Sentiment Analysis and Trend Detection in Twitter

María del Pilar Salas-Zárate[1(✉)], José Medina-Moreira[2,3], Paul Javier Álvarez-Sagubay[3], Katty Lagos-Ortiz[2,3], Mario Andrés Paredes-Valverde[1], and Rafael Valencia-García[1]

[1] Departamento de Informática y Sistemas, Universidad de Murcia, 30100 Murcia, Spain
{mariapilar.salas,marioandres.paredes,valencia}@um.es

[2] Universidad Agraria del Ecuador, Avenida 25 de Julio, Guayaquil, Ecuador
{jmedina,klagos}@uagraria.edu.ec

[3] Universidad de Guayaquil Cdla. Universitaria Salvador Allende, Guayaquil, Ecuador
{paul.alvarezs,jose.medinamo,katty.lagoso}@ug.edu.ec

Abstract. Social networks such as Twitter are considered a rich resource of information about actual world actions of all types. Several efforts have been dedicated to trend detection on Twitter i.e., the current popular topics of conversation among its users. However, despite these efforts, sentiment analysis is not taken into account. Sentiment analysis is the field of study that analyzes people's opinions and moods. Therefore, applying sentiment analysis to tweets related to a trending topic also enables to know if people are talking positively or negatively about it, thus providing important information for real-time decision making in various domains. On the basis of this understanding, we propose SentiTrend, a system for trend detection on twitter and its corresponding sentiment analysis. In this paper, we present the SentiTrend's architecture and functionality. Also, the evaluation results concerning the effectiveness of our approach to trend detection and sentiment analysis are presented. Our proposal obtained encouraging results with an average F-measure of 80.7 % for sentiment classification, and an average F-measure 80.0 % and 75.5 % for trend detection.

Keywords: Twitter · Social media analysis · Sentiment analysis · Trend detection

1 Introduction

The messages posted in social networks provide a solid background about the ideas and opinions not only about the users of the social networks but also about the environment where they live. This information can be used and consumed by a wide range of institutions and organisms for strategic decision making.

Nowadays, Twitter is one of the most popular online social networking and microblogging services that enables its users to send and read text-based posts of up to 140 characters, known as tweets. Millions of users use Twitter to keep in touch with friends, meet new people and discuss about everything [1]. Companies are increasingly using Twitter to advertise and recommend products, brands, and services; to build and maintain reputations; to analyze users' sentiment regarding their products or those of their competitors; to respond to customers' complaints; and to improve decision making and business intelligence [2].

R. Valencia-García et al. (Eds.): CITI 2016, CCIS 658, pp. 63–76, 2016.
DOI: 10.1007/978-3-319-48024-4_6

Several pieces of research have been conducted in recent years in order to automatically process the information on social networks [3–6]. An outstanding issue which provides research opportunities is trending topic detection. In the context of Twitter, trending topics represent the popular "topics of conversation", among its users [7]. Monitoring and analyzing this rich and continuous flow of user-generated content can yield valuable information. However, most works about trend detection fail to take sentiment into consideration. Sentiment analysis gives an effective and efficient means to expose public opinion timely which gives vital information for decision making in various domains.

In this work, we propose an approach, known as SentiTrend, to trending topic detection on Twitter and its subsequent polarity detection. SentiTrend collects messages from Twitter and processes them in order to determine their trending topic based on the TF-IDF (Term Frequency–Inverse Document Frequency) model. Then, an estimated positive, negative or neutral sentiment value is assigned to each tweet related to the trending topic detected. The task of assigning a sentiment value to a tweet is done using a free software from Stanford University, known as Stanford Classifier. The Stanford Classifier is a Java-based implementation of a maximum entropy classifier, which takes data and applies probabilistic classification [8].

On the other hand, it is worth mentioning that studies exclusively deal with the English language, perhaps owing to the lack of resources in other languages. Considering that the Spanish language has a much more complex syntax than many other languages, and that it is the third most widely spoken language in the world, we firmly believe that the computerization of Internet domains in this language is of utmost importance. For this reason, this work is mainly motivated in the Spanish language.

This paper is structured as follows: Sect. 2 presents the state of the art on sentiment analysis and trend detection on social networks. Section 3 presents the architecture and functionality of our proposal. Section 4 shows a set of experiments carried out to validate the proposed approach concerning the effectiveness of our approach to trend detection and sentiment analysis. Finally, Sect. 5 describes our conclusions and future work.

2 Related Work

2.1 Sentiment Analysis

In recent years, several researchers have introduced methods for sentiment classification. Most of these efforts are based on two approaches: the semantic orientation approach and the machine learning approach. Both approaches have their advantages and drawbacks. The semantic orientation approach is based on opinion words, namely, words that are commonly used in expressing positive or negative sentiment. Opinion words are typically contained in a dictionary called opinion lexicon.

For example, Ghosh & Animesh [9] presented a rule-based method that can be used to identify the sentiment polarity of opinion sentences. They use SentiWordNet to calculate the overall sentiment score of each sentence. The results obtained in this work indicate that SentiWordNet could be used as an important resource for sentiment classification tasks. Peñalver-Martínez et al. [10], meanwhile, propose an innovative opinion mining methodology that takes advantage of new semantic Web-guided solutions to enhance the results obtained with traditional natural language processing techniques, sentiment analysis processes and Semantic Web technologies. Their proposal is specifically based on three different stages: (1) an ontology-based mechanism for feature identification, (2) a SentiWordNet-based technique to assign a polarity to each feature, and (3) a new approach for opinion mining based on vector analysis. Montejo-Ráez et al. [11] presented an unsupervised approach for polarity classification in Twitter. They integrated SentiWordNet to compute the final value of polarity. The synsets values are weighted with the PageRank scores obtained in the random walk process over WordNet.

However, tweets are not considered ‘‘normal’’ pieces of text, since the 140-character threshold imposes limitations in the length. A further peculiarity of the tweets is the extensive usage of jargon expressions, abbreviations, and emoticons. A disadvantage is the fact that jargon expressions are often domain dependent. These factors lead to a low recall when the lexicon-based method is applied on informal corpora of text, like posts from micro-blogs [3].

An alternative approach is the application of machine learning techniques. This approach is based on using a collection of data to train classifiers. The drawback of the machine learning-based methods is mainly focused on the manual labeling required over massive sets of tweets. However, several pieces of research showed that the machine learning approach outperforms the semantic orientation approach [12].

For example, Mohammad et al. [13] propose a basic automatic system to classify tweets and determine who is feeling certain emotion, and towards whom. They trained a Support Vector Machine (SVM) classifier for that. Sidorov et al. [14], meanwhile, examine how classifiers work while carrying out opinion mining of Spanish Twitter data. They explore how different settings (n-gram size, corpus size, the number of sentiment classes, balanced vs. unbalanced corpus, various domains) affect the precision of the machine learning algorithms and experiment with Naïve Bayes, Decision Tree, and Support Vector Machines. Some other works [15, 16] combine NLP (Natural Language Processing) and machine learning techniques in order to increase the effectiveness of their method. Salas-Zárate et al. [15] present a method that uses a hybrid feature extraction method based on POS (part-of-speech) pattern and dependency parsing. The features obtained are enriched semantically through common sense knowledge bases. Then, a feature selection method is applied to eliminate the noisy and irrelevant features. Finally, a set of classifiers is trained in order to classify unknown data. Habernal et al. [16] present in-depth research on supervised machine learning methods for sentiment analysis of Czech social media. They explore different pre-processing techniques and employ various features and classifiers. The authors also experiment with five different feature selection algorithms and investigate the influence of named entity recognition and preprocessing on sentiment classification performance.

Finally, some other more recent proposals are based on psycholinguistic tools for sentiment analysis such as LIWC [17, 18].

2.2 Trend Detection

There are several studies that have addressed trend detection on social networks. For example, [19] present a trend prediction method for news event or news topics on Twitter.

Experimental results show that the method is simple and effective. The authors also propose and analyze several possible reasons for the trend rising and falling of news topics on Twitter. Kaleel & Abhari [6] propose a system for event detection and trending from tweet clusters which are discovered using LSH (Locality Sensitive Hashing) technique. Specifically, the key issues addressed by the authors are: (1) construction of a dictionary using incremental term TF–IDF in high dimensional data to create tweet feature vector, (2) leveraging LSH to find truly interesting events, (3) trending the behavior of events based on time, geo-locations and cluster size, and (4) speed-up the cluster discovery process while retaining the cluster quality. Ding et al. [20], meanwhile, focus on automated personalization of tweets for popular trending topics. The main objective is to classify the tweets information as “Like” or “Dislike” on a particular topic based on personal preferences. Martinez-Romo & Araujo [1] present a methodology based on two main aspects: the detection of spam tweets in isolation and without previous information of the user; and the application of a statistical analysis of language to detect spam in trending topics. Mathioudakis & Koudas [21] present TwitterMonitor, a system that performs trend detection over the Twitter stream. The system identifies emerging topics on Twitter in real time and provides meaningful analytics that synthesize an accurate description of each topic. Users interact with the system by ordering the identified trends using different criteria and submitting their own description for each trend.

We should state that our work differs from the existing works for two reasons: (1) Our approach is based on the Spanish language in comparison to most studies that exclusively deal with the English language, and (2) Our approach obtains an estimated positive, negative or neutral sentiment value for each tweet related to the trending topic detected.

3 SentiTrend Architecture

SentiTrend consists of a Back-End and a Front-End layer (see Fig. 1). On the one hand, The Back-End is divided into five main components: (1) Retrieve module, (2) Data extraction and verification module, (3) Pre-processing module, (4) Trend detection module, and (5) Sentiment analysis module. On the other hand, the Front-End consists

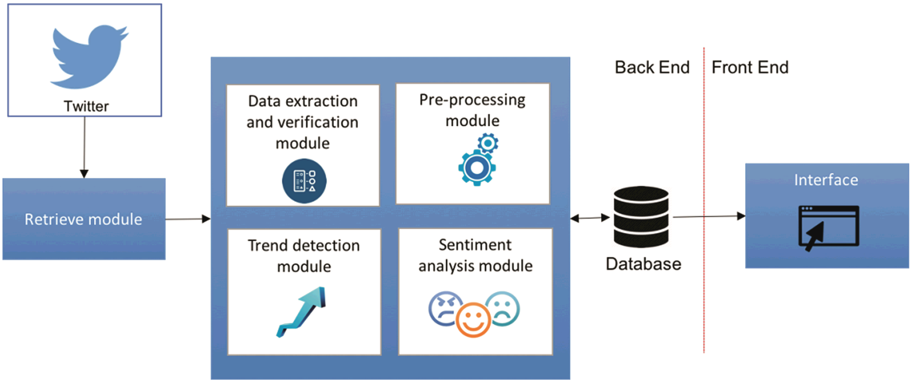


Fig. 1. SentiTrend’s architecture.

of a web application developed in Java that allows users to view the trending topics detected, and the tweets corresponding to the selected trending topic including the percentage of positive, negative or neutral tweets.

3.1 Back-End Layer

As was previously mentioned, the Back-End layer is divided into five main components:

1. Retrieve module. This module is responsible for establishing, maintaining a connection with Twitter and retrieving tweets.
2. Data extraction and verification module. It avoids storing repetitive tweets and extracts valuable information from the tweets such as user, text, followers, etc.
3. Pre-processing module: This module carries out the data cleansing of each tweet by means of NLP techniques. In other words, it removes from the tweets information such as hyperlinks, emoticons, among others.
4. Trend detection module. It performs the trend detection based on a TF-IDF model.
5. Sentiment Analysis module. This module classifies tweets as positive, negative, or neutral.

A detailed description of the modules contained in the architecture shown above is provided in the following sections.

Retrieve module. This module handles establishing and maintaining the connection with Twitter servers to retrieve tweets. We use Twitter4 J, a Java library that gives access to the Twitter API and assists in integrating the Twitter service into any Java application. In order to obtain useful results, we have established two search filters: (1) Track, and (2) Locations. The first filter consists of a comma-separated list of phrases which will be used to determine which tweets will be delivered on the stream. The second filter consists of a comma-separated list of longitude, latitude pairs specifying a set of bounding boxes to filter Tweets. Each bounding box should be specified as a pair of longitude and latitude pairs, with the southwest corner of the bounding box coming first. For example, to obtain the tweets from Spain, we need the following coordinates:

```
upper right point:
Latitude:43.834527
Length:1.423828
lower left point:
Latitude:36.119713
Length: -9.47461
```

Data extraction and verification module. In this module, information about each tweet is extracted. Also, a verification process is carried out. Next, a detailed description of the process performed is presented.

1. It obtains a tweet of the tail of tweets.
2. This module retrieves the tweet information namely, id, date, number of retweets, text, language, the user who wrote it, number of the user's followers and hashtags and users that appear in the tweet.

3. It verifies if a tweet is original or a retweet
   - (a) If a tweet is original, it verifies if it exists in the database with its id
     - (i) If the tweet is not in the database, a sentiment classification is carried out ("sentiment analysis module"), and all information is stored in the database.
     - (ii) Otherwise, data such as number of followers, number of retweets are updated in the database.
   - (b) If a tweet is a retweet, the original tweet is obtained, as well as its id, date, number of retweets, text, language, the user who wrote it, the user'sfollowers, the user name, hashtags, and users named in the tweet.
     - (i) If the original tweet is not in the database, a sentiment classification is carried out ("sentiment analysis module"), and all information is stored in the database.
     - (ii) Otherwise, data such as number of followers and number of retweets are updated in the database.

Pre-processing module. The pre-processing module carries out the data cleansing of each tweet by means of NLP techniques [22, 23].

The system carries out the following steps before extracting features from the text of the tweet.

- Slang words translation: Tweets often contain slang words. Slang word translation means converting the slang words like lol, omg, among others, into their standard form.
- Tokenization: The sentences are divided into words or tokens by removing white spaces and other symbols or special characters.
- Case Normalization: The process is to turn the entire tweet into lowercase.
- Stemming: It is the process of reducing all the remaining words to their respective stems. It is worth remarking that stemming finds the stem, and not the root of the words.
- The removal of Stop Words: A stop word is defined as a word that contains no meaning or relevance in and of itself. All words that appeared as the most frequent in at least 80 % were classified as stop words. If a word was identified as a stop word, it was removed.
- Identify presence of URL using a regular expression ("https?://\\S+\\s") and remove all the URLs from the tweet.
- Remove all the private usernames identified by @user and the symbol # of hashtags.

Trend detection module. This module performs trend detection through two main phases. Firstly, a set of simple and composite features are extracted. This process is performed by using n-grams (like unigrams, bigrams and trigrams) [24]. For example, the features obtained from the sentence "big bang theory" are the following:

```
unigrams: "big", "bang", "theory".
bigrams: "big bang", "bang theory".
trigrams: "big bang theory".
```

Secondly, in order to calculate the weight of the words, a TF.IDF model is used. TF-IDF is a statistical measure that is used to estimate the importance of a word in a document or in a collection of documents [6, 25]. Having said that, term frequency can be defined as:

$$tf_{ij} = \frac{n_{ij}}{\mathrm{N}} \tag{1}$$

where n_{ij} is the number of times word i occurs in document j and N is the total number of words in document j.

$$N = \sum_{k} n_{kj} \tag{2}$$

The second definition is often referred to as the normalized term frequency. Inverse document frequency is defined as

$$idf_i = \log\left(\frac{D}{d_i}\right) \tag{3}$$

where d_i is the number of documents that contain word i and D is the total number of documents.

Therefore, the TF-IDF score for a word w in a document d is calculated by:

$$tf - idf = tf_{ij} * idf_i \tag{4}$$

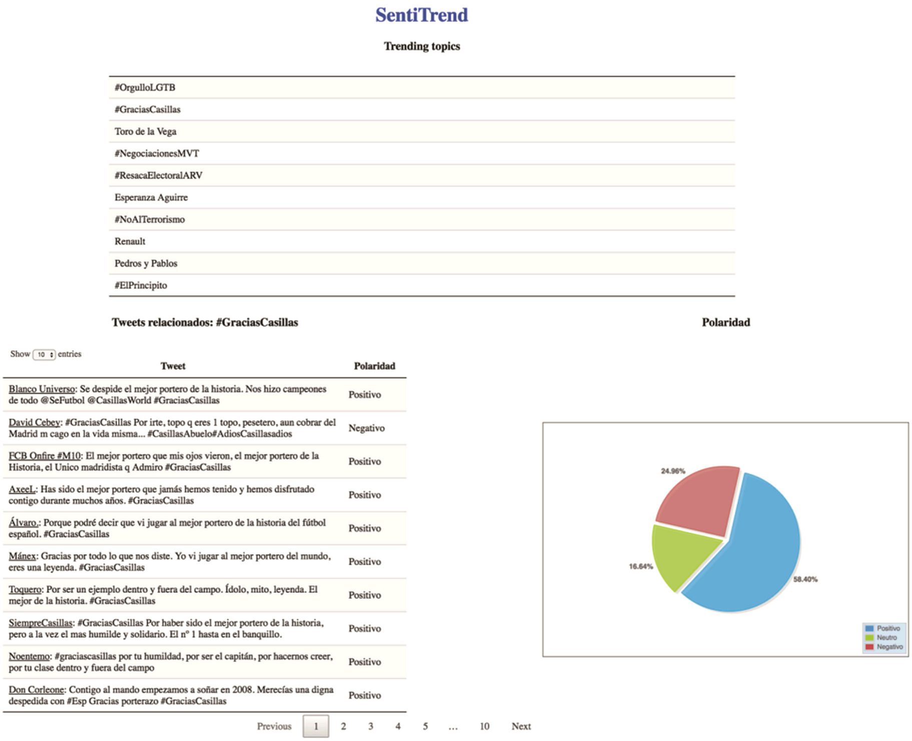
Sentiment analysis module. The last module provides the negative, positive or neutral polarity of the tweets. Aiming to perform such a task, this module needed the previous development of a Machine Leaning-based module able to determine the polarity of a tweet, i.e. to determine if a tweet is positive, negative or neutral concerning a topic. The development of this module involved two main phases. Firstly, a corpus consisting of 1000 positive tweets, 1000 negative tweets, and 1000 neutral tweets was obtained. We used the Twitter API to collect the tweets. After downloading the tweets, each tweet was individually processed as described in a "pre-processing module" section. Also, we performed a manual review of the filtered tweets in order to make sure that the obtained tweets are relevant to our study. Finally, each tweet was classified by hand in order to ensure the quality of the corpus. This time-consuming task was performed in a period of 12 months by a group of five people with a great experience in the sentiment classification domain. We do not share this dataset the Twitter policy does not all us to share tweets contents.

Secondly, the corpus mentioned above was used to training a classifier, more specifically, we use the Stanford classifier, a Java-based implementation of a maximum entropy classifier, