process to collect additional synsets

DOI: 10.1002/asi

used in the final computation of the polarity score, as described in the previous section.

With this set of experiments, we were able to explore the contribution of each step to the final performance of the unsupervised polarity classification system. For the two last banks, tweet preprocessing was studied with the same aim: to find out the effects of possible configurations of the system. For the last one, graph construction choices were studied. In the following sections, we provide further details of the method used in these experiments, along with the results obtained.

STS Corpus

For our experiments, we used the STS corpus presented in Go et al. (2009). This study is one of the first about SA in Twitter. These authors carried out a supervised approach to solve the polarity classification problem. The supervised approached is characterized by requiring a large set of training data. In SA, this request is more important because the machine learning algorithms need more data to discover the differences between positive and negative examples. Thus, Go et al. (2009) built a training corpus for polarity classification of English tweets. Due to the short text size (140 characters maximum), a large amount of labeled tweets was needed. Manually tagging a large set of tweets is a very complicated and time-consuming task. Therefore, Go et al. chose to label the tweets with emoticons. They considered as positive those tweets with positive emoticons (☺, ☺, :), :D, or =)), and as negative those with negative emoticons (☹, ☹ : ( ). The result was a training set of 1.6 million tweets with the same number of positive and negative tweets, which were gathered between April 6, 2009, and June 25, 2009.

In addition to the training data set, the supervised approach requires a test set to prove the quality of the method, so Go et al. (2009) also built a test corpus. The test data set may be much smaller than the training one, so in this case, it is possible to manually generate a label test set. Therefore, they built a test corpus of 177 negative tweets and 182 positive tweets which were manually marked. They searched the Twitter Search API with queries related to consumer products, companies, and people. To evaluate our unsupervised method, we only selected the test set of the STS corpus because our method does not require any training set, which is one advantage of the nonsupervised approaches.

Baseline

As baseline results, we computed the polarity score by means of SentiWordNet positivity and negativity values. To this end, sense disambiguation has to be performed in some way, as SentiWordNet considers synsets to polarity score, not terms. Therefore, for the baseline, we applied MF disambiguation; that is, we considered the most common sense

TABLE 3. Results of the baseline.

Accuracy	Precision	Recall	F-score
0.423398	0.667593	0.423325	0.518112

(first sense in WordNet) as the sense of the terms in the tweet. In this way, the computation is rather simple:

$$p(t)=\frac{1}{|t|}\sum_{s\in t}\frac{1}{|s|}\sum_{i\in s}\left(p_i^+-p_i^-\right) \quad (2)$$

When compared with Equation 1, in this case we have no weighting in the polarity difference, and the synsets are determined by MF. The results obtained are shown in Table 3.

As can be seen, these results are discouraging. SentiWordNet seems a useless resource for unsupervised classification in Twitter. The bank of results shows how this performance can be improved by means of polarity score weighting.

Applying WSD

The first set of experiments uses Equation 1 in carrying out a WSD process. In this case, no expansion is performed, but synsets polarity scores are weighted with the disambiguation value obtained by means of the UKB algorithm. Two possibilities are explored as noted earlier: to consider different sentences or to join all of them in one unique sentence. This affects only the final calculation; with joined sentences, there is no average of polarities over sentences, so the equation could be simplified as follows:

$$p(t)=\frac{1}{|t|}\sum_{i\in t}\left(p_i^+-p_i^-\right)w_i \quad (3)$$

where

- $p(t)$ is the polarity of tweet t, t being a list of synsets.
- $|t|$ is the number of subjective synsets in tweet t.
- s is a sentence in t, being itself a set of synsets.
- $|s|$ is the number of synsets in sentence s.
- i is a synset in t.
- p_i^+ is the positive polarity of synset i.
- p_i^- is the negative polarity of synset i.
- w_i is the weight of synset i.

The results obtained are shown in Table 4.

As can be seen, the tokenization of sentences has a negative effect on the final performance of the system due to the bias introduced by short expressions as in the sample tweet shown earlier. A short sentence with a positive term receives the same relevance as a long one with several negative terms. With this finding, we realized that in the case of Twitter it is preferable not to consider tokenization at sentence level, so the computation of the polarity would follow Equation 3.

