Like It or Not: A Survey of Twitter Sentiment Analysis Methods

Article *in* ACM Computing Surveys · June 2016

DOI: 10.1145/2938640

CITATIONS
201

READS
8,106

2 authors:

Anastasia Giachanou
University of Lugano
33 PUBLICATIONS **368** CITATIONS

SEE PROFILE

Fabio Crestani
University of Lugano
402 PUBLICATIONS **4,665** CITATIONS

SEE PROFILE

Like It or Not: A Survey of Twitter Sentiment Analysis Methods

ANASTASIA GIACHANOU and FABIO CRESTANI, Università della Svizzera Italiana

Sentiment analysis in Twitter is a field that has recently attracted research interest. Twitter is one of the most popular microblog platforms on which users can publish their thoughts and opinions. Sentiment analysis in Twitter tackles the problem of analyzing the tweets in terms of the opinion they express. This survey provides an overview of the topic by investigating and briefly describing the algorithms that have been proposed for sentiment analysis in Twitter. The presented studies are categorized according to the approach they follow. In addition, we discuss fields related to sentiment analysis in Twitter including Twitter opinion retrieval, tracking sentiments over time, irony detection, emotion detection, and tweet sentiment quantification, tasks that have recently attracted increasing attention. Resources that have been used in the Twitter sentiment analysis literature are also briefly presented. The main contributions of this survey include the presentation of the proposed approaches for sentiment analysis in Twitter, their categorization according to the technique they use, and the discussion of recent research trends of the topic and its related fields.

CCS Concepts: • **Information systems → Sentiment analysis;**

Additional Key Words and Phrases: Sentiment analysis, opinion mining, microblogs, twitter

ACM Reference Format:
Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of Twitter sentiment analysis methods. ACM Comput. Surv. 49, 2, Article 28 (June 2016), 41 pages.
DOI: http://dx.doi.org/10.1145/2938640

1. INTRODUCTION

Recent years have witnessed the rapid growth of social media platforms (e.g., Facebook, Twitter, Google+, and several blogs) in which users can publish thoughts and opinions on any topic. The increasing popularity of social media platforms has changed the web from a static repository of information into a dynamic forum with continuously changing information. Social media platforms gave the capability to people to express and share their thoughts and opinions on the web in a very simple way. Thus, the so-called *User Generated Content* varies a lot, from simple "likes" in status updates in Facebook to long publications in blogs.

User-generated information is a good source of opinion and can be valuable for a variety of applications that require understanding the public opinion about a concept. One typical example that illustrates the importance of public opinion refers to enterprises that can capture the views of customers about their products or their competitors.[1,2]

[1]https://www.brandwatch.com/sentiment-analysis-feature/.
[2]http://www.trackur.com/.

This research was partially funded by the Swiss National Science Foundation (SNSF) under the project OpiTrack.
Authors' addresses: A. Giachanou and F. Crestani, Faculty of Informatics, Università della Svizzera italiana (USI), Via Giuseppe Buffi 13, 6900 Lugano, Switzerland; emails: {anastasia.giachanou, fabio.crestani}@usi.ch.

DOI: http://dx.doi.org/10.1145/2938640

This information can be used to improve the quality of their services or products accordingly. In addition, it is possible for a government to understand the public view regarding different social issues and to act promptly. Another example is that potential customers of a product can use the opinionated information to decide whether to buy the product or not. Finally, users can view the evolution of sentiment in various topics in which they are interested.[3,4]

Until recently, the main sources of opinionated information were friends and specialized websites. Now, consumers can consult past experiences and opinions published by other users before buying a specific product. However, mining opinions and sentiment from social media is very challenging due to the vast amount of data generated by the different sources. The opinionated information about a topic is hidden within the data and therefore it is nearly impossible for a person to look through the different sources and extract useful information. For that reason, researchers have started investigating and developing approaches that can automatically detect the text polarity and can effectively mine opinionated information even within a huge amount of data.

Opinion Mining (OM) and *Sentiment Analysis* (SA) are two emerging fields that aim to help users find opinionated information and detect the sentiment polarity. OM and SA are commonly used interchangeably to express the same meaning. However, some researchers state that they aim to tackle two slightly different problems. According to Tsytsarau and Palpanas [2012], OM is about determining whether a piece of text contains opinion, a problem that is also known as *subjectivity analysis*, whereas the focus of SA is the sentiment polarity detection by which the opinion of the examined text is assigned a positive or negative sentiment. More formally, OM or SA is the "computational study of opinions, feelings and subjectivity in text" [Pang and Lee 2008].

According to a more complete definition given by Liu [2012] "an opinion is a quintuple $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k." To illustrate the different parts of the definition, we use an example. Consider the following review posted on *10.06.2015* by the user *Helen*:

The picture quality of my new Nikon V3 camera is great.

In this example, *Nikon V3* is the entity for which the opinion is expressed, *picture quality* is the aspect of the entity, the sentiment of the opinion is *positive*, the opinion holder is the user *Helen*, and the time that the opinion is expressed is *10.06.2015*. The opinion quintuple *(Nikon_V3, picture_quality, positive, Helen, 10.06.2015)* can be generated after analyzing this example.

OM and SA have been studied on many media, including reviews, forum discussions, and blogs. Recently, researchers have started to analyze opinions and sentiments expressed in microblogs as they contain a large number of opinionated text. One of the most popular microblogs is Twitter,[5] which has managed to attract a large number of users who share opinions, thoughts, and, in general, any kind of information about any topic of their interest. The information that is posted on Twitter frequently contains opinion about products, services, celebrities, events, or anything that is of user's interest. Due to its increasing popularity, Twitter has recently attracted the interest of many researchers who analyzed Twitter data for a variety of different tasks such as making predictions [Bollen et al. 2011], detecting users' sentiment towards different

[3]http://www.opinioncrawl.com/.
[4]https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/.
[5]http://twitter.com/.

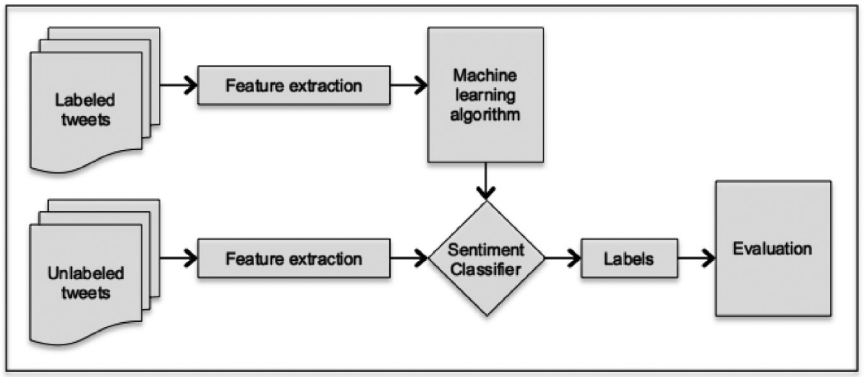

Fig. 1. Typical process for sentiment classification.

topics [Go et al. 2009], detecting users' emotions [Mohammad 2012], and detecting irony [Reyes et al. 2013].

Twitter Sentiment Analysis (TSA) tackles the problem of analyzing the messages posted on Twitter in terms of the sentiments they express. Twitter is a novel domain for SA and very challenging. One of the main challenges is the length limitation, according to which tweets can be up to 140 characters. In addition, the short length and the informal type of the medium have caused the emergence of textual informalities that are extensively encountered in Twitter. Thus, methods proposed for TSA should take into account these unique characteristics.

The majority of TSA methods use a method from the field of machine learning, known as classifier. Figure 1 shows the most typical TSA process. The first step includes collecting tweets and labeling them by sentiment. The labeled tweets represent the training data. Although Twitter API (Application Programming Interface) facilitates the process of collecting tweets, assigning labels is challenging and should be addressed carefully. More details about how tweets can be collected and annotated can be found on Section 5.2. The next step focuses on extracting a set of features that are used to train the classifier. The selected features and their combination may influence the performance of the classifier. Section 2.3 presents more details for feature selection. Both the labeled data and the selected features are forwarded to the machine-learning algorithm and are used to build the classifier model. Section 3 presents the various TSA approaches. In the last step, the classifier assigns labels to tweets that are not annotated. The correctness of those annotations determine the performance of the classifier. The evaluation metrics are described on Section 2.4.

There are numerous of articles focused on SA and, more recently, some on TSA. This creates the need for a survey article to summarize the proposed approaches and the recent research trends. Two long and detailed surveys on SA were presented some time ago by Pang and Lee [2008] and Liu and Zhang [2012]. Pang and Lee [2008] provide a comprehensive overview of the SA approaches using various types of data. However, this survey was published some time ago and does not cover recent developments and trends in the field. More recently, Liu and Zhang [2012] provided a comprehensive and detailed description of all the important concepts and topics related to SA. Their survey is organized based on the different SA applications for each of which they provide explanatory definitions and related approaches. One important difference is that our survey is specifically focused on TSA. We provide a shorter introduction to the general SA concepts and definitions that are still illustrative enough for beginners to

obtain a pedagogical overview of the field. However, we provide detailed descriptions for the most important TSA concepts, including a description of Twitter and TSA feature selection followed by a number of illustrative examples. In addition, we present the recent developments on TSA, including the use of deep learning.

Tsytsarau and Palpanas [2012] have also presented a very informative survey on SA. Similarly to Liu and Zhang [2012], they organized their survey based on the main tasks of SA for each of which they present definitions and discuss problems and the various approaches. The categorization of the articles is illustrated using tables and graphs to facilitate comparison of works. This survey briefly discusses opinion mining in microblogs, which began receiving attention only recently. Instead, in our survey, we only focus on TSA articles and categorize them by approach. Similarly to Tsytsarau and Palpanas [2012], we use tables to facilitate the comparison of the different approaches. However, we also present recent developments (i.e., deep learning) and critically discuss the TSA approaches depending on their strengths and limitations. Another difference is that we provide detailed presentation of the available TSA sentiment lexicons and datasets. In addition, we describe the process of creating own Twitter datasets using Twitter API. Moreover, there is one survey focused on research performed on Twitter presented by Martínez-Cámara et al. [2012]. However, Martínez-Cámara et al. discuss a number of different tasks applied on Twitter data and they provide only a brief overview of TSA. Also, only a small subset of the described articles are related to TSA, whereas our survey provides a comprehensive overview of this field.

To the best of our knowledge, there is a lack of detailed surveys on TSA. This survey provides a comprehensive overview of the TSA area, including recent trends and proposed approaches. First, it categorizes the presented studies based on an approach including recent developments such as the use of deep learning. This can help researchers and newcomers to obtain a panoramic view of the particular field. The presented approaches are also critically discussed based on their advantages and disadvantages. Second, it discusses fields related to TSA that have attracted research interest. These include Twitter opinion retrieval, tracking sentiments over time, irony detection, and emotion detection. Third, this survey discusses different resources that are frequently used in TSA. We summarize different approaches proposed for the construction of sentiment lexicons for Twitter, present and briefly describe the available evaluation datasets, describe the process of creating Twitter datasets, and summarize the different processes applied to annotate the existing datasets. Finally, this survey identifies and discusses some open issues and directions for future research. We believe that there are sufficient differences to make this survey distinct and complimentary to the other surveys we mentioned.

The rest of the article is organized as follows. Section 2 attempts to give a general overview of the TSA by presenting Twitter, TSA challenges, and feature selection methods. Section 3 presents and analyses the proposed TSA methods and the corresponding articles. The fields related to TSA are presented in Section 4. Research resources are presented in Section 5. Section 6 discusses the open problems and indicates future directions. Finally, the conclusions of this survey are summarized in Section 7.

2. TWITTER SENTIMENT ANALYSIS: A GENERAL VIEW

2.1. Twitter

Microblogging is a network service with which users can share messages, links to external websites, images, or videos that are visible to users subscribed to the service. Messages that are posted on microblogs are short in contrast to traditional blogs.

Currently, a number of different microblogging platforms are available, including Twitter,[6] Tumblr,[7] FourSquare,[8] Google+,[9] and LinkedIn.[10]

One of the most popular microblogs is Twitter, which was launched in 2006 and since then has attracted a large number of users. Currently, Twitter has 284 million users who post 500 million messages per day.[11] Due to the fact that it provides an easy way to access and download published posts, Twitter is considered one of the largest datasets of user generated content. Twitter is characterized by some specific features that are listed below:

—*Tweet*: A tweet is a single message posted on Twitter. The content of a tweet, which can be, at maximum, 140 characters, can vary from personal information or personal opinion on products or events to others such as links, news, photos, or videos.
—*User/Username*: A user has to be registered with the platform to post tweets. The user selects a pseudonym (username) during registration, which will be afterwards used to post messages.
—*Mention*: Mentions in a tweet indicate that the post mentions another user. To make this reference to a username, users use the symbol @ followed by the specific username they refer to (@username). Mentions are placed anywhere in the body of the tweet.
—*Replies*: Replies in a tweet are used to indicate that the post is an answer to another tweet and are usually employed to create conversations. Similarly to mentions, they are created using the @ symbol followed by the username they refer to. Replies are placed next to the username that creates the reply.
—*Follower*: Followers refer to the users that follow a user's tweets and activity. Following other users is the main way to connect to other users in Twitter. Users on Twitter receive updates from those they follow and they send their updates to those who follow them.
—*Retweet*: Retweets refer to the tweets that are re-distributed. When a user finds a tweet interesting, then he or she can re-post it by using the retweeting functionality. The retweeting is considered a powerful tool for disseminating information. The tweet that is shared remains unchanged and is usually marked with the abbreviation RT followed by the author's username (RT @username). The retweet may also contain a short comment.
—*Hashtag*: Hashtags are used to indicate the relevance of a tweet to a certain topic. Hashtags that are created using the # character followed by the topic name (#topic) have emerged from the need to label information on the messages that were posted. Tags are generated by users spontaneously and can be utilized to get all the tweets with the same hashtag. Hashtags that appear in a high number of tweets are characterized as trending topics.
—*Privacy*: Twitter gives the option to a user to decide if his/her tweets will be visible to everyone or only to his/her approved Twitter followers.

Figure 2 shows an example of a tweet. The tweet is taken from the *New York Times Fashion* user. We observe that it contains some of the mentioned characteristics such as username (*NYT Fashion*) and the indication that it is a retweet. The tweet is a reply to the user *NYTFashion* (@*NYTFashion*) and contains one mention that is referred to

[6]http://twitter.com/.
[7]http://tumblr.com/.
[8]http://foursquare.com/.
[9]http://plus.google.com.
[10]http://inkedin.com/.
[11]https://about.twitter.com/company.

Fig. 2. A picture of a tweet taken from the New York Times Fashion user.

the user *nytimes* (*@nytimes*). Also, two hashtags indicate topics related to this tweet (*#Periscope* and *#metgala*).

2.2. Sentiment Analysis Challenges

Detecting sentiment in Twitter is a non-trivial task and differs considerably from detecting sentiment in conventional text such as blogs and forums. Researchers who try to develop effective TSA methods have to confront a number of challenges that emerge from the special characteristics of Twitter. One of the most important challenges is the informal type of medium and the length limitation. Also, they have to deal with the dynamic and evolving content. Here, we present the most important TSA challenges:

—*Text Length*: One of the unique characteristics of tweets is their short length, which can be up to 140 characters. This makes TSA differ from the previous research of sentiment analysis of longer text such as blogs or movie reviews. Bermingham and Smeaton [2010] performed a study with the aim to examine if the short length of tweets makes this task more difficult compared to longer texts. To this end, they compared Support Vector Machine (SVM) and Multinomial Naïve Bayes (MNB) classifiers for blog and Twitter SA. Their results showed that SVM performs better than MNB on blogs, but MNB achieved better results than SVM on microblogs. They concluded reporting that classifying tweets is a much easier task than classifying longer documents such as blogs.

—*Topic Relevance*: Most of the work that is done on TSA aims to classify the sentiment orientation of a tweet without considering the topical relevance. To capture the topic relevance of a tweet, many researchers simply consider the presence of a word as an indicator of the topical relevance. Also, other studies consider the hashtag symbol as a strong indicator of the tweet's relevance towards a specific topic. Considering the short length of the tweets, those approaches can be partially right as in most of the cases the sentiment will target that specific topic.

—*Incorrect English*: Due to its informal type of communication and the length limitation, the language used in Twitter is very different from the language used in other text genres (web, blogs, news etc.). Tweets contain textual peculiarities including emphatic uppercasing, emphatic lengthening, abbreviations and the use of slangs and neologisms. Brody and Diakopoulos [2011] presented a study focused on emphatic lengthening and its impact on TSA. According to this study, emphatic lengthening is very frequent in Twitter, occurring in approximately one of every six tweets.

—*Data Sparsity*: Tweets contain a lot of noise due to the extensive use of incorrect English and misspellings. This phenomenon, which is known as data sparsity, has an impact on the overall performance of sentiment analysis. The main reason for data sparsity in Twitter is the fact that a great percentage of tweets' terms occur fewer than 10 times [Saif et al. 2012a] in the entire corpus. One interesting study focused on reducing data sparseness of tweets was presented by Saif et al. [2012a], who proposed semantic smoothing to reduce the sparseness.

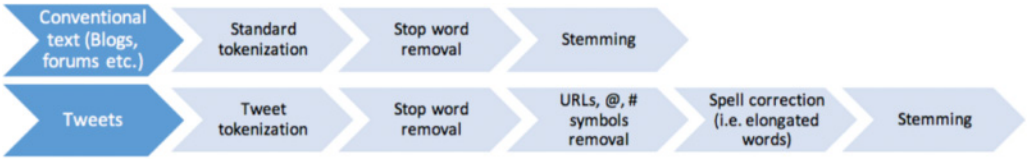

Fig. 3. Typical scenarios of preprocessing on standard text and tweets.

—*Negation*: Presence of negation words plays an important role in detecting the sentiment polarity of a message. The detection and the proper handling of negations is not trivial and remains a challenge. Detecting negations is important because they may cause the flip of a message's polarity (positive becomes negative or vice versa). A large number of researchers adopted a simple technique by simply reversing the polarity when a negation word is detected. A more advanced approach for handling negation in TSA was presented by Kiritchenko et al. [2014], who developed two separate lexicons, one with terms that usually appear in context with negations and one with terms that appear in context without negations. Kiritchenko et al. showed that negation of positive terms tend to imply negative sentiment, whereas in the case of negative terms the sentiment remains negative in the negated context.
—*Stop Words*: Stop words are common words that have low discrimination power (e.g., the, is, and who), and they are usually filtered out before processing the text. Typical pre-compiled stop-words lists are not suitable for Twitter and may even influence the TSA performance. For example, the word "like" generally is considered a stop word; however, it has an important sentiment discrimination power when doing TSA. To this end, there has been some work focused on building stop-words lists for Twitter. Saif et al. [2014] presented a study in which they examined the impact of removing stop words on TSA effectiveness. They used tweets from six different datasets on which they applied six different stop-word identification methods with the aim to examine how they affect the TSA performance. Their analysis included observations of the fluctuations on the level of data scarcity, the size of the classifiers feature space and classification performance.
—*Tokenization*: Another challenge related to TSA is the tokenization of the sentences. Rather than splitting on whitespace, Owoputi et al. [2013] proposed a Twitter-specific tokenizer. The tokenizer was proved to be effective when dealing with Twitter data.
—*Multilingual Content*: Tweets are written in a wide variety of languages, sometimes mixed even in the same message. The difficulty for language detection increases as a result of the tweets' short length. One of the studies that focused on multilingual TSA was presented by Narr et al. [2012], who developed a language-independent classifier that was evaluated over tweets in four languages. They showed that the proposed classifier performs effectively in multiple languages without needing extra process per additional language.
—*Multimodal Content*: In some cases, tweets contain multimodal content such as images or videos. Image and video analysis may be valuable for TSA, as it can provide useful information on determining who is the opinion holder or on the entity extraction. However, extracting features from multimodal content for TSA is still an under-explored area.

All of these challenges are very important and have to be considered for TSA. However, some of these challenges (negation, stop-word removal, and multimodal content) are encountered when addressing SA on all kinds of textual data. Figure 3 shows a typical preprocessing scenario of standard text and tweets. We observe that

additional steps are required in case of tweets in order to deal with some of their unique characteristics.

2.3. Feature Selection for Twitter Sentiment Analysis

The majority of SA and TSA methods detect sentiment based on a feature set. The selected features and their combination play an important role for detecting the sentiment of a text. Several types of textual features have been examined in the literature for online reviews and news articles including part-of-speech (POS) tags and information from sentiment lexicons. In the domain of microblogs, we can identify four different classes of textual features: *semantic*, *syntactic*, *stylistic*, and *Twitter-specific* features. Semantic, syntactic, and stylistic features include well-known features and have been used in the existing literature of SA of other genres such as reviews, blogs, and forums. Semantic features include terms that reveal negative or positive sentiment usually taken from sentiment lexicons or the semantic concept of terms. Syntactic features that are frequently applied include n-grams and part-of-speech tags. Stylistic features refer to the writing style used in Twitter, whereas the last class include features that emerged from the unique characteristics of tweets such as retweets or hashtags.

In most of the cases, TSA feature selection is based on approaches that were previously shown to be effective in other domains. An et al. [2014] used the chi-squared measure, which is a common statistical test for feature selection, whereas Kiritchenko et al. [2014] used a Pointwise Mutual Information (PMI) measure to identify terms that reveal sentiment. Also, a number of researchers analyzed the impact of different features on TSA and managed to establish feature selection criteria [Pak and Paroubek 2010; Agarwal et al. 2011; Kouloumpis et al. 2011]. In the remainder of this section, we present the most common features that have been used for TSA.

—*Semantic Features*: The most frequently used semantic features are *opinion words*, *sentiment words*, *semantic concepts*, and *negation*. Opinion words refer to words or phrases that are characterized as indicative of opinion, whereas sentiment words are indicative of positive or negative sentiment. Opinion and sentiment words and phrases are of the most used features in SA and can be extracted manually or semi-automatically from opinion and sentiment lexicons, respectively. In TSA, researchers have leveraged lexicons developed for other domains such as SentiWordNet [Esuli and Sebastiani 2006] and Multi-perspective Question Answering (MPQA) Opinion Corpus [Wilson et al. 2005] to either produce scores in an unsupervised way or as additional features for training machine-leaning methods in a supervised way. Other researchers examined the usefulness of the semantic concepts that are hidden in tweets. For example, Saif et al. [2012b] leveraged the entities related to the terms used in tweets and their sentiment to improve the performance of TSA. An additional feature that is very important for TSA and in general for SA is negation, which may flip the polarity of text. In fact, features that imply negations have been examined by a number of researchers [Pak and Paroubek 2010; Tai and Kao 2013; Kiritchenko et al. 2014].

—*Syntactic Features*: Syntactic features are the most used features, together with semantic features. These are typically *unigrams*, *bigrams*, *n-grams*, *terms' frequencies*, *POS*, *dependency trees*, and *coreference resolution*. With the intention to explore the impact of different terms on sentiment analysis, a number of studies assigned a binary weighting score (presence/absence) to the terms, whereas others use more advanced weighting schema considering frequencies of terms. In the literature, a classifier trained only on unigrams is frequently used as a baseline for comparison reasons. In addition to the terms, POS tags (e.g., nouns, verbs, and adjectives) is another syntactic feature that can capture information that indicates opinion as in

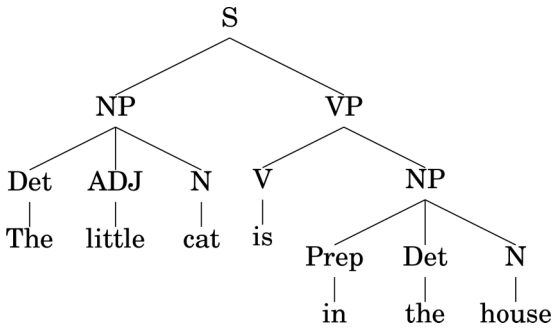

Fig. 4. The parse tree of the sentence "The little cat is in the house." In this example, S refers to a sentence, NP to a noun phrase, VP to a verb phrase, Det to a determiner, ADJ to an adjective, V to a verb, Prep to a preposition, and N to a noun.

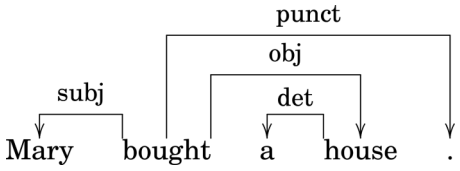

Fig. 5. The dependency tree of the sentence "Mary bought a house."

case of adjectives considered to be related with opinion. In the case of POS, results are controversial since there are studies that do not report improvements using them [Go et al. 2009; Kouloumpis et al. 2011], whereas others report at least small improvements [Pak and Paroubek 2010; Barbosa and Feng 2010; Agarwal et al. 2011]. Another feature is the dependency trees based on the notion that the words and other linguistic units are connected to each other by directed links. Dependency trees produce syntactical relations of the terms within a sentence. In dependency trees the verb is the center of the sentence and the rest linguistic units are connected with the verb with syntactical relations. One example of the syntactical relations is a word that is a subject in relation to a verb. This feature is important for entity-level sentiment analysis, as it indicates relations between the opinion words and the opinion target. Coreference resolution that occurs when two or more expressions refer to the same person or thing is an additional syntactic feature that has been examined for TSA [Zhang et al. 2011]. Trees are frequently used to illustrate the relations after the syntactic analysis of a sentence. Figure 4 shows the parse tree of the sentence "The little cat is in the house," on which one can easily view the POS tags of the sentence, whereas Figure 5 shows the dependency tree of the sentence "Mary bought a house," on which one can view the syntactical relations within the sentence.

—*Stylistic Features*: These include features emerging from the non-standard writing style that is used in Twitter. Some examples are *emoticons*, *intensifiers*, *abbreviations*, *slang terms*, and *punctuation marks*. One important feature is the presence of emoticons, whose usefulness has been extensively examined in the literature. Wikipedia[12] is a common source for obtaining an emoticon list [Agarwal et al. 2011]. Figure 6 illustrates a list of the most common emoticons. Another stylistic characteristic of tweets is the use of intensifiers, which are used to increase the emphasis of what is written and which include repeated characters, emphatic lengthening, and

[12]http://en.wikipedia.org/wiki/List_of_emoticon.

Fig. 6. A list with the most common emoticons.

emphatic uppercase. Also, the use of punctuation marks (e.g., exclamation marks etc.) is very common in Twitter.

—*Twitter-Specific Features*: Researchers have also examined some features that are specific of Twitter. These are *hashtags*, *retweets*, *replies*, *mentions*, *usernames*, *followers*, and *URLs*. A number of researchers have analyzed the impact of those features on TSA by considering their presence/absence or their frequency in a tweet [Barbosa and Feng 2010; Jiang et al. 2011].

Feature selection is not a trivial task and a thorough analysis is needed to detect the most useful features for each domain. To this end, Agarwal et al. [2011] and Kouloumpis et al. [2011] analyzed the usefulness of different features for TSA. Agarwal et al. [2011] proposed a feature-based model and performed a comprehensive set of experiments to examine the usefulness of various features, including POS and lexicon features. The analysis showed that the most useful combination is the one of POS with the polarity of words. Kouloumpis et al. [2011] also analyzed the impact of different features on TSA. This study was mostly focused on semantic and stylistic features, including emoticons, abbreviations, and the presence of intensifiers. Combinations of features that reveal the polarity of the terms with the n-grams managed to achieve the best performance. However, this study showed that POS had a negative impact on TSA, in contrast with the conclusions of the study performed by Agarwal et al. [2011].

The most typical feature selection process is to isolate words and other features and apply different feature selection and dimensionality reduction techniques with the aim to identify the most informative. One limitation of this approach is that it requires additional steps to handle other phenomena such as negation handling or sarcasm detection. In addition, when the list of candidate features grows a lot, finding the best feature combination is not always feasible. To address these limitations, researchers started exploring algorithms that are capable of obtaining learning representations of the data and these make it easier to extract useful information for building the classifier [Bengio et al. 2013]. To this end, researchers have started recently exploring deep-learning methods based on word embeddings that allow sentence structure and semantics understanding [Maas et al. 2011; Irsoy and Cardie 2014; Tang et al. 2014a].

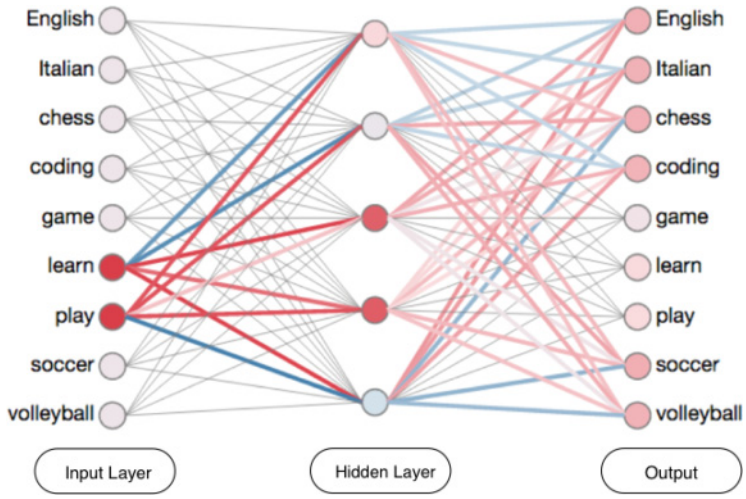

Fig. 7. A visualization of relations of word *eat* compared to the rest of the words. This figure was created with the help of *wevi: word embedding visual inspector* tool.

Table I. Example of a Confusion Matrix Showing the Performance of a Sentiment Analysis Method

	Predicted as *Positive*	Predicted as *Negative*
Are *Positive*	TP	FN
Are *Negative*	FP	TN

Word embeddings use numbers to represent the words where each of these numbers is a dimension. Once word embeddings have been trained, they can be used to extract words similarities or other relations. An example is illustrated in Figure 7 that shows the relations and similarities of the words *learn* and *play* in relation to the other words after the training phase. This figure was created with the help of *wevi: word embedding visual inspector* tool.[13]

2.4. Evaluation Metrics for Twitter Sentiment Analysis

TSA can be considered as a classification problem, since the goal in the typical scenario is to classify the opinion expressed in a tweet as positive or negative. The most frequently used evaluation metrics are *accuracy*, *precision*, *recall*, and *F-score*, adopted from traditional classification problems.

To better understand the metrics, we introduce one example. Consider that we want to evaluate the performance of a classifier or, in general, a method on its ability to classify a text as expressing positive or negative sentiment. Table I describes the performance of this method on a set of test data for which the sentiment is known. This table, also called a *confusion matrix*, shows the number of *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP), and *False Negatives* (FN) instances that are used to compare the predictions of the method with the ground truth. TP represents the number of instances that were predicted as positive and were indeed positive, whereas FP is the number of instances incorrectly predicted as positive. TN and FN have a corresponding meaning for the negative class.

Based on this matrix, we now present the most popular evaluation metrics having been used in the SA and TSA literature.

[13]https://ronxin.github.io/wevi/.

—*Accuracy*: Accuracy is the most frequently used evaluation metric and measures how often the method being evaluated made the correct prediction. It is calculated as the sum of the true predictions divided by the total number of predictions. That is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

—*Precision*: Precision represents the exactness of the method and is calculated as the ratio of instances that were predicted as positive and were indeed positive divided by the total number of instances that were predicted as positive. That is:

$$Precision = \frac{TP}{TP + FP}.$$

—*Recall*: Recall, which is also known as sensitivity, denotes the fraction of positive instances that were predicted to be positive and is calculated as:

$$Recall = \frac{TP}{TP + FN}.$$

—*F-score*: Usually, calculating recall and precision is not enough. A combination of the two is more appropriate to evaluate the performance of the methods. The F-score is the metric that combines recall and precision. This metric is also known as *harmonic F-score*, *F1-score*, or *F-measure accuracy* and is calculated as:

$$F\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

Finally, when the sentiment classification is formulated as a multi-class problem (i.e., it aims to classify a tweet as positive, negative, or neutral), it is common practice to calculate the positive, negative, and neutral F-score. However, there are approaches that do not predict the neutral class. This does not mean that the task is reduced to predicting only positive and negative tweets. These approaches should still be evaluated on the whole ground truth that includes neutral tweets.

3. TWITTER SENTIMENT ANALYSIS APPROACHES

In the literature, SA has been applied at three different levels: *document*, *sentence*, and *entity* levels. SA at the document level aims to identify the sentiment polarity expressed in the whole document. The sentence level SA aims to classify each sentence as positive or negative, whereas entity-level SA detects the sentiment polarity of a specific entity/target of a particular object.

Due to the length limitation, the majority of tweets contains a single sentence. Therefore, for the task of TSA there is no fundamental difference between document and sentence level. In case of tweets, SA can be applied on two levels: *message / sentence* and *entity* levels.

Four different classes can be identified in the literature of TSA:

—*Machine Learning*
—*Lexicon-Based*
—*Hybrid (Machine Learning & Lexicon-Based)*
—*Graph-Based*

The machine-learning approach employs a machine-learning method and a number of different features to build a classifier that can detect tweets that express opinion or sentiment. The lexicon-based approach uses a manually or automatically built list of

positive and negative terms to derive the polarity of the message or the entity under investigation. Methods that follow the hybrid approach combine machine-learning and lexicon-based methods to achieve a better performance [Khuc et al. 2012]. In contrast to machine-learning and lexicon-based approaches that can be applied to any type of text, the graph-based methods are applied to content with any type of social relations. The graph-based approach exploits social network properties to achieve better performance for TSA [Speriosu et al. 2011].

3.1. Machine-Learning Methods

The majority of the proposed methods that deal with TSA employs a classifier from the field of machine learning that is trained on various features of tweets. Some of the most applied classifiers are the Naïve Bayes (NB), Maximum Entropy (MaxEnt), Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Random Forest (RF), and Conditional Random Field (CRF). In the following, we review these approaches, classifying them in either supervised methods or ensembles.

3.1.1. Supervised Learning. One of the first studies dealing with TSA was carried out by Go et al. [2009], who treated the problem as a binary classification, classifying the tweets as either positive or negative. Due to the difficulty of manually tagging the sentiment of tweets, they employed distant supervision to build a machine-learning classifier. Go et al. used the technique that was demonstrated by Read [2005] to collect the data. To this aim, they used emoticons to differentiate between negative and positives tweets. Retweeted posts and messages containing both positive and negative emoticons were filtered out. The final training data set consisted of 1,600,000 messages that were equally distributed into the two categories of polarity. They examined NB, MaxEnt, and SVM classifiers, the same methods applied by Pang et al. [2002] for SA of movie reviews. Bigrams, unigrams, and POS tags were used as features. The authors drew interesting results such as that adding negation as an explicit feature with unigrams and using POS tags are not useful for polarity classification. They also reported that the most effective method was using NB with bigrams as features, which managed to achieve an accuracy of 82.7%.

Similarly to Go et al. [2009], Pak and Paroubek [2010] also used emoticons as labels to annotate about 300,000 tweets. Unlike Go et al., Pak and Paroubek [2010] tackled the problem as a multiclass classification task and classified tweets as positive, negative, or neutral. Neutral tweets were gathered from newspapers and magazines accounts, such as New York Times and Washington Posts. The authors also presented the frequency distribution of terms after a linguistic analysis. One conclusion was that the pronoun for the first person and the adjectives connected to that usually occurred in opinionated messages. They compared the performance of SVM, MNB, and CRF using different features including unigrams, bigrams, n-grams, and the position of n-grams. Their results showed that the best combination was MNB together with n-grams and POS tags. They also observed that the performance increased with more training data.

Barbosa and Feng [2010] tackled the TSA problem with a two-step classifier. The first step determined whether the message was opinionated or not while the second step aimed to further classify the tweet as positive or negative. Barbosa and Feng used information from three different sentiment detection tools to annotate a collection of tweets. Tweets assigned different sentiment polarity from the detection tools were removed resulting in a training dataset of 200,000 tweets. They used meta-index and syntax features to train the classifiers. POS tags and words' polarity using the MPQA lexicon [Wiebe et al. 2006] are examples of meta-index features, whereas syntax features include retweets, hashtags, URLs, emoticons, and so on. Also, they normalized the frequency of each feature by the number of tokens in the tweet. The best results

occurred with the SVM classifier obtaining an accuracy of 81.9% for the subjectivity detection and 81.3% for the polarity detection. An interesting conclusion was that the syntax features were more important for the subjectivity detection, whereas the meta-features were so for the polarity detection.

Davidov et al. [2010] also presented a supervised approach that was similar to a k-Nearest Neighbors algorithm (kNN). In contrast to the previous approaches, they leveraged the hashtags and emoticons in tweets for collecting training data. Apart from the traditional features, Davidov et al. also used hashtags, smileys, punctuations, and frequent patterns and achieved an average harmonic F-score of 86.0% for binary classification for their kNN-like classification strategy.

One of the most used classifiers for addressing TSA is the SVM classifier. Bakliwal et al. [2012] employed an SVM classifier trained on 11 features to address TSA. They employed different pre-processing techniques one by one in order to measure their effectiveness. Spelling correction, stemming, and stop-words removal managed to increase the accuracy of the classifier. Two different datasets were used to evaluate their approach, the Stanford dataset [Go et al. 2009] and the Mejaj [Bora 2012]. The best results were achieved by the combination of NLP- and Twitter-specific features.

Similarly, Mohammad et al. [2013] employed an SVM classifier on the dataset given by a SemEval-2013 evaluation campaign [Nakov et al. 2013]. They represented each tweet as a feature vector that included word/character n-grams, POS, capital words, hashtags, lexicons, punctuation, emoticons, emphatic lengthening, and negation. They observed that the SVM classifier trained using those features performed better than the baseline trained on unigrams. Their method obtained an F-score of 69.02% for the message-level analysis and 88.93% for the term-level task. The authors concluded that the most useful features are the lexicon features and the n-grams.

A linear-kernel SVM method was proposed by Kiritchenko et al. [2014] for TSA. The proposed system was based on a supervised statistical text classification approach. Kiritchenko et al. utilized a variety of surface-form, sentiment, and semantic features, the majority of which were derived from tweet-specific lexicons. The linear-kernel SVM managed to outperform the MaxEnt classifier.

A three-step cascaded classifier framework for TSA was presented by Asiaee et al. [2012]. In the first step, they identified the tweets of the topic of interest. In the second step, they identified the tweets with sentiment, whereas in the last step the tweets were annotated with sentiment polarity. They studied the TSA performance of a number of classical methods and also proposed new algorithms including kNN, NB, weighted SVM, and Dictionary Learning. One interesting result of this study is that the performance of the classification was improved in a low-dimensional space.

The selection of features is very important for the effectiveness of the supervised methods. To this end, a number of researchers analyzed the impact of different features on TSA [Agarwal et al. 2011; Aisopos et al. 2011; Kouloumpis et al. 2011; Saif et al. 2012b; Hamdan et al. 2013]. Agarwal et al. [2011] performed a study to analyze the different TSA features and to examine the performance of different methods on TSA. In total, they compared three methods; the baseline method that considered only unigrams, the partial tree kernels, and the feature-based model. They showed that both the tree-kernel and the feature-based model outperformed the baseline. In addition, after extensive analysis of the usefulness of the features, they concluded that the most helpful features for TSA were those that refer to the sentiment polarity of a term.

Aisopos et al. [2011] suggested the use of n-gram graphs to improve classification accuracy. The authors employed two classification algorithms: MNB and a C4.5 tree classifier that were evaluated on about 3 million tweets. The training data annotation was based on the presence or absence of emoticons. Extensive experiments showed

that the best model is a C4.5 tree classifier trained on a 4-gram graph with distance-weighted edges. Their best model obtained 66.77% accuracy for binary classification and 50.67% for the three-way classification.

Another study that investigated the usefulness of different features was presented by Kouloumpis et al. [2011], who specifically focused on the linguistic features. The authors used AdaBoost to detect the polarity of the sentiment. Training data were collected using the existing hashtags in tweets that indicated sentiment. The best performance was achieved by combining n-grams, lexicon, and microblogging features, whereas POS was not a good indicator of sentiment.

A conceptual semantic approach was proposed by Saif et al. [2012b], who examined a set of semantic features. The semantic features consisted of semantic concepts (e.g., person, city) that represented extracted entities (e.g., Steve Jobs, London). Saif et al. [2012b] incorporated the semantic features into an NB classifier. Their results showed an average increase of an F-harmonic accuracy score for identifying negative and positive sentiment of around 6.5% and 4.8% over the baselines of unigrams and POS features, respectively.

Hamdan et al. [2013] proposed to use many features and resources with the aim to achieve a good performance on TSA. Examined features included concepts from DBPedia, verb groups and adjectives from WordNet, and senti-features from SentiWordNet. Hamdan et al. [2013] also employed a dictionary of emotions, abbreviations, and slang words to improve the accuracy of TSA. Their method managed to improve F-measure accuracy by 2% and 4% by considering these features compared to the SVM trained on unigrams and NB classifier, respectively.

Instead of applying TSA at a tweet level, Jiang et al. [2011] used a machine-learning approach to address the task of aspect-based TSA. The proposed method combined target-independent and target-dependent features and manually defined rules to detect the syntactic patterns that showed if a term was related to a specific object. They also employed a binary SVM for subjectivity and polarity classification. They utilized microblog-specific features such as retweets, replies, and mentions to create a graph that reflects the similarities of tweets. Jiang et al. [2011] reported an accuracy of about 68% for target identification and of about 79% for sentiment classification. Utilizing target identification improved the accuracy of sentiment classification, resulting in an accuracy of 85%.

Finally, Aston et al. [2014] studied the sentiment analysis problem in tweet data streams. They examined different supervised methods with limitations on memory and processing time. Tweets were represented using character n-grams. The large number of features was reduced by selecting the top N features of a gram. The following six different evaluation algorithms were used for the selection of the top features: Chi Squared, Filtered Feature, Gain Ratio, Info Gain, One R, and Relief. In their study, they compared three different versions of the Perceptron classifier (Best Learning Rate, Voted) as well as different combinations of them. The combination of the methods Best Learning Rate and Voted resulted in the best performance, that is, an F-score of 85% and 78% for subjectivity and sentiment analysis, respectively.

Table II summarizes the articles that employed a supervised method to address TSA. The first column shows information for the reference. The second column refers to the purpose of the article. The objective can be *Twitter Sentiment Analysis* (TSA) or *Entity Twitter Sentiment Analysis* (entity-TSA). The third column shows the algorithms used in each study. This column provides a full list of the different algorithms that were examined and adopted in the article in addition to the proposed method. Features employed by researchers (if reported in the article) are presented in the fourth column. The last column refers to the dataset(s) used in the corresponding study. If the authors used any of the datasets that are presented in Section 5, then we mention the

Table II. Summary of the Articles Employed a Supervised Method to Address TSA

Study	Task	Algorithms	Features	Dataset
Go et al. [2009]	TSA	NB, MaxEnt, SVM	unigrams, bigrams, POS	STS
Pak and Paroubek [2010]	TSA	MNB, SVM, CRF	unigrams, bigrams, trigrams, POS	own
Barbosa and Feng [2010]	TSA	SVM	meta-features (POS, polarity-MPQA), tweet syntax (i.e., retweet, hashtags, emoticons, links etc.)	own
Davidov et al. [2010]	TSA	kNN	word and n-gram based, punctuation-based, pattern-based	OC
Bakliwal et al. [2012]	TSA	SVM, NB	words' polarity, unigrams, bigrams, emoticons, hashtags, URLs, targets etc.	STS, Mejaj [Bora 2012]
Mohammad et al. [2013]	TSA	SVM	word/character n-grams, POS, caps, lexicons, punctuation, negation, tweet-based	SemEval-2013
Kiritchenko et al. [2014]	TSA	linear kernel SVM, MaxEnt	word/character n-grams, POS, caps, punctuation, emoticons, automatic sentiment lexicons, polarity, emphatic lengthening	SemEval-2013
Asiaee et al. [2012]	TSA	dictionary learning, WSVM, NB, kNN,		DETC
Agarwal et al. [2011]	TSA	SVM	POS, unigrams, DAL lexicon, caps, exclamation etc.	own
Aisopos et al. [2011]	TSA	MNB, C4.5 tree	n-grams	own
Kouloumpis et al. [2011]	TSA	AdaBoost	N-gram with lexicon features, twitter-based, POS	STS, ETC
Saif et al. [2012b]	TSA	NB	unigrams, POS, sentiment-topic, semantic features	STS, HCR, OMD
Hamdan et al. [2013]	TSA	SVM, NB	unigrams, concepts (DBPedia), verb groups/adjectives (WordNet) and senti-features (SentiWordNet)	SemEval-2013
Jiang et al. [2011]	entity-TSA	SVM	unigrams, emoticons, hashtags, punctuation the General Inquirer lexicon	own
Aston et al. [2014]	TSA	Perceptron with Best Learning Rate, Voted Perceptron, Ensemble Method	character n-grams	Sanders

abbreviation of the dataset. In case the authors created a different dataset to evaluate their method, then we specify this with the term own. In those cases, there are not many details described for the dataset, or the dataset is not available publicly.

3.1.2. Classifier Ensembles. Recently, the concept of combining classifiers has been proposed as a new direction for improving the performance of individual classifiers. This approach, known as *ensembles classifiers*, has been also applied to TSA.

Lin and Kolcz [2012] applied ensemble classifiers on large-scale datasets crawled from Twitter. They linearly combined LR classifiers trained from hashed byte 4-grams and applied ensembles of different sizes, formed by different models and trained on different sets. They managed to improve accuracy of sentiment analysis by employing

classifiers ensembles. One limitation was that they evaluated their approach only using a single algorithm, with a single dataset.

On the contrary, da Silva et al. [2014] employed more classifiers compared to Lin and Kolcz [2012]. They combined RF, SVM, MNB, and LR. Two different approaches were explored for feature representation: bag-of-words and feature hashing. Their results showed that classifier ensembles formed by MNB, SVM, RF, and LR could improve classification accuracy. Also, training ensembles on bag-of-words and lexicons was more effective than using feature hashing. Their approach was evaluated using the Sanders, STS, OMD, and HCR datasets.

A different approach was considered by Hassan et al. [2013], who proposed a bootstrapping ensemble framework. In addition to TSA, the framework could also cope with class imbalance, data sparsity, and representational richness issues. They used some of the most common features, including unigrams and bigrams, POS, and semantic features. In addition, the authors claimed that the proposed framework could be used to build sentiment time series.

Ensemble classifiers have also been applied to address expression-level TSA [Rodríguez-Penagos et al. 2013; Clark and Wicentowski 2013]. Rodríguez-Penagos et al. [2013] proposed to combine machine-learning and rule-based approaches. They examined the combination of CRF, SVM, and a heuristic approach. For the heuristic approach, they used sentiment words, negation markers, and quantifiers. Clark and Wicentowski [2013] suggested the combination of NB classifiers. They used N-grams, sentiment words, POS, and special tokens (i.e., emoticons) to train the classifiers. Each classifier was trained on a single feature. They managed to obtain an F-measure of 0.672 at the expression-level classification.

Finally, Kouloumpis et al. [2011] and Aston et al. [2014] explored methods from both supervised and classifier ensembles approaches and for this reason can be also categorized in a classifier ensembles approach. Kouloumpis et al. [2011] explored the effectiveness of the meta-classifier AdaBoost and showed that it outperforms SVM. Aston et al. [2014] explored the effectiveness of different versions of the Perceptron classifier of their combination. Their results were consistent with those obtained by Kouloumpis et al. [2011], since their best performance was also achieved with classifier ensembles and, more specifically, when they combined the Best Learning Rate and Voted methods.

Table III summarizes the articles that employed classifier ensembles to address TSA.

3.1.3. Deep Learning. Deep learning is one of the fastest-growing fields of machine learning and is applied to solve perceptual problems such as image recognition and understanding natural languages. Deep learning uses neural networks to learn many levels of abstraction. In text-related tasks, deep-learning approaches typically include two steps. First, they learn word embeddings from the text collection and these are then applied to produce the representations of the documents. In relation to sentiment analysis, deep learning is used to learn word embeddings from large amounts of text data [Maas et al. 2011]. Recently, Tang et al. [2015a] used deep learning to learn semantic representations of user and products, whereas Tang et al. [2015b] used deep learning for review prediction.

Deep learning has also been explored for TSA. Tang et al. [2014b] proposed to learn sentiment specific word embeddings (SSWE) from tweets that were collected using distant supervision. In their study, they developed three neural networks to learn SSWE that were then used as features for TSA. The methods were evaluated on the SemEval-2013 dataset. The best result in terms of F1 score was 86.58% and was obtained by combining SSWE with sentiment lexicons and the same features with those used by Mohammad et al. [2013]. SSWE was also evaluated on SemEval-2014 [Tang et al. 2014a] and obtained the second ranking. The SSWE approach obtained an F1

Table III. Summary of the Articles Employed Classifier Ensembles to Address TSA

Study	Task	Algorithms	Features	Dataset
Lin and Kolcz [2012]	TSA	Logistic Regression, Majority vote	feature hashing	own
da Silva et al. [2014]	TSA	RF, SVM, MNB, LR	bag-of-words, feature hashing	Sanders, STS, OMD, HCR
Hassan et al. [2013]	TSA	RBF Neural Network, RandomTree, REP Tree, NB, Bayes Net, LR and SVM, bootstrap model	unigrams, bigrams, POS, semantic features	Sanders
Clark and Wicentowski [2013]	entity-TSA	NB, Weighted voting scheme	N-gram, lexicon, polarity strength	SemEval-2013
Rodríguez-Penagos et al. [2013]	TSA	CRF, SVM, heuristic method, Majority vote, upper bound, ensemble vote	N-gram, POS, tweet-based features, SentiWordNet	SemEval-2013
Kouloumpis et al. [2011]	TSA	AdaBoost	N-gram with lexicon features, twitter-based, POS	STS, ETC
Aston et al. [2014]	TSA	Perceptron with Best Learning Rate, Voted Perceptron, Ensemble Method	character n-grams	Sanders

Table IV. Summary of the Articles Employed Deep Learning to Address TSA

Study	Task	Algorithms	Features	Dataset
Tang et al. [2014b]	TSA	$SSWE_h$, $SSWE_r$, $SSWE_u$	sentiment lexicons and features used in Mohammad et al. [2013]	SemEval-2013
Tang et al. [2014a]	TSA	SSWE, Coooolll	sentiment lexicons, emoticons, negation, punctuation, n-grams, cluster, lengthening, caps	SemEval-2013, SemEval-2014
Dong et al. [2014]	entity-TSA	AdaRNN	dependency tree, unigrams, bigrams	own
Vo and Zhang [2015]	entity-TSA	Target-ind, Target-dep	sentiment lexicons, embeddings, pooling functions	[Dong et al. 2014]

score of 87.61% when it was combined with a number of features, including sentiment lexicons, emoticons, and emphatic lengthening.

Dong et al. [2014] proposed an Adaptive Recursive Neural Network (AdaRNN) for entity-level TSA. This method used a dependency tree in order to find the words syntactically related with the target and to propagate the sentiment from sentiment words to the targets. AdaRNN was evaluated on a manually annotated dataset consisting of 6248 training and 692 testing tweets and managed to obtain an F1 score of 65.9%.

The dataset created by Dong et al. was also used by Vo and Zhang [2015], who proposed using a rich set of automatic features. According to their approach, the tweet is split into a left and right context in relation to a specific target. Word embeddings were then used to model the interactions of the two contexts that were used to detect the sentiment towards the target. The authors explored a range of pooling functions to automatically extract rich features. Their approach outperformed the AdaRNN by obtaining an F1 score of 69.9%.

Table IV summarizes the articles that employed deep learning to address TSA.

3.2. Lexicon-Based Methods

Lexicon-based methods leverage lists of words annotated by polarity or polarity score to determine the overall opinion score of a given text. The main advantage of these methods is that they do not require training data. Lexicon-based approaches have been extensively applied on conventional text such as blogs, forums, and product reviews [Turney 2002; Taboada et al. 2011; Ding et al. 2008]. However, they are less explored in TSA compared to machine-learning methods. The main reason is the uniqueness of the text on Twitter that not only contains a large number of textual peculiarities and colloquial expressions such as yolo and gr8 but also has a dynamic nature with new expressions and hashtags emerging from time to time.

One of the most well-known lexicon-based algorithms developed for social media is SentiStrength [Thelwall et al. 2010]. SentiStrength can effectively identify the sentiment strength of informal text including tweets using a human-coded lexicon that contains words and phrases that are frequently confronted in social media. Apart from the sentiment lexicon that contains about 700 words, SentiStrength uses a list of emoticons, negations, and boosting words to assign the sentiment to a text. Initially, the algorithm was tested on MySpace comments. The algorithm was extended by Thelwall et al. [2012] by introducing idiom lists and new sentiment words in the lexicon and by strength boosting using emphatic lengthening. SentiStrength was compared with many machine-learning approaches and tested on six different datasets, including a dataset with tweets posts.

Ortega et al. [2013] proposed a three-step technique for TSA. Pre-processing was performed in the first step and polarity detection in the second step. In the last step, they performed rule-based classification. Polarity detection and rule-based classification were based on WordNet and SentiWordNet. Their approach managed to achieve good results when evaluated on the SemEval-2013 dataset [Nakov et al. 2013]. However, the authors did not compare their method with existing techniques to prove its effectiveness.

The SemEval-2013 dataset was also used by Reckman et al. [2013] to evaluate a rule-based system. Their system was based on handwritten rules, each of which had the form of a pattern. This system performed very well on TSA and was one of the top-performing systems on SemEval-2013.

An interesting method was presented by Hu et al. [2013a], who proposed an unsupervised sentiment analysis method based on emotional signals. The emotional signals were divided into two categories: emotion correlation and emotion indication. The Emotional Signals for unsupervised Sentiment Analysis (ESSA) approach was built on the orthogonal nonnegative matrix tri-factorization model. Two different datasets were used for the evaluation of the ESSA approach, the STS [Go et al. 2009] and the OMD [Shamma et al. 2009] datasets. Extensive experiments indicated the effectiveness of the proposed framework as well as the roles of different emotional signals in sentiment analysis.

Saif et al. [2016] presented SentiCircles, a lexicon-based approach to address TSA. SentiCircles updated the pre-assigned scores and polarity of words in sentiment lexicons by considering the patterns of words that co-occur in different contexts. The SentiCircles approach was evaluated on three different datasets: OMD [Shamma et al. 2009], HCR [Speriosu et al. 2011], and STS-Gold [Saif et al. 2013]. Extensive experiments proved the effectiveness of the method that outperformed the methods based on MPQA and SentiWordNet.

A number of works proposed to expand the list of sentiment words by leveraging semantic relationships such as synonyms and antonyms. The most typical scenario is to define a small amount of sentiment words, frequently annotated by hand, and then to expand this initial list by adding words with similar semantics [Kim and Hovy

Table V. Summary of the Articles Employed a Lexicon-Based Method to Address TSA

Study	Task	Algorithms	Features	Dataset
Thelwall et al. [2012]	TSA	SentiStrength	polarity, emoticons, negations, emphatic lengthening, boosting words etc.	SS-Tweet
Ortega et al. [2013]	TSA	clustering-based word sense disambiguation (WSD), lexicon-based classifier	WordNet, SentiWordNet	SemEval-2013
Reckman et al. [2013]	TSA	rule-based		SemEval-2013
Hu et al. [2013a]	TSA	ESSA	emoticons, sentiment lexicon (MPQA), textual similarity, word co-occurrence	STS, OMD
Saif et al. [2016]	entity-TSA	SentiCircles	SentiWordNet, MPQA, Thelwall-Lexicon	OMD, HCR, STS-Gold
Feng et al. [2011]	TSA	lexicon-based	connotation lexicons	SemEval-2007, STS

2004] frequently obtained from WordNet. The problem with this approach is that the expansion of the opinion information is restricted and dependent on the initial list of seed words. To overcome this problem, Feng et al. [2011] proposed using connotation lexicons to enclose subtle dimensions of a word's sentiment. In their work, they first defined a list of seed words and then they used a graph-based algorithm based on PageRank and HITS to learn the connotation lexicon together with the connotative predicates. The evaluation of their approach based on SemEval-2007 and STS datasets demonstrated promising results for using connotation lexicons on TSA.

Table V summarizes the articles that employed a lexicon-based method to address TSA.

3.3. Hybrid Methods

A number of researchers combined machine-learning and lexicon-based approaches. An interesting study was presented by Zhang et al. [2011], who proposed a hybrid method to address entity-based TSA. For each of the entities *Obama, Harry Potter, Tangled, iPad, and Packers* they computed a sentiment score based on their proximity to words from a sentiment lexicon. They proposed a rule-based algorithm that also considered comparative judgments, negation, and expressions that were likely to change the orientation of a phrase. To collect more annotated data and enhance the recall of the proposed method, they identified additional subjective terms using Chi-square. The SVM classifier was then applied for sentiment polarity detection.

Another interesting hybrid method was presented by Ghiassi et al. [2013], who combined dynamic artificial neural network with n-gram. Emoticons and tweets that contained the word *love* or *hate* or their synonyms were used as features to build the two classifiers: SVM and a Dynamic Architecture for Artificial Neural Networks (DAN2). The proposed approach was tested on a collection of tweets crawled using the subject *Justin Bieber*. The results showed that DAN2 managed to outperform SVM.

In the approach presented by Kumar and Sebastian [2012], a log-linear classifier was combined with a dictionary-based method that calculated the semantic orientation of the adjectives and of the verbs/adverbs, respectively. A simple linear equation was then used for the calculation of a tweet's overall sentiment. They also performed preprocessing that included removal of URLs, replies and hashtags, spelling correction,

Table VI. Summary of the Articles Combined Machine-Learning and Lexicon-Based Methods to Address TSA

Study	Task	Algorithms	Features	Dataset
Zhang et al. [2011]	entity-TSA	SVM	unigrams, emoticons, hashtags, lexicon [Ding et al. 2008]	own
Ghiassi et al. [2013]	TSA	n-gram analysis, SVM, DAN2	emoticons, tweets containing the words 'love' or 'hate'	own
Kumar and Sebastian [2012]	TSA	corpus-based, dictionary-based, log-liner regression	punctuation, WordNet, emoticons, POS	own
Khuc et al. [2012]	TSA	lexicon-based, Online Logistic Regression	sentiment lexicon, POS, bigrams	own
Khan et al. [2014]	TSA	EEC, IPC, SWNC	emoticons, positive and negative words, SentiWordNet dictionary	own

replacement of emoticons by their polarity, and POS tagging. The authors claimed that their proposed system was able to effectively detect the tweets' polarity.

Khuc et al. [2012] also combined a lexicon-based approach with a classifier to improve TSA accuracy. They considered the MapReduce framework to create a co-occurrence matrix based on bigram phrases. The cosine similarity between words is then computed and the edges with low cosine score are removed. Then, they combined the score generated using a simple lexicon-based approach with a classifier. For the machine-learning algorithm, they used the Online LR approach. Their experiments showed that the hybrid approach outperformed the simple lexicon-based classifier that was only based on words/phrases that indicated sentiment.

The framework presented by Khan et al. [2014] was built by a three-step process, the last step of which was based on a hybrid TSA method. The first step included data acquisition using the Twitter API, followed by pre-processing of tweets. Pre-processing included detection of slangs and abbreviations, lemmatization and correction, and stopwords removal. The pre-processed tweets were passed to the sentiment classifier. The Polarity Classification Algorithm (PCA) sentiment classifier detected sentiment on a tweet based on Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC), and SentiWordNet Classifier (SWNC). A set of emoticons, a list of sentiment words, and SentiWordNet dictionary were used by EEC, IPC, and SWNC classifiers, respectively. The experiments showed that the final hybrid classification managed to outperform the performance of using any of the EEC, IPC, or SWNC classifiers.

Table VI summarizes the articles that combined machine-learning and lexicon-based methods to address TSA.

3.4. Graph-Based Methods

Although machine-learning methods achieve a decent performance on TSA, they require a large number of annotated data. Label propagation is a method that can reduce the demand of the annotated data. To this end, a number of researchers utilized the Twitter social graph under the assumption that people influence one another. Label propagation is a semi-supervised method in which labels are distributed to nodes using the connection graphs.

Speriosu et al. [2011] were some of the firsts to apply a label propagation method for TSA. The proposed method leveraged the Twitter follower graph under the assumption that people influence one another. Users, tweets, unigrams, bigrams, hashtags, and

Table VII. Summary of the Articles Employed a Graph-Based Method to Address TSA

Study	Task	Algorithms	Features	Dataset
Speriosu et al. [2011]	TSA	LexRatio, MaxEnt, LProp	N-gram, hashtags, emoticons, lexicon [Wilson et al. 2005], Twitter follower graph	STS, OMD, HCR
Cui et al. [2011]	TSA	graph propagation	emoticons, punctuation, SentiWordNet	STS
Wang et al. [2011]	TSA	SVM-voting, Loopy Belief Propagation, Relaxation Labeling, Iterative Classification Algorithm	unigrams, punctuation, emoticons, lexicon	own
Tan et al. [2011]	TSA	SVM Vote, HGM-NoLearning, HGM-Learning	followers/followees, textual features	own

emoticons were used as nodes for the construction of the graph. The proposed label propagation method outperformed a lexicon-based approach and a MaxEnt classifier.

Cui et al. [2011] also tackled TSA with a label propagation method based on analysis of emotion tokens. Cui et al. first extracted the emotion tokens from tweets. A graph propagation method was then used to assign polarities to the tokens. In the last step, they analyzed and classified the emotion tokens. The emotion tokens included emoticons, repeating punctuations, and repeating letters. Their approach managed to perform well in analyzing sentiment of messages written in any natural language.

Instead of using emotion tokens, Wang et al. [2011] proposed a graph-based model that leveraged co-occurrence of hashtags to classify the sentiment of certain hashtags. They proposed different algorithms (Loopy Belief Propagation, Relaxation Labeling, Iterative Classification Algorithm) that were compared to an SVM voting. The SVM was trained with several features, including unigrams, punctuation, and emoticons. The Loopy Belief Propagation algorithm managed to achieve the best performance in terms of accuracy compared to the other tested methods.

Tan et al. [2011] leveraged users' social relations to address user-level TSA. Their study showed that connected users share the same sentiment. Also, they empirically proved that if two users have the same sentiment, then it is more likely to have a connection in a social network. They compared three methods: SVM Vote, Heterogeneous Graph Model with Direct estimation from simple statistics (HGM-NoLearning), and Heterogeneous Graph Model with SampleRank (HGM-Learning). The authors evaluated their method on tweets about politicians and showed that user-level TSA could be significantly improved when considering the users' connections within a social network.

Table VII summarizes the articles employed a graph-based method to address TSA.

3.5. Other Methods

In the TSA literature there are some techniques that cannot be roughly categorized in any of the above categories. Formal Concept Analysis (FCA), proposed by Kontopoulos et al. [2013], is one of those techniques. Kontopoulos et al. used concept analysis to build an ontology domain model. They proposed a method in which tweets were broken down into a set of aspects that were relevant to the subject. Their model was applied and evaluated on the domain of smart phones. Considering that the model detected

Table VIII. Summary of the Articles Employed Other Methods to Address TSA

Study	Task	Algorithms	Features	Dataset
Kontopoulos et al. [2013]	entity-TSA	FCA	WordNet, OpenDover[14]	own
Korenek and Šimko [2014]	entity-TSA	appraisal theory, SVM	tweet-specific, linguistic, appraisal features	Sanders
Hu et al. [2013b]	TSA	SANT	unigrams	STS, OMD

aspects of the domain and assigned scores to them, they managed to obtain a more detailed analysis of sentiments towards a specific topic.

On the contrary, Korenek and Šimko [2014] leveraged appraisal theory to determine the sentiment of the main entity of a tweet. An appraisal dictionary with a list of annotated terms was created. The proposed approach was evaluated on Sanders dataset and outperformed the baseline [Go et al. 2009] by obtaining an accuracy of 87.57%.

Another approach that falls into this category is the Sociological Approach to handling Noisy and short Texts (SANT) proposed by Hu et al. [2013b]. SANT was based on the characteristics of Twitter as networked data. In particular, the authors presented a method that incorporated Sentiment Consistency and Emotional Contagion theories into the supervised learning process. Experimental results showed that these social theories were effective for TSA.

Table VIII summarizes the articles that employed other methods that cannot be roughly categorized to address TSA.

3.6. Discussion

From the above, we notice that the machine-learning approach is the most popular on TSA. The majority of the approaches employ a traditional machine-learning method, which is trained on a set of features. In an attempt to improve the performance and generate more precise results, some researchers combined several classifiers. The classifiers ensembles tend to perform better than using a single classifier.

In general, the machine-learning methods have some limitations. First, their performance depends on the number of training data, and, for this reason, they usually require a large amount of annotated tweets to obtain a high performance. However, annotating tweets is expensive due to the fact that the content of Twitter is continuously changing. Distant supervision is one alternative to get a large amount of annotated tweets. However, the annotation quality using this approach is low and can harm the performance of the classifier. Label propagation is another alternative that can reduce the demand of the annotated tweets. Therefore, there are works categorized as graph-based that leveraged the Twitter social graph under the assumption that people influence one another.

Another limitation of machine-learning approaches is that they are domain dependent. That means that a classifier can perform very well when it is applied on the same domain to the one it was trained. However, its performance decreases when it is applied to a different domain. That means that the classifier needs to be retrained in order to perform well on a different domain.

The effectiveness of the traditional machine-learning approaches depends on the set of selected features. The majority of the works determine a set of features on which the classifier is trained. This approach is not able to capture some phenomena, such as negation detection, that may change the sentiment of a tweet. Recently, researchers

[14]http://opendover.nl/.

started exploring algorithms that are capable of learning representations of data to overcome these limitations [Bengio et al. 2013]. To this end, researchers have started recently exploring deep-learning methods based on word embedding methods that allow sentence structure and semantics understanding [Maas et al. 2011; Irsoy and Cardie 2014; Tang et al. 2014a].

On the other hand, a number of works applied lexicon-based methods that rely on sentiment lexicons. One strength of these approaches is that they do not require annotated data. However, they rely on static lists of words. That means that a word that is not in the lexicon is not considered. Especially for Twitter, which has a continuously changing content, the lists have to be updated frequently.

Another limitation is that the lexicons are context independent and do not consider that words' sentiments depend on context. One typical example is the word "small" that expresses opposite opinions in the following two sentences: "*The size of the phone is small and fits in my pocket*" (positive) and *"The buttons on the keyboard are very small*" (negative). There have been attempts to train algorithms that alleviate some of these limitations [Thelwall et al. 2010], but they necessitate frequent retraining.

In an attempt to overcome the limitations of the machine-learning and lexicon-based methods, a number of works proposed hybrid approaches. One strength of hybrid methods is that they overcome some of the limitations of ML approaches using the lexicon-based methods and vice versa. For example, they can avoid manual labeling of training data by using the results of a lexicon-based method. However, hybrid methods require a high computational complexity.

The last category of methods is the graph-based method that includes approaches that exploit the Twitter social graph and its attributes. These approaches do not require a large amount of manually annotated data and they leverage connections of users and tweets (i.e., followers, replies, and past tweets of a user) to automatically collect more annotated data. However, these methods are domain specific due to the fact that the sentiment lexicons and the exploited relations are domain specific.

4. RELATED FIELDS

There are some tasks related to TSA that have recently attracted the interest of researchers. In this section, we discuss these tasks as well as how they are addressed.

4.1. Twitter-Based Opinion Retrieval

Twitter-based opinion retrieval aims to identify tweets that are relevant to a user's query and also express opinion about it. Opinion retrieval that is a sub-field of OM combines approaches from information retrieval and opinion mining [Paltoglou and Giachanou 2014]. Opinion retrieval in Twitter is under-explored in the literature. Some exceptions are Luo et al. [2013a] and Luo et al. [2013b]. Luo et al. [2013a] were the first to explore opinion retrieval in Twitter. They proposed a learning rank model to address the problem of retrieving relevant and opinionated tweets towards a user's query by leveraging social and opinionatedness information. They showed that integrating links, mentions, and author information in the model could improve the opinion retrieval performance.

A learning to rank method was also applied by Luo et al. [2013b]. However, this study considered the problem of propagated opinion retrieval that aimed to identify tweets that were relevant to a topic, express an opinion about it, and then be retweeted. To calculate the opinionatedness of tweets, Luo et al. [2013b] proposed to use social and structural information. Additionally, they used a set of features including retweetability, opinionatedness, and textual quality of the tweet. They reported significant improvements of using those features over the baselines.

Recently, Giachanou et al. [2016] proposed to use topic-specific stylistic variations assuming that the number of stylistic variations depends on the topic that is discussed in a tweet. Giachanou et al. first applied topic modeling to extract the topic of a tweet and then calculated the opinion based on the tweet's stylistic variations and opinionated terms. They explored the usefulness of emoticons, exclamation marks, emphatic lengthening, and opinionated hashtags. The reported results showed that the importance of stylistic variations in indicating opinionatedness depends on the tweet's topic as the proposed approach significantly outperformed the baselines.

4.2. Tracking Sentiments Over Time

The development of models that focus on *tracking sentiments over time* has also been a hot topic and has been recently applied on tweets, too. An et al. [2014] presented a study with the aim to understand whether mining social media data can be used to yield insights on climate change sentiment. They combined classical sentiment analysis algorithms, data-mining techniques, and time series methods with the aim to detect and track sentiment regarding climate change from Twitter feeds. The authors claimed that sudden change in sentiment polarity may be caused by major climate events. However, they reported an important variation in sentiment polarity that implied significant uncertainty in overall sentiment.

Bollen and Pepe [2011] performed a sentiment analysis of all public tweets posted from the August 1 to December 20, 2008. Bollen and Pepe [2011] mapped every day to a six-dimensional mood vector (tension, depression, anger, vigor, fatigue, confusion). A psychometric instrument called the Profile of Mood States (POMS) was used to extract and analyze those moods. They compared the results to the timeline of cultural, social, economic, and political events that occurred during the same period. After analyzing the impact of world global events on the mood of microblog posts, they found that the mood level in posts was correlated with cultural, political, and other events.

In another study, O'Connor et al. [2010] investigated the relation between opinion expressed in tweets and the public opinion obtained by polls. To this end, authors retrieved relevant tweets to some specific topics and then estimated the sentiment score of every day. They used a simple lexicon-based approach and the MPQA sentiment lexicon [Wiebe et al. 2006] to assign sentiment score to tweets. Sentiment time series were then produced by smoothing the daily positive versus negative ratio with a moving average window of the past k days. They found that there was a strong correlation between the smoothed time series and the polling data on costumer confidence and political opinion.

Bifet and Frank [2010] proposed a data-stream mining approach that could follow the changes of class distributions and allowed sentiment analysis in real time. The proposed approach could monitor the evolution of the impact of words on class predictions. To learn a linear classifier, Bifet and Frank used the stochastic gradient descent (SGD) method that had a similar performance with MNB.

Finally, focusing on visual sentiment analysis, Hao et al. [2011] explored three different approaches on a large volume of tweets. Hao et al. proposed a topic-based text-stream analysis method to determine the topic of discussion based on a number of opinionated attributes. Pixel cell-based sentiment calendars and high-density geomaps visualization techniques are then used to visualize a large number of tweets and facilitate data exploration.

4.3. Irony Detection on Tweets

Irony is a communication phenomenon that has been well studied in linguistics, psychology, and cognitive science [Gibbs 1986]. Irony is a way of communicating the opposite of the literal meaning and therefore can cause communicational

misunderstandings. Humans can easily detect irony. However, in terms of text mining, automatic *irony detection* is very difficult and has many challenges [Pang and Lee 2008]. In SA, the recognition of irony is very important, given the fact that it may flip the polarity of the sentiment of a message [Pang and Lee 2008].

In the field of irony detection, a few studies have focused on tweets. Reyes et al. [2013] presented one of the first studies for irony detection on tweets. In that study, Reyes et al. [2013] analyzed irony in terms of a multidimensional model of textual elements. They collected a corpus of 40,000 tweets using the hashtags #irony, #education, #humor, and #politics. Their model based on four different types of features (signatures, unexpectedness, style, and emotional scenarios), managed to obtain an F1-score of around 70% in distinguishing tweets having the hashtag #irony from those having the hashtags #education, #humor, or #politics.

The corpus built by Reyes et al. [2013] was also used by Barbieri and Saggion [2014]. For their study, they created three different datasets (irony vs education, irony vs humor, and irony vs politics). They built a model based on seven different types of features (frequency, written-spoken style, intensity, structure, sentiments, synonyms, and ambiguity). Their results showed that frequency, structure and synonyms were the most important features for detecting irony in all three datasets.

Liebrecht et al. [2013] used the hashtag #sarcasme to create a dataset of 78,000 tweets in Dutch. A balanced Winnow algorithm was used to classify the tweets as sarcastic or not with 75% accuracy. They found that some strong clues for detecting irony in tweets were the existence of the words irony, sarcasm, and cynicism; positive words; and intensifiers in the respective tweets. They also claimed that it was more likely to detect irony in tweets with a positive polarity.

Maynard and Greenwood [2014] investigated the use of sarcasm in tweets and studied its impact on TSA. Hashtags were used to determine if a tweet was sarcastic or not. One of the conclusions of the study was that detecting sarcastic tweets could improve sentiment detection by nearly 50 percentage points. However, the authors concluded that even when a tweet was correctly identified as being sarcastic, accuracy of sentiment analysis was still far from perfect.

4.4. Emotion Detection on Tweets

Another problem that is related to TSA is *emotion detection*. The difference between sentiment and emotion is that sentiment reflects a feeling, whereas emotion reflects an attitude [Tsytsarau and Palpanas 2012]. According to Plutchik [1980], there are eight basic emotions: anger, joy, sadness, fear, trust, surprise, disgust, and anticipation. Emotion detection aims at identifying various emotions from text. Considering the abundance of opinions and emotions expressed in microblogs, emotion detection in Twitter has attracted the interest of the research community. Here, we present some of the studies focused on emotion detection on tweets.

Mohammad [2012] proposed to consider hashtags that show an emotion (e.g., #anger, #surprise) for emotion detection. After creating a corpus that could be used for emotion detection, Mohammad [2012] conducted experiments that showed that the self-labeled hashtag annotations were consistent and matched with the annotations of the trained judges. Also, he created an emotion lexicon that could be used as available source of information when detecting emotions in text.

Roberts et al. [2012] used the list of the six Ekman's basic emotions (joy, anger, fear, sadness, surprise, disgust) proposed in Ekman [1992]. Roberts et al. [2012] extended the original list with an additional emotion: love. In their study, they created a series of binary SVM emotion classifiers achieving F-measures ranging from 0.642% (for anger) to 0.740% (for fear).

A semi-supervised learning method for emotion recognition in tweets was presented by Sintsova et al. [2014]. Based on a general purpose emotion lexicon, Sintsova et al. [2014] constructed the Balanced Weighted Voting classifier to correctly detect domain-specific emotional tweets. The classifier was evaluated on tweets about sports. The experimental results showed that the Balanced Weighted Voting managed to outperform their baseline based on NB.

Janssens et al. [2014] investigated the impact of using weak labels compared to strong labels on emotion recognition. The weakly annotated label set was created employing the hashtags of the tweets, while the strong label set was created by the use of crowdsourcing. Both label sets were used separately as input for five classification algorithms (Stochastic Gradient Descent, SVM, MNB, Nearest Centroid, Ridge) to determine the classification performance of the weak labels. The results indicated 9.25% decrease in F1-score when using weak labels.

4.5. Tweet Sentiment Quantification

Tweet sentiment quantification has recently attracted attention and this is reflected by the fact that is included as a new task in SemEval-2016 evaluation. Unlike sentiment analysis that aims to classify individual tweets, tweet sentiment quantification estimates the distribution of tweets across the different classes. Tweet sentiment quantification can be viewed as a different task from the one of sentiment classification and needs to be evaluated with different measures [Gao and Sebastiani 2015].

All these differences between classification and quantification are discussed by Gao and Sebastiani [2015]. Gao and Sebastiani aimed to differentiate between sentiment classification and sentiment quantification, and they argued that the latter is more appropriate when the goal is to estimate the class prevalence. In their study, they performed a series of experiments on various tweet collections and showed that quantification-specific algorithms outperform, at prevalence estimation, state-of-the-art classification algorithms.

Amati et al. [2014] modified Hopkins and Kings approach to estimate the sentiment distribution towards different topics. They proposed to use a features that are learned during the training phase. These features composed the sentiment dictionary. Their experiments showed that their proposed approach can be effectively applied for real time sentiment estimation.

Table IX summarizes the articles that focused on fields related to TSA. The objective of a study (second column) can be any of the *Twitter-Based Opinion Retrieval* (TOR), *Tracking Sentiments over Time* (TST), *Irony Detection* (ID), *Emotion Detection* (ED), and *Tweet Sentiment Quantification* (TSQ).

5. RESEARCH RESOURCES

In this section, we present and analyze the resources that are commonly used for TSA. We present sentiment lexicons developed for microblogs and datasets used for TSA. Also, we discuss the process of crawling tweets for creating new datasets and possible annotation processes.

5.1. Sentiment Lexicons

Building sentiment lexicons is closely related to the task of sentiment analysis. The sentiment lexicons contain a list of words annotated by their sentiment. Two of the most well-known and used lexicons are the SentiWordNet [Baccianella et al. 2010] and MPQA [Wiebe et al. 2006] lexicons. Early works in TSA used those lexicons as part of their methods [Barbosa and Feng 2010; O'Connor et al. 2010]. However, because the language used in microblogs differs considerably from the one that is used in other text genres [Baeza-Yates and Rello 2011], researchers proposed the construction of new and

Table IX. Summary of the Articles Focused on Fields Related to TSA

Study	Task	Algorithms	Features	Dataset
Luo et al. [2013a]	TOR	learning to rank/RankSVM	BM25/VSM score, tweet-specific, MPQA, TwitterSenti etc.	own
Luo et al. [2013b]	TOR	learning to rank/SVM light	retweetability, opinionatedness and textual quality	Luo et al. [2013a]
Giachanou et al. [2016]	TOR	unsupervised/topic specific variations	stylistic features, AFINN	Luo et al. [2013a]
An et al. [2014]	TST	supervised/NB, SVM		own
Bollen and Pepe [2011]	TST	lexicon-based/simple scoring	lexicon	own
O'Connor et al. [2010]	TST	lexicon-based/simple scoring	MPQA lexicon	OC
Bifet and Frank [2010]	TST	Stochastic Gradient Descent	unigrams	ETC
Hao et al. [2011]	TST	Pixel Sentiment Calendar, Pixel Sentiment Geo Map	Density distribution, negativity, influence	own
Reyes et al. [2013]	ID	supervised/NB, decision tree	signatures, degrees of unexpectedness, stylistic features, emotional scenarios	own
Barbieri and Saggion [2014]	ID	supervised/decision tree	frequency, written-spoken style, intensity, structure, sentiments (SentiWordNet), synonyms	Reyes et al. [2013]
Liebrecht et al. [2013]	ID	supervised/Balanced Winnow	unigrams, bigrams, trigrams	own
Maynard and Greenwood [2014]	ID	hashtag tokenization	hashtags	own
Mohammad [2012]	ED	supervised/SVM with Sequential Minimal Optimization	unigrams, bigrams	own
Roberts et al. [2012]	ED	supervised/SVM	unigram, bigrams, trigrams, punctuation, WordNet lexicon, topic scores, significant words	own
Sintsova et al. [2014]	ED	semi-supervised/Balanced Weighted Voting, NB, PMI-based	unigrams, bigrams, N-grams	own
Janssens et al. [2014]	ED	supervised/Stochastic Gradient Descent, SVM, MNB, Nearest Centroid, Ridge	N-grams, TF-IDF	own
Gao and Sebastiani [2015]	TSQ	SVM(KLD), SVM(HL)	same as Kiritchenko et al. [2014]	SemEval-2013-2015, Sanders, SS-Tweet, OMD, HCR, GASP
Amati et al. [2014]	TSQ	modification of Hopkins and Kings approach, linear regression	terms	Tweets2011, own

more specialized lexicons. Inspired by Affective Norms for English Words [Bradley and Lang 1999], Nielsen [2011] proposed a new sentiment lexicon, known as the AFINN (after the author's name Finn Årup Nielsen) lexicon, which was built for microblogs. The AFINN lexicon contains acronyms and slang words such as *lol* and *yolo*. Each of

the 2,477 English words is manually annotated with a score that ranges from −5 to +5. The score indicates the sentiment strength of the term, with −5 being most negative and +5 being most positive.

A graph-based approach was proposed by Tai and Kao [2013] for creating a sentiment lexicon for Twitter. Tai and Kao [2013] applied a graph-based semi-supervised label propagation method to assign polarities to words. They constructed graphs using words as nodes and used the semantic similarity between two words to weight edges. They conducted their experiments on about 650,000 tweets that were crawled using two keywords: *bullish* and *bearish*. Their experimental results indicated that their proposed method of automatically generating sentiment lexicon on Twitter is more effective compared to the general purpose lexicons.

A graph-based method for building the sentiment lexicon was also considered by Khuc et al. [2012]. Their approach was based on tweets containing emoticons. After normalizing the tweets, they used a POS tagger to extract nouns, adjectives, adverbs, verbs, interjections, abbreviations, hashtags, and emoticons from tweets. Co-occurrences between words were mapped on a graph on which the sentiment scores were propagated from seed nodes to other nodes. The new lexicon was shown to have a good quality and to be effective for TSA.

Minocha and Singh [2012] also focused in constructing sentiment lexicon from Twitter data. Instead of using a graph-based approach, they used an ontology tree generated by the 16 categories of the Open Directory Project. Their evaluation showed that the proposed domain specific lexicon is effective for TSA.

5.2. Datasets

This section discusses topics that are related to evaluation datasets for TSA. First, we briefly describe the process of creating new Twitter datasets. Then, we describe the existing datasets and different annotation processes for annotating tweets by sentiment.

5.2.1. Crawling Your Own Data. Twitter provides an easy way for researchers and developers to access and collect data via the two Twitter APIs: REST and Streaming.[15] REST API provides short-lived connections that are rate limited and someone can request and download a certain amount of tweets. Considering that Twitter does not give access to tweets that are older than a week, that means that REST API access is limited to tweets posted a week before at most. On the other side, Streaming API supports long-lived connections via different HTTP endpoints and downloads data almost in real time. That means that it can receive the latest tweets that contain specific terms or those that were posted by a specific user.

The Twitter APIs provide an easy way to collect a large amount of tweets that have specific characteristics such as tweets containing specific terms or emoticons, posted by a specific user or from a specific location. The tweets are returned in JSON format,[16] a widely used format for storing and exchanging data, and, therefore, they can be easily parsed by many programming languages. The metadata returned by the Twitter APIs include information such as publication date, author's username, location, hashtags, retweets, followers, and many other data.

A large number of researchers use the Twitter APIs to crawl tweets. The majority of the researchers prefer the Streaming API because it provides unlimited and real-time access to tweets that meet a specific requirement. Usually researchers use lists of emoticons, entities [Zhang et al. 2011] or hashtags to crawl tweets.

[15]https://dev.twitter.com/overview/documentation.
[16]http://www.json.org/.

Although a large amount of tweets can be crawled and downloaded via Twitter APIs, using those data for scientific research is challenging. One of the main problems is caused by the change of terms of service introduced in March 2011. According to that, Twitter does not allow the redistribution of collected tweets and this is also the case for annotated tweets datasets. Researchers are only allowed to share tweet IDs instead of actual data. However, since users may delete or privatize their tweets, the annotated datasets become partly inaccessible over time. This has the effect that when the percentage of inaccessible tweets is high, then the dataset cannot be used anymore since the results of different methods are not comparable anymore.

5.2.2. Evaluation Datasets. Due to the increasing popularity of TSA, several evaluation datasets have been built. The datasets usually contain a number of tweets crawled from Twitter together with tweets' sentiment. *Positive*, *negative*, and *neutral* are the most common labels [Nakov et al. 2013]. However, there are datasets that contain annotations only for positive and negative tweets [Go et al. 2009]. Additional labels for annotating tweets include *mixed*, *irrelevant*, or *other* [Speriosu et al. 2011]. Annotating tweets with a sentiment strength was considered by Thelwall et al. [2012]. Other evaluation datasets contain sentiment labels about the different aspects/entities appearing in the tweets [Speriosu et al. 2011]. A very interesting overview of eight publicly available datasets for TSA was presented by Saif et al. [2013]. In the following, we present and briefly describe the available evaluation datasets developed for TSA.

Edinburgh Twitter Corpus (ETC): One of the most well-known datasets for TSA is the Edinburgh Twitter Corpus [Petrović et al. 2010], which spans from the November 11, 2009, to February 1, 2010. The 96 million tweets of the collection were collected through Twitter's streaming API and is a representative set of the tweets posted during that period. A number of researchers have used this dataset, including Kouloumpis et al. [2011]. This dataset has been also used to address other tasks applied on Twitter such as microblog search [Naveed et al. 2011] and topic modeling [Zhao et al. 2011].

Stanford Twitter Sentiment (STS): The STS corpus[17] was introduced by Go et al. [2009] and provides both training and testing sets. The tweets were collected between April 6 and June 25, 2009, on the condition to contain at least one emoticon. The collected tweets were then automatically annotated as positive or negative based on the their emoticons. According to their annotation process, a tweet is considered positive if it contains emoticons such as :), :-), :), :D, or =), whereas a tweet is negative if it contains :(, :-(, or : (. This process resulted in a training set of 1.6 million annotated tweets. The testing set has 182 positive and 177 negative tweets that were manually annotated. The STS corpus has been extensively used for tasks such as subjectivity classification and TSA. For example, Bravo-Marquez et al. [2013], used this dataset to evaluate their method on the subjectivity classification task while other researchers have used this dataset for evaluating their methods on sentiment analysis [Go et al. 2009; Saif et al. 2012a; Speriosu et al. 2011; Bakliwal et al. 2012].

Sanders Corpus (Sanders): The Sanders[18] dataset consists of 5,513 manually annotated tweets with respect to four different targets: *Apple*, *Google*, *Microsoft*, *Twitter*. Each tweet was annotated as positive, negative, neutral, or irrelevant given its topic, resulting in 570 positive, 654 negative, 2,505 neutral, and 1,786 irrelevant tweets. A number of researchers have used the dataset for subjectivity and TSA [Bravo-Marquez et al. 2013; Liu et al. 2012; Deitrick and Hu 2013].

[17]http://help.sentiment140.com/.
[18]http://www.sananalytics.com/lab/twitter-sentiment/.

O'Connor's Corpus (OC): The O'Connor's Corpus contains 1 billion tweets crawled during 2008 and 2009. The initial dataset was not annotated by sentiment. Researchers have used hashtags and emoticons as noisy labels to annotate the dataset by sentiment and to use it for TSA [O'Connor et al. 2010; Davidov et al. 2010].

Health Care Reform (HCR): The HCR dataset[19] was presented by Speriosu et al. [2011] and consists of 2,515 manually annotated tweets of which 541 are positive, 1,381 negative, 470 neutral, 79 irrelevant, and 44 unsure tweets. The tweets were collected in March 2010 and are about the topic *hcr*. The collected tweets were manually annotated for polarity regarding to one out of the following different targets: *health care reform*, *Obama*, *Democrats*, *Republicans*, *Tea Party*, *conservatives*, *liberals*, and *Stupak*. The tweets were then split into three different sets, the training, the development, and the testing, each consisting of about 840 tweets. The dataset has been used for both subjectivity and TSA [Saif et al. 2012b; Speriosu et al. 2011].

Obama-McCain Debate (OMD): The OMD dataset[20] was built from tweets collected during the presidential debate in September 2008 [Shamma et al. 2009]. The 3,238 collected tweets were manually labelled as positive, negative, mixed, or other by at least three and up to eight annotators using the Amazon Mechanical Turk. This dataset has been used by many researchers for TSA [Hu et al. 2013a, 2013b; Saif et al. 2012b; Speriosu et al. 2011].

Sentiment Strength Twitter Dataset (SS-Tweet): The Sentiment Strength Twitter Dataset[21] was constructed by Thelwall et al. [2012] with the aim to evaluate SentiStrength.[22] The dataset contains 4,242 manually annotated tweets. Tweets that express a negative sentiment were labeled with a number between −1 (not negative) and −5 (extremely negative), whereas tweets that express positive sentiment with a number between 1 (not positive) and 5 (extremely positive).

SemEval Datasets: SemEval (Semantic Evaluation) is an ongoing series of evaluations of computational semantic analysis systems that has evolved from the SensEval word sense disambiguation evaluation series. The task of sentiment analysis was introduced for the first time in SemEval 2013. The first collection released by SemEval is the *SemEval-2013* dataset [Nakov et al. 2013], which consists of tweets that contain named entities and words that are also present in SentiWordNet. Sentiment labels were assigned to tweets using Amazon Mechanical Turk, resulting in 15,196 labeled tweets. SemEval-2013 consists of 5,810 positive, 6,979 objective/neutral, and 2,407 negative tweets and has been used to evaluate methods for tweet-level [Mohammad et al. 2013; Martínez-Cámara et al. 2013; Remus 2013] and expression-level subjectivity detection [Mohammad et al. 2013; Chalothorn and Ellman 2013; Kökciyan et al. 2013]. *SemEval-2014* is an extension of the SemEval-2013 collection and was built as a continuation of the TSA task. This collection contains 982 positive, 202 negative, and 669 neutral tweets in addition to SemEval-2013 tweets. A similar annotation process followed the SemEval-2013 [Rosenthal et al. 2014] process. In 2015, the *SemEval-2015* collection was released as an extension of the SemEval-2013 and SemEval-2014 collections. SemEval-2015 contains 1899 positive, 1008 negative, and 190 neutral additional tweets [Rosenthal et al. 2015]. *SemEval-2016* is also an extension of the past SemEval collections. In addition to TSA, this collection contains data for some new tasks, including tweet quantification that aims to estimate the distribution of tweets

[19]https://bitbucket.org/speriosu/updown/.
[20]https://bitbucket.org/speriosu/updown/.
[21]http://sentistrength.wlv.ac.uk/documentation/.
[22]http://sentistrength.wlv.ac.uk/.

across the different classes instead of classifying individual tweets [Gao and Sebastiani 2015].

STS-Gold: STS-Gold was constructed in Saif et al. [2013] as a subset of the STS dataset [Go et al. 2009]. The dataset contains 2,034 annotated tweets of which 632 are positive and 1,402 are negative. Also, the dataset contains 58 manually annotated entities annotated by three different human annotators. The dataset has been used for TSA [Saif et al. 2014, 2016].

Dialogue Earth Twitter Corpus (DETC): This corpus was constructed as part of the Dialogue Earth Project[23] and contains three datasets (WA, WB, and GASP). The WA and the WB sets contain tweets about weather, whereas the third set contains tweets about gas prices. The three sets contain 4,490, 8,850, and 12,770 tweets, respectively, which were manually annotated with sentiment labels. Each tweet was asigned with one of the following labels: positive, negative, neutral, not related, or cannot tell. The corpus has been used to evaluate methods for sentiment analysis on twitter data [Asiaee et al. 2012].

One interesting observation is the tendency of researchers to crawl their own data and evaluate their methods on them. This is a consequence of the lack of benchmarks in the field and the difficulty in building one. The majority of the approaches use a number of words or phrases to crawl tweets [Speriosu et al. 2011; Nakov et al. 2013], whereas others crawl a sample of the posted tweets regardless their topic [Petrović et al. 2010]. Crawling tweets using a set of keywords has the advantage that the topic of the tweet is known *a priori*. However, one limitation is that there is high chance tweets on the same topic are not crawled due to word mismatch. In addition, careful annotation is needed to detect the tweets that express sentiment about the specific target. On the contrary, crawling a sample of the posted tweets can lead to a more representative collection. However, the collection will have a lot of noise and filtering is required. We would like to note that the decision of the crawling approach is mainly based on the addressed task. For example, for entity-based TSA using a set of keywords to crawl tweets is an adequate approach, whereas tasks such as TST or burst detection require a sample of posted tweets that span over time.

5.2.3. Annotation. One of the main challenges in evaluating approaches that address Twitter-based sentiment analysis is the absence of benchmark datasets. In the literature, a large number of researchers have used the Twitter API to crawl tweets and create their own datasets, whereas others evaluate their methods on collections that were created by previously reported studies. One major challenge in creating new datasets is how the tweets should be annotated. There are two approaches that have been followed for annotating the tweets according to their polarity: *manual annotation* and *distant supervision*.

One of the most popular platforms used for manual annotation of tweets is the *Amazon Mechanical Turk* platform[24] which is a crowdsourcing service used to coordinate the use of human intelligence for tasks that computers are currently unable to do. This approach has been used for manual annotation of several datasets [Go et al. 2009; Shamma et al. 2009; Nakov et al. 2013]. Due to the fact that it is a costly process, it has been usually applied only for the annotation of the test set [Go et al. 2009]. This approach can support the annotation of several hundreds to few thousands of tweets. The quality of the annotations is measured by calculating the kappa coefficient of inter-annotator agreement.

[23]www.dialogueearth.org.
[24]https://www.mturk.com/mturk/welcome.

Another popular annotation approach is *distant supervision*, also known as indirect crowdsourcing. This approach is very common for creating training sets that require large number of annotated data. Emoji, emoticons, and hashtags are usually employed as noisy sentiment labels. Go et al. [2009] used tweets with emoticons to annotate the training tweet set. Tweets with positive emoticons (i.e., :), :-), and :D) were considered positive, whereas those containing negative emoticons (i.e., :(and :-() were considered as negative. Pak and Paroubek [2010] also considered emoticons as noisy labels to annotate the tweets used for training. Hashtags have been also extensively used as indicators of positive or negative sentiment. Davidov et al. [2010] considered emoticons and hashtags for creating the training set. They used 50 hashtags that were indicative of positive (#happy, #love, etc.) or negative (#hate, #bored, etc.) sentiment. A slightly different approach was followed by Barbosa and Feng [2010], who used Twitter sentiment classification web sites (e.g., *Twendz*, *Twitter Sentiment*, and *TweetFeel*) to collect their data.

Datasets annotated using distant supervision usually contain several thousand or even hundreds thousand of annotated tweets. This amount of tweets is much larger than the one that can be collected using the manual annotation. Even though distant supervision can be used to collect a large number of annotations, the annotations are not always accurate. This is also supported by Liu and Zhang [2012], who reported that manually labeled data can lead to superior results.

6. OPEN ISSUES

After analyzing the literature of TSA, we can observe that it is still an open domain for research and there are issues that require further exploration. The most important open issues are the following: *use of deep learning*, *lack of benchmarks*, *data sparsity*, *multilingual content*, *tracking sentiments over time*, and *multidisciplinary research*.

Use of deep learning: One of the most important limitations of machine-learning approaches is that their effectiveness depends on the set of selected features. However, typical feature selection approaches are not able to address some phenomena, such as negation detection. One solution to that is using algorithms that are capable of learning representations of the data, known as deep-learning algorithms. The effectiveness of these algorithms on TSA is still under-explored since only a few researchers have considered this direction [Tang et al. 2014a, 2014b; Vo and Zhang 2015; Dong et al. 2014]. Using deep-learning algorithms to address TSA is an area that needs to be better explored. One interesting direction would be to examine the effectiveness of the recursive neural network deep-learning algorithms on TSA and negation handling in tweets, since they are effective on SA of standard text.

Lack of benchmarks: One of the most important problems in this domain is the lack of benchmark datasets. It has been noticed that researchers collect their own data and evaluate their methods on them. The comparison of different methods is very difficult or even impossible. One exception to this is the SemEval collections. However, the SemEval datasets contain a few thousand annotated tweets stressing the difficulty iof creating large collections. In addition, Twitter does not allow the redistribution of tweets, a restriction that makes the construction of benchmark datasets more challenging. Lack of benchmark datasets is also observed in fields that are related to TSA such as TOR for which there are only two attempts to create annotated datasets [Luo et al. 2013a; Paltoglou and Buckley 2013]. Even the creation of benchmarked collections on these fields is very challenging, but it is not impossible. The contribution of different research groups where each is responsible for annotating a percentage of the data can be a solution to build larger collections.

Data sparsity: Data sparsity occurs to a large extent in Twitter due to the large amount of informal textual peculiarities. Dealing with data sparsity is very important, since it can influence the performance of TSA. Reducing data sparsity was explored by Saif et al. [2012a]. However, considering the importance of minimizing data sparsity, we believe that this problem needs to be further investigated.

Multilingual content: Tweets are written in a wide variety of languages and sometimes more than one language is used in the same tweet. However, only few researchers have tried to address multilingual TSA [Narr et al. 2012]. Also, a few datasets have been built for this problem. We believe that addressing multilingual TSA is an interesting and important field of research that needs to be further investigated.

Tracking sentiments over time: Detecting sentiment towards a topic and tracking its evolution over time is a field that has received a little attention. Tracking sentiment towards a specific topic and identifying sentiment changes is very important for various applications. For example, companies can track the sentiment towards their products and act promptly in case of negative sentiment emergence. Topic models that jointly combine topic and sentiment detection of tweets and time could be proposed to deal with this problem. Similar approaches have been already proposed for standard text and therefore it would be interesting to explore if they can be applied on Twitter.

Multidisciplinary research: The combination of research from different fields is still under-explored. Applying sentiment analysis methods on economics research or on human and social science domains can yield interesting results. For example, it is possible to explore how geographic places or meteorological variables and events influence the level of happiness within a society [Mitchell et al. 2013; Curini et al. 2014]. Additionally, sentiment analysis could be applied on a marketing domain to predict the success of a product or of movies [Asur and Huberman 2010]. Another interesting direction would be to apply sentiment analysis on health domains to explore how emotions correlate with well being or with health in general.

Finally, we should note that recently Twitter announced that it may remove the length limitation of 140 characters. This is an important change and will influence different aspects of TSA. However, the extent of this impact depends on how much the language of tweets will change. For example, currently there is a great use of abbreviations, which may be significantly reduced after the length restriction is lifted. Using whole words instead of abbreviations may facilitate research on TSA. On the other hand, we believe that users will continue to use some characteristics of the language, such as emoticons and slang language, that are mostly related to the informal communication style of Twitter. Another possible change will be the number of topics discussed in a single tweet. Currently, the majority of tweets are about a single topic due to the length limitation [Giachanou and Crestani 2016]. That implies that if a tweet contains opinion, then the opinion is towards the specific topic that is discussed. In case users start writing longer tweets, then additional features such as proximity or more sophisticated topic extraction methods will be necessary for effective sentiment analysis.

7. CONCLUSIONS

Recent years have witnessed an increasing research interest in analyzing tweets according to the sentiment they express. This interest is a result of the large amount of messages that are posted everyday in Twitter andthat contain valuable information for the public mood for a number of different topics. Methods from the machine-learning field are applied for TSA on a more frequent rate compared to the rest of the approaches, with the SVM and NB classifiers being the most prevalent. Unigram-based

SVM is usually considered as a baseline to which the proposed methods are compared. In addition, lexicons are utilized in a large set of proposed methods to support detecting words that indicate sentiment. SentiWordNet and MPQA are the most used lexicons that are usually extended with words that are used in Twitter.

This survey presented an overview on the recent updates of TSA. More than 50 articles were categorized and briefly described. After the analysis, it is clear that TSA is still an open field for research. We reviewed the most prominent approaches for TSA and discussed prevalent methods for Twitter-based opinion retrieval, tracking sentiments over time, irony detection, emotion detection, and tweet sentiment quantification. In addition, we presented different resources (datasets and sentiment lexicons) that have been used in TSA. Our survey gave a comprehensive review of the proposed TSA methods, discussed related trends, identified interesting open problems, and indicated promising future research directions.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their valuable comments.

REFERENCES

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media (LSM'11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 30–38.

Fotis Aisopos, George Papadakis, and Theodora Varvarigou. 2011. Sentiment analysis of social media content using n-gram graphs. In *Proceedings of the 3rd ACM SIGMM International Workshop on Social Media (WSM'11)*. ACM, New York, NY, 9–14. DOI:http://dx.doi.org/10.1145/2072609.2072614

Giambattista Amati, Marco Bianchi, and Giuseppe Marcone. 2014. Sentiment estimation on twitter. In *Proceedings of the 5th Italian Information Retrieval Workshop (IIR'14)*, Vol. 1127. CEUR Workshop Proceedings, 39–50.

Xiaoran An, R. Auroop Ganguly, Yi Fang, B. Steven Scyphers, M. Ann Hunter, and G. Jennifer Dy. 2014. Tracking climate change opinions from twitter data. In *Proceedings of the Workshop on Data Science for Social Good Held in Conjunction with KDD 2014*. ACM, New York, NY.

T. Amir Asiaee, Mariano Tepper, Arindam Banerjee, and Guillermo Sapiro. 2012. If you are happy and you know it... tweet. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM, New York, NY, 1602–1606. DOI:http://dx.doi.org/10.1145/2396761.2398481

Nathan Aston, Jacob Liddle, and Wei Hu. 2014. Twitter sentiment in data streams with perceptron. *J. Comput. Commun.* 2 (2014), 11–16. DOI:http://dx.doi.org/10.4236/jcc.2014.23002

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT'10)*. IEEE Computer Society, Washington, DC, 492–499. DOI:http://dx.doi.org/10.1109/WI-IAT.2010.63

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Language Resources and Evaluation Conference (LREC'10)*. European Language Resources Association (ELRA), 2200–2204.

Ricardo Baeza-Yates and Luz Rello. 2011. How bad do you spell? The lexical quality of social media. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media: The Future of the Social Web Workshop (ICWSM'11)*. AAAI Press.

Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. 2012. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'12)*. Association for Computational Linguistics, Stroudsburg, PA, 11–18.

Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in twitter: Feature analysis and evaluation. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14)*. European Language Resources Association (ELRA), 4258–4264.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)*. Association for Computational Linguistics, Stroudsburg, PA, 36–44.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828.

Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: Is brevity an advantage? In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. ACM, New York, NY, 1833–1836. DOI:http://dx.doi.org/10.1145/1871437.1871741

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th International Conference on Discovery Science (DS'10)*. Springer-Verlag, Berlin, 1–15.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *J. Comput. Sci.* 2, 1 (2011), 1–8. DOI:http://dx.doi.org/10.1016/j.jocs.2010.12.007

Johan Bollen and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*. AAAI Press, 450–453.

Nibir Nayan Bora. 2012. Summarizing public opinions in tweets. *Int. Jo. Comput. Ling. Appl.* 3, 1 (2012), 41–55.

Margaret M. Bradley and Peter J. Lang. 1999. *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Technical Report. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. 2013. Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In *Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'13)*. ACM, New York, NY, 2:1–2:9. DOI:http://dx.doi.org/10.1145/2502069.2502071

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: Using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 562–570.

Tawunrat Chalothorn and Jeremy Ellman. 2013. Tjp: Using twitter to analyze the polarity of contexts. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 375–379.

Sam Clark and Richard Wicentowski. 2013. SwatCS: Combining simple classifiers with estimated accuracy. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 425–429.

Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: Bridging the gap among multilingual twitter sentiment analysis. In *Proceedings of the 7th Asia Conference on Information Retrieval Technology (AIRS'11)*. Springer-Verlag, Berlin, 238–249. DOI:http://dx.doi.org/10.1007/978-3-642-25631-8_22

Luigi Curini, Stefano Iacus, and Luciano Canova. 2014. Measuring idiosyncratic happiness through the analysis of twitter: An application to the Italian case. *Social Indicators Research* 121, 2 (2014), 525–542. DOI:http://dx.doi.org/10.1007/s11205-014-0646-2

Nádia F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Supp. Syst.* 66 (July 2014), 170–179. DOI:http://dx.doi.org/10.1016/j.dss.2014.07.003

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING'10)*. Association for Computational Linguistics, Stroudsburg, PA, 241–249.

William Deitrick and Wei Hu. 2013. Mutually enhancing community detection and sentiment analysis on twitter networks. *J. Data Anal. Inform. Process.* 1, 3 (2013), 19–29.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*. ACM, New York, NY, 231–240. DOI:http://dx.doi.org/10.1145/1341531.1341561

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Stroudsburg, PA, 49–54.

Paul Ekman. 1992. An argument for basic emotions. *Cogn. Emotion* 6, 3/4 (1992), 169–200. DOI:http://dx.doi.org/10.1080/02699939208411068

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Language Resources and Evaluation Conference (LREC'06)*. European Language Resources Association (ELRA), 417–422. DOI:http://dx.doi.org/10.1.1.61.7217

Song Feng, Ritwik Bose, and Yejin Choi. 2011. Learning general connotation of words using graph-based algorithms. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*. Association for Computational Linguistics, Stroudsburg, PA, 1092–1103.

Wei Gao and Fabrizio Sebastiani. 2015. Tweet sentiment: From classification to quantification. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM'15)*. ACM, New York, NY, 97–104. DOI:http://dx.doi.org/10.1145/2808797.2809327

Manoochehr Ghiassi, James Skinner, and David Zimbra. 2013. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Syst. Appl.* 40, 16 (2013), 6266–6282. DOI:http://dx.doi.org/10.1016/j.eswa.2013.05.057

Anastasia Giachanou and Fabio Crestani. 2016. Opinion retrieval in twitter: Is proximity effective? In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC'16)*. ACM, New York, NY.

Anastasia Giachanou, Morgan Harvey, and Fabio Crestani. 2016. Topic-specific stylistic variations for opinion retrieval on twitter. In *Proceedings of the 38th European Conference on Advances in Information Retrieval (ECIR'16)*. Springer International Publishing, Berlin, 466–478. DOI:http://dx.doi.org/10.1007/978-3-319-30671-1_34

Raymond W. Gibbs. 1986. On the psycholinguistics of sarcasm. *J. Exper. Psychol. Gen.* 115 (1986), 3–15. Issue 1. DOI:http://dx.doi.org/10.1037/0096-3445.115.1.3

Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification Using Distant Supervision*. Technical Report. Standford.

Hussam Hamdan, Frederic Béchet, and Patrice Bellot. 2013. Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*, Vol. 2. Association for Computational Linguistics, 455–459.

Ming Hao, Christian Rohrdantz, Halldór Janetzko, Umeshwar Dayal, Daniel A. Keim, Lars-Erik Haug, and Mei-Chun Hsu. 2011. Visual sentiment analysis on twitter data streams. In *Proceedings of the 2011 IEEE Symposium on Visual Analytics Science and Technology (VAST'11)*. IEEE, 277–278.

Ammar Hassan, Ahmed Abbasi, and Daniel Zeng. 2013. Twitter sentiment analysis: A bootstrap ensemble framework. In *Proceedings of the International Conference on Social Computing (SocialCom'13)*. IEEE Computer Society, 357–364. DOI:http://dx.doi.org/10.1109/SocialCom.2013.56

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013a. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. ACM, New York, NY, 607–618. DOI:http://dx.doi.org/10.1145/2488388.2488442

Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013b. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*. ACM, New York, NY, 537–546. DOI:http://dx.doi.org/10.1145/2433396.2433465

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. Association for Computational Linguistics, Stroudsburg, PA, 720–728.

Olivier Janssens, Steven Verstockt, Erik Mannens, Sofie Van Hoecke, and Rik Van De Walle. 2014. Influence of weak labels for emotion recognition of tweets. In *Proceedings of the 2nd International Conference in Mining Intelligence and Knowledge Exploration (MIKE 2014)*. Springer, Berlin, 108–118. DOI:http://dx.doi.org/10.1007/978-3-319-13817-6_12

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT'11)*. Association for Computational Linguistics, Stroudsburg, PA, 151–160.

Farhan Hassan Khan, Saba Bashir, and Usman Qamar. 2014. TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Supp. Syst.* 57 (Jan. 2014), 245–257. DOI:http://dx.doi.org/10.1016/j.dss.2013.09.004

Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, and Jay Ramanathan. 2012. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC'12)*. ACM, New York, NY, 459. DOI:http://dx.doi.org/10.1145/2245276.2245364

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA, Article 1367. DOI:http://dx.doi.org/10.3115/1220355.1220555

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *J. Artif. Intell. Res.* 50 (2014), 723–762.

Nadin Kökciyan, Arda Çelebi, Arzucan Özgür, and Suzan Üsküdarl. 2013. BOUNCE: Sentiment classification in twitter using rich feature sets. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 554–561.

Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, and Nick Bassiliades. 2013. Ontology-based sentiment analysis of twitter posts. *Expert Syst. Appl.* 40, 10 (2013), 4065–4074.

Peter Korenek and Marián Šimko. 2014. Sentiment analysis on microblog utilizing appraisal theory. *World Wide Web* 17, 4 (July 2014), 847–867. DOI:http://dx.doi.org/10.1007/s11280-013-0247-z

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*. AAAI Press, 538–541.

Akshi Kumar and Teeja Mary Sebastian. 2012. Sentiment analysis on twitter. *Int. J. Comput. Sci. Issues* 9, 4 (2012), 372–378.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA'13)*. Association for Computational Linguistics, Stroudsburg, PA, 29–37.

Jimmy Lin and Alek Kolcz. 2012. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data (SIGMOD'12)*. ACM, New York, NY, 793–804. DOI:http://dx.doi.org/10.1145/2213836.2213958

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5, 1 (May 2012), 1–167. DOI:http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining Text Data*. Springer, New York, NY, 415–463.

Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. 2012. Emoticon smoothed language models for twitter sentiment analysis. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence Emoticon*. AAAI Press, Palo Alto, CA, 1678–1684.

Zhunchen Luo, Miles Osborne, and Ting Wang. 2013a. An effective approach to tweets opinion retrieval. *World Wide Web* 18, 3 (2013), 545–566. DOI:http://dx.doi.org/10.1007/s11280-013-0268-7

Zhunchen Luo, Jintao Tang, and Ting Wang. 2013b. Propagated opinion retrieval in twitter. In *Proceedings of the 14th International Conference on Web Information Systems Engineering (WISE'13)*. Springer, Berlin, 16–28. DOI:http://dx.doi.org/10.1007/978-3-642-41154-0_2

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT'11)*. Association for Computational Linguistics, Stroudsburg, PA, 142–150.

Eugenio Martínez-Cámara, Teresa M. Martín-Valdivia, Alfonso L. Ureña López, and Arturo Montejo-Ráez. 2012. Sentiment analysis in twitter. *Nat, Lang, Eng,* 20, 01 (Nov. 2012), 1–28. DOI:http://dx.doi.org/10.1017/S1351324912000332

Eugenio Martínez-Cámara, Arturo Montejo-Ráez, Teresa M. Martín-Valdivia, and Alfonso L. Ureña López. 2013. Sinai: Machine learning and emotion of the crowd for sentiment analysis in microblogs. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14)*. European Language Resources Association (ELRA), 4238–4243.

Akshay Minocha and Navjyoti Singh. 2012. Generating domain specific sentiment lexicons using the web directory. *Adv, Comput,: Int, J,* 3, 5 (2012), 45–51.

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PloS One* 8, 5 (2013), e64417.

Saif M. Mohammad. 2012. #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval'12)*. Association for Computational Linguistics, 246–255.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 321–327.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 312–320.

Sascha Narr, Hulfenhaus Michael, and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. In *Proceedings of the Knowledge Discovery and Machine Learning at LWA 2012 (KDML'12)*.

Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Searching microblogs: Coping with sparsity and document quality. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. ACM, New York, NY, 183–188.

Finn Å Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis of microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Volume 718 (ESWC'11)*. CEUR Workshop Proceedings, 93–98.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. AAAI Press.

Reynier Ortega, Adrian Fonseca, and Andres Montoyo. 2013. SSA-UO: Unsupervised twitter sentiment analysis. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, 501–507.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of Conference of the North American Chapter of the Association of Computational Linguistics on Human Language Technologies (NAACL-HLT 2013)*. The Association for Computational Linguistics, 380–390.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th on International Language Resources and Evaluation Conference (LREC'10)*. European Language Resources Association (ELRA), 1320–1326.

Georgios Paltoglou and Kevan Buckley. 2013. Subjectivity annotation of the microblog 2011 realtime adhoc relevance judgments. In *Proceedings of the 35th European Conference on Advances in Information Retrieval (ECIR'13)*. Springer-Verlag, Berlin, 344–355. DOI: http://dx.doi.org/10.1007/978-3-642-36973-5_29

Georgios Paltoglou and Anastasia Giachanou. 2014. Opinion retrieval: Searching for opinions in social media. In *Professional Search in the Modern World: COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*. Springer International Publishing, CBerlin, 193–214. DOI: http://dx.doi.org/10.1007/978-3-319-12511-4_10

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135. DOI: http://dx.doi.org/10.1561/15000000011

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP'02)*. Association for Computational Linguistics, Stroudsburg, PA, 79–86. DOI: http://dx.doi.org/10.3115/1118693.1118704

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media (WSA'10)*. Association for Computational Linguistics, Stroudsburg, PA, 25–26.

Robert Plutchik. 1980. Emotion: Theory, research, and experience: Vol. 1. theories of emotion. In *Approaches to Emotion*, R. Plutchik and H. Kellerman (Eds.). Academic Press, New York, NY, 3–33.

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

Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. 2013. Rule-based detection of sentiment phrases using SAS sentiment analysis. In *2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: 7th International Workshop on Semantic Evaluation*, Vol. 2. Association for Computational Linguistics, 513–519.

Robert Remus. 2013. ASVUniOfLeipzig: Sentiment analysis in twitter using data-driven machine learning techniques. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Confer-*

ence on Lexical and Computational Semantics (SemEval'13). Association for Computational Linguistics, 450–454.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.* 47, 1 (March 2013), 239–268. DOI: http://dx.doi.org/10.1007/s10579-012-9196-x

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie, and Sanda M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Language Resources and Evaluation Conference (LREC'12)*. European Language Resources Association (ELRA), 3806–3813.

Carlos Rodríguez-Penagos, Jordi Atserias, Joan Codina-Filbà, David García-narbona, Jens Grivolla, Patrik Lambert, and Roser Saurí. 2013. FBM: Combining lexicon-based ML and heuristics for social media polarities. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*, Vol. 2. Association for Computational Linguistics, 483–489.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. Association for Computational Linguistics, 451–463.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. SemEval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*. Association for Computational Linguistics and Dublin City University.

Hassan Saif, Miriam Fernández, and Harith Alani. 2014. Automatic stopword generation using contextual semantics for sentiment analysis of twitter. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track at the 13th International Semantic Web Conference (ISWC'14)*. CEUR-WS.org, 281–284.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the STS-gold. In *Proceedings of the 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM'13)*. CEUR-WS.org.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14)*. European Language Resources Association (ELRA), 810–817.

Hassan Saif, Yulan He, and Harith Alani. 2012a. Alleviating data sparsity for twitter sentiment analysis. In *Workshop on Making Sense of Microposts (#MSM2012): Big Things Come in Small Packages at the 21st International Conference on the World Wide Web (WWW'12)*. CEUR-WS.org, 2–9.

Hassan Saif, Yulan He, and Harith Alani. 2012b. Semantic sentiment analysis of twitter. In *Proceedings of the 11th International Conference on The Semantic Web - Volume Part I (ISWC'12)*. Springer-Verlag, Berlin, 508–524. DOI: http://dx.doi.org/10.1007/978-3-642-35176-1_32

Hassan Saif, Yulan He, Miriam Fernández, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. *Inform. Process. Manag.* 52, 1 (2016), 5–19. DOI: http://dx.doi.org/10.1016/j.ipm.2015.01.005

David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the First SIGMM Workshop on Social Media (WSM'09)*. ACM, New York, NY, 3–10. DOI: http://dx.doi.org/10.1145/1631144.1631148

Valentina Sintsova, Claudiu Musat, and Pearl Pu. 2014. Semi-supervised method for multi-category emotion recognition in tweets. In *2014 IEEE International Conference on Data Mining Workshop (ICDM'14)*. IEEE, 393–402. DOI: http://dx.doi.org/10.1109/ICDMW.2014.146

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP (EMNLP'11)*. Association for Computational Linguistics, Stroudsburg, PA, 53–63.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Comput. Ling.* 37, 2 (2011), 267–307.

Yen-Jen Tai and Hung-Yu Kao. 2013. Automatic domain-specific sentiment lexicon generation with label propagation. In *Proceedings of the International Conference on Information Integration and Web-Based Applications & Services (IIWAS'13)*. ACM, New York, NY, 53–62. DOI: http://dx.doi.org/10.1145/2539150.2539190

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM, New York, NY, 1397–1405. DOI: http://dx.doi.org/10.1145/2020408.2020614

Duyu Tang, Bing Qin, and Ting Liu. 2015a. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL'15)*. The Association for Computer Linguistics, 1014–1023.

Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015b. User modeling with neural network for review rating prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 1340–1346.

Duyu Tang, Furu Wei, Bing Qin, Ting Liu, and Ming Zhou. 2014a. Coooolll: A deep learning system for twitter sentiment classification. In *Proceedings of the 8th International Workshop on Semantic Evaluation*. Association for Computational Linguistics and Dublin City University, 208–212.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014b. Learning sentiment-specific word embedding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*. The Association for Computer Linguistics, 1555–1565.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *J. Am. Soc. Inform. Sci. Technol.* 63, 1 (2012), 163–173.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *J. Am. Soc. Inform. Sci. Technol.* 61, 12 (2010), 2544–2558.

Mikalai Tsytsarau and Themis Palpanas. 2012. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.* 24, 3 (2012), 478–514.

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

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*. AAAI Press, 1347–1353.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. 2011. Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM'11)*. ACM, New York, NY, 1031–1040. DOI:http://dx.doi.org/10.1145/2063576.2063726

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2006. Annotating expressions of opinions and emotions in language. *Lang. Res. Eval.* 39, 2–3 (2006), 165–210. DOI:http://dx.doi.org/10.1007/s10579-005-7880-9

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. 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, Stroudsburg, PA, 347–354. DOI:http://dx.doi.org/10.3115/1220575.1220619

Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. 2011. *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis*. Technical Report.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Springer-Verlag, Berlin, 338–349. DOI:http://dx.doi.org/10.1007/978-3-642-20161-5_34

Received July 2015; revised April 2016; accepted April 2016