DOI: 10.1002/asi

TABLE 4. Results without expansion.

Sentences joined	Weight	Accuracy	Precision	Recall	F-score
No	1.0	0.476323	0.645814	0.475213	0.547532
No	PPV	0.490251	0.644361	0.490035	0.556701
Yes	1.0	0.529248	0.720177	0.528497	0.609625
Yes	PPV	0.545961	0.695121	0.546145	0.611693

PPV = Personalized PageRank vectors.

TABLE 5. Results with expansion.

Configuration	Accuracy	Precision	Recall	F-score
WN_e64	0.586111	0.586783	0.630183	0.630183
WN-A_e60	0.586111	0.687892	0.587524	0.633759
WN+G_e50	0.583333	0.686919	0.584699	0.631701
WN-A+G_e64	0.649025	0.650431	0.673904	0.673904

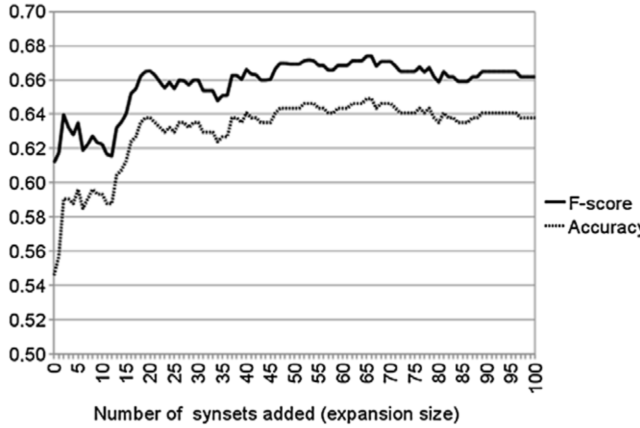

FIG. 2. Performance variation according to expansion size for WN−A+G graph configuration.

Important evidence from these results is the fact that the weighting of SentiWordNet scores (PPV values) improves the results shown in the previous section. Therefore, disambiguation is interesting, but also a proper weighting of SentiWordNet polarities. As we can see, when the sense is determined by the UKB algorithm but the weight of the synset does not affect the polarity in SentiWordNet (i.e., w_i is set to 1.0 always), the results are not as good as when considering the random walk final value for weighting the synset polarity.

Applying Expansion

Due to the improvement obtained by applying the ranking scores from the random walk approach when disambiguating terms, this set of experiments proves that this algorithm can be used to increase the number of concepts (synsets) in the tweet and use them for proposing a more accurate polarity score. Equation 3 is again used here, but the set of concepts (t in the equation; i.e., the list of synsets) is enlarged with those synsets resulting from the propagation of weights from original synsets due to a random walk (a PageRank-like method) over the WordNet graph. Again, several variants have been explored, and Table 5 presents the results obtained with some of these different configurations. For each possible variation of the graph, only the expansion with the highest performance is shown here. These configurations are:

- *WN_e64:* Graph based on all WordNet relations. Expansion with 64 additional synsets.
- *WN–A_e60:* Graph based on all WordNet relations, but without the antonymy one. Expansion with 60 additional synsets.
- *WN+G_e50:* Graph based on all WordNet relations plus new relations built from the text of the WordNet glosses. Expansion with additional 50 synsets.
- *WN–A+G_e64:* Graph based on all WordNet relations plus glosses relations and without antonymy connections. Expansion with additional 64 synsets.

From Table 5, we can conclude that discarding antonyms, considering the new relations built from the text of the WordNet glosses, and expanding with 64 concepts allow us to reach an F-score of 0.673904, which is an impressive improvement of 23.12% over the other baseline for an unsupervised approach.

Of course, the effect of the number of additional synsets considered was analyzed. Runs from 0 to 100 additional terms were launched. We can see this effect graphically in Figure 2. As can be easily identified, in general the number of additional synsets increases the performance (here, in terms of F-score and accuracy) of the polarity classification system, up to a maximum value. This graph corresponds to the configuration WN−A+G and, as given in Table 5; a maximum value of performance is reached when expanding with 64 concepts. To visually identify the convenience of certain graph relations, Figure 3 is presented. From this graph, we can see that WN−A+G outperforms other configurations.

Further Experiments on POS Tag Selection

The use of POS tags as features for polarity classification in Twitter has been studied by several researchers. Go et al. (2009) analyzed the contribution of POS tags in combination with unigrams and bigrams, but the experiments with those lexical features did not outperform those that only used unigrams and bigrams. Nevertheless, Pak and Paroubek (2010) and Agarwal, Xie, Vovsha, Rambow, and Passonneau (2011) argued the importance of POS tags with or without word-prior polarity. In Saif, He, and Alani (2012b), the use of POS tags improves the unigrams' baseline results, but the best results are achieved when the tweets are represented by semantic features and not by POS-tags

DOI: 10.1002/asi

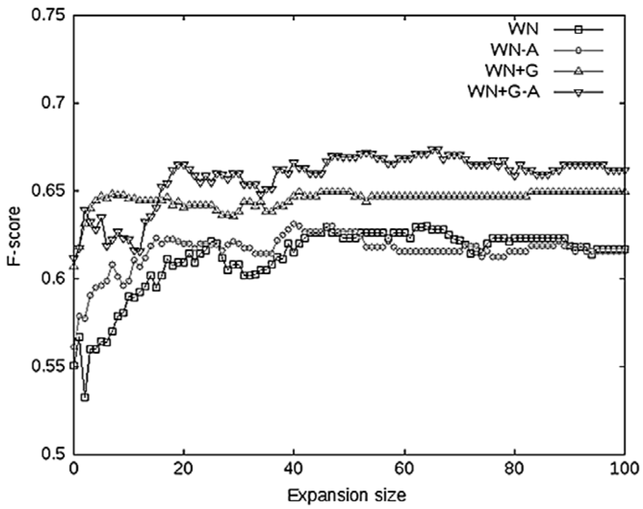

FIG. 3. Effect of graph relations on F-score.

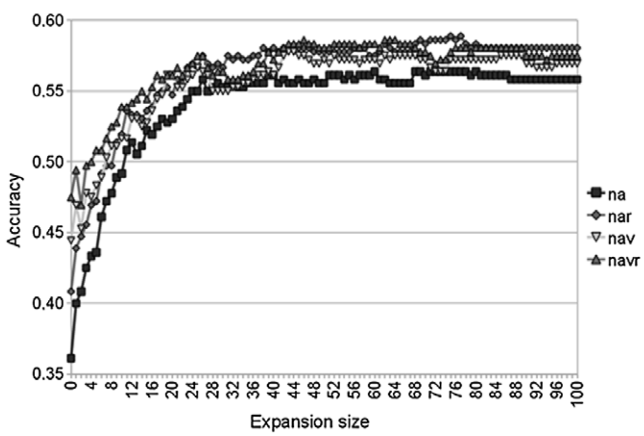

FIG. 4. Effect of part-of-speech-based term-selection on accuracy.

ones. Regarding the aforementioned articles, there are some differences in the research community regarding the contribution of POS tags, so we performed some experiments on this issue and found, as shown in Figure 4, that POS selection was worthless (despite the tiny improvement when verbs were not considered). In Figure 4, "n" represents nouns, "v" verbs, "a" adjectives, and "r" adverbs. As can be seen by the density of the lines, no significant improvement can be identified. These experiments were performed on the whole WordNet graph, with sentence splitting, so the accuracy values are not as high as for the best configuration (no sentence splitting, WordNet graph with glosses relations and without antonymy connections).

Results Comparison

The main reason for choosing the STS corpus was that it has been used in other studies, so it allows us to compare our unsupervised method with those applied to the same corpus. Go et al. (2009) presented the STS corpus and also performed several experiments in which they explored augmenting different n-gram features in conjunction with POS tags in the training of three supervised classifiers: support vector machine (SVM), naïve Bayes (NB), and maximum entropy (MaxEnt). The best results were achieved with the algorithm maximum entropy and taking as features a combination of unigrams and bigrams.

Although Go et al. (2009) studied the way of representing tweets and the most suitable machine leaning algorithm, Bifet and Frank (2010) focused their study on the analysis of the performance of three data-flow algorithms. Before they applied the algorithms, they removed the stop words from the tweets, represented them as a set of unigrams, and utilized term presence (boolean weight) to weight the unigrams. The machine learning methods tested were multinominal naïve Bayes, stochastic gradient descent, and Hoeffding tree (Domingos & Hulten, 2000). When the STS test corpus was used, the algorithm that achieved the best results was the multinominal naïve Bayes, with 0.8245 of accuracy.

Speriosu et al. (2011) compared several methods for polarity classification in Twitter in three corpora, one of them the STS corpus. The method that achieved the best results is based on the combination of several knowledge sources with a noise-supervised label propagation algorithm. The method consists of a graph that has users, tweets, unigrams, bigrams, hashtags, and emoticons as its nodes, which are connected based on the link existence among them. A label propagation method is then applied, where the polarity labels are propagated from a small set of seed nodes whose real labels are known throughout the graph.

Saif et al. (2012a) also followed a supervised approach to tackle the problem of SA in Twitter. Their main contribution is two sets of features to alleviate the data-sparsity problem in Twitter: semantic features and sentiment-topic features. The semantic features are obtained through an entity-clustering process where name entities are clustered in high-level semantic concepts. To discover the sentiment-topic features, these authors used the joint sentiment-topic model (Lin & He, 2009). When they measured the quality of the algorithm selected (naïve Bayes), the model that performed best was the one based on interpolation semantic features; however, when they used the STS test corpus, the sentiment-topic model performed better.

To complete the comparison, the results of the aforementioned studies and our best one are shown in Table 6. In this table, the approach followed (supervised/unsupervised), the algorithm, and the accuracy values are highlighted.

Differences between our method and the other ones are not small. However, note that our approach is nonsupervised, so it does not need any previous training process and only a minimum base of knowledge to classify the polarity of tweets. This is paramount because it solves one of the problems with SA in Twitter: the lack of a representative manually labeled corpus of tweets for building machine

DOI: 10.1002/asi

TABLE 6. Results reached by other systems and our best run over the STS corpus.

Article	Approach	Method	Features	Accuracy
Go et al. (2009)	Supervised	Maximum entropy	Unigram + bigram	0.8300
Bifet & Frank (2010)	Supervised	Multinominal naïve Bayes	Unigrams	0.8245
Speriosu et al. (2011)	Supervised	Label propagation	Unigrams, bigrams, and microblog features	0.8470
Saif et al. (2012a)	Supervised	Naïve Bayes	Semantic	0.8410
			Sentiment-topic	0.8630
Our method	Unsupervised	SentiSynset expansion	Unigrams	0.6490

learning models. Another advantage of the unsupervised model is its domain-independent nature, so we can easily apply and adapt it to any domain.

Conclusions and Future Work

Until now, most articles published about polarity classification in Twitter have followed a supervised approach, and the main differences are in the use of lexical, syntactic, or microblogging features, and most recently, the use of some semantic features. In this article, we present an unsupervised approach to tackling the problem of polarity classification based on the combined use of two knowledge resources: WordeNet and SentiWordNet. Another characteristic of the method presented is its context-dependent nature, accomplished by the application of the PageRank algorithm to WordNet, with the goal of obtaining a rank set of synsets related to the concepts and the context of the tweet. The result of this process is the expansion of the terms that appear in the tweet, so it is more likely to capture the exact meaning that the user expresses in the post.

Thus, after this intensive experimentation, our results reveal the following three main facts:

- The weighting of the polarities' scores from SentiWordNet improves the performance of the proposed unsupervised method.
- Expanding with additional terms clearly influences the results, with a significant gain in measurement values.
- The selection of the graph is relevant, and this has to be chosen carefully because all remaining graph-based computations rely on it.

The aim of this research line is to continue improving the accuracy of the method, so we have planned several experiments on the different parts of the algorithm. The text of tweets is usually informal and badly constructed, with poor grammar structure and misspellings; therefore, before applying our method, we should solve these problems. Furthermore, an important problem of tweets is the data sparsity, which also should be alleviated. We are working to improve the quality of our cleaning process because it could be influential in the next step, the disambiguation process. We also are studying different disambiguation methods to obtain the most accurate WordNet synsent for each term. The polarity measure method is another part of the algorithm that we are considering, with the goal of calculating the most accurate polarity value.

Another challenge of SA is the multilingual domain, so we are considering the adaptation of our method in a multilingual environment. With that goal in mind, we are considering several multilingual frameworks and parallel corpora.

Acknowledgments

This work was partially supported by a grant from the Fondo Europeo de Desarrollo Regional (FEDER), TEXT-COOL 2.0 Project (TIN2009-13391-C04-02), and ATTOS Project (TIN2012-38536-C03-0) from the Spanish government. In addition, this article was partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development through the FIRST Project (FP7-287607). This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use that may be made of the information contained therein.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the Asssociation for Computational Linguistics (ACL 2011) Workshop on Languages in Social Media (pp. 30–38).

Agirre, E., & Edmonds, P. (2006). Introduction. In E. Agirre & P. Edmonds (Eds.), Word sense disambiguation (pp. 1–28). the Netherlands: Springer.

Agirre, E., & Soroa, A. (2009). Personalizing PageRank for word sense disambiguation. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), Athens, Greece (pp. 33–41). Stroudsburg, PA: Association for Computational Linguistics.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC10 0) (pp. 2200–2204).

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), Honolulu, HI (pp. 127–135). Stroudsburg, PA: Association for Computational Linguistics.

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in Twitter streaming data. Proceedings of the 13th International Conference on Discovery Science (DS'10) (pp. 1–15). Berlin, Germany: Springer-Verlag.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic (pp. 440–447). Stroudsburg, PA: Association for Computational Linguistics.

Bradley, M.M., & Lang, P.J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings (Tech. Report). University of Florida, Center for Research in Psychophysiology.

Bollen, J., Pepe, A., & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In L. Adamic, R.A. Baeza-Yates, & S. Counts (Eds.), Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Spain (pp. 450–453).

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., & Haruechaiyasak, C. (2012). Discovering consumer insight from Twitter via sentiment analysis. Journal of Universal Computer Science, 18(8), 973–992.

Denecke, K. (2008). Using SentiWordnet for multilingual sentiment analysis. Data Engineering Workshop (ICDEW 2008), Institute of Electrical and Electronic Engineers (IEEE) 24th International Conference on Data Engineering (pp. 507–512). New York, NY: Institute of Electrical and Electronic Engineers.

Diakopoulos, N.A., & Shamma, D.A. (2010). Characterizing debate performance via aggregated twitter sentiment. In Proceedings of the Special Interest Group on Computer–Human Interaction (SIGCHI) Conference on Human Factors in Computing Systems (CHI ’10), Atlanta, GA (pp. 1195–1198). New York, NY: ACM.

Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. Proceedings of the Sixth Association for Comuputing Machinery Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD) International Conference on Knowledge Discovery and Data Mining, Boston, MA (pp. 71–80). New York, NY: ACM.

Esuli, A., & Sebastiani, F. (2007). PageRanking WordNet synsets: An application to opinion mining. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic (pp. 424–431).

Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision (CS224N–Final Project Report). Stanford University, Stanford, CA.

Hassan, A., & Radev, D. (2010). Identifying text polarity using random walks. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden (pp. 395–403). Stroudsburg, PA: Association for Computational Linguistics.

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, Portland, OR (pp. 123–131).

Hernández, S., & Sallis, P. (2011). Sentiment-preserving reduction for social media analysis. In C. San Martin & S.-W. Kim (Eds.), Lecture Notes in computer science: Vol. 7042. Progress in pattern recognition, image analysis, computer vision, and applications (pp. 409–416). Berlin, Germany: Springer.

Jansen, B.J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Microblogging as online word of mouth branding. Proceedings of the Computer–Human Interaction Extended Abstracts on Human Factors in Computing Systems (CHI EA ’09), Boston, MA (pp. 3859–3864). New York, NY: ACM.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. Proceedings of the Ninth WebKDD and First SNAKDD 2007 Worksop on Web Mining and Social Network Analysis (pp. 56–65). New York, NY: ACM.

Kim, E., Gilbert, S., Edwards, M.J., & Graeff, E. (2009). Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. Web Ecology Project (Publ. No. 03, August). Retrieved from http://webecologyproject.org

Krishnamurthy, B., Gill, P., & Arlitt, M. (2008) A few chirps about Twitter. Proceedings of the First Workshop on Online Social Networks (pp. 19–24). New York, NY: ACM.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Proceddings of the Fifth Annual International Conference on Systems Documentation (SIGDOC ’86) (pp. 24–26). New York, NY: ACM.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. Proceedings of the 18th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management, Hong Kong, China (pp. 375–384). New York, NY: ACM.

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167.

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. Proceedings of the Association for Computational Linguistics (ACL-02) Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Vol. 1 (ETMTNLP ’02) (pp. 63–70). Stroudsburg, PA: Association for Computational Linguistics.

Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A., & Montejo-Ráez, A. (2012). Sentiment analysis in Twitter. Natural Language Engineering. doi:10.1017/S1351324912000332

Mihalcea, R. (2005, October). Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 411–418).

Miller, G.A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39–41.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., & Ureña-López, L.A. (2012). Random walk weighting over SentiWordNet for sentiment polarity detection on Twitter. Proceedings of the Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2012) (pp. 3–10). Stroudsburg, PA: Association for Computational Linguistics.

Ogawa, T., Ma, Q., & Yoshikawa, M. (2011). News bias analysis based on stakeholder mining. IEICE Transactions on Information and Systems, 94(3), 578–586.

Padró, L., & Stanilovsky, E. (2012). FreeLing 3.0: Towards wider multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012), Istanbul, Turkey.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Standford InfoLab.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10), Valletta, Malta (pp. 19–21). Paris, France: European Language Resources Association.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL ’04). Barcelona, Spain. Stroudsburg, PA: Association for Computational Linguistics.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 79–86).

Perea-Ortega, J.M., Martín-Valdivia, M.T., Ureña-López, L.A., & Martínez-Cámara, E. (In press). Improving polarity classification of bilingual parallel corpora combining machine learning and semantic orientation approaches. Journal of the American Society for Information Science and Technology.

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal of Informetrics, 3(2), 143–157.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007, June). Semeval-2007 task-17: English lexical sample, srl and all words. Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007) (pp. 87–92).

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proceedings of the Association for Computational Linguistics Student Research Workshop (ACLstudent '05) (pp. 43–48). Stroudsburg, PA: Association for Computational Linguistics.

Saif, H., He, Y., & Alani, H. (2012a). Alleviating data sparsity for Twitter sentiment analysis. Proceedings of Making Sense of Microposts (MSM2012) (pp. 2–9).

Saif, H., He, Y., & Alani, H. (2012b). Semantic sentiment analysis of Twitter. Lecture Notes in Computer Science. The Semantic Web—International Seamntic Web Conference (ISWC 2012) Vol. 7649 (pp. 508–524). Berlin, Germany: Springer.

Sinha, R., & Mihalcea, R. (2007). Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. Proceedings of the International Conference on Semantic Computing (ICSC 2007) (pp. 363–369). New York, NY: Institute of Electrical and Electronic Engineers.

Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. Proceedings of the First Workshop on Unsupervised Learning in Natural Languagge Processing, Edinburgh, Scotland (pp. 53–63).

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 62(2), 406–418.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. Journal of the American Society for Information Science and Technology, 63(1), 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478–514.

Turney, P.D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of Association for Computational linguistics (ACL '02) (pp. 417–424). Shroudsburg, PA: Association for Computational Linguistics.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. Proceedings of the Sixth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'05) Mexico City, Mexico (pp. 486–497). Berlin, Germany: Springer-Verlag.

Wijaya, D., & Bressan, S. (2008). A random walk on the red carpet: Rating movies with User Reviews and PageRank. Proceedings of the 17th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management, Napa Valley, CA (pp. 951–960).

Wu, Q., & Tan, S. (2011). A two-stage framework for cross-domain sentiment classification. Expert Systems with Applications, 38(11), 14269–14275.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining lexicon based and learning-based methods for twitter sentiment analysis (Tech. Report HPL-2011-89) Retrieved from http://www.hpl.hp.com/techreports/2011/HPL-2011-89.pdf.

Zirn, C., Niepert, M., Stuckenschmidt, H., & Strube, M. (2011). Fine-grained sentiment analysis with structural features. Proceedings of the Fifth International Joint Conference on Natural Language Processing (iJcnLP-2011) (pp. 336–344). Chiang Mai, Thailand: Asian Federation of Natural Language Processing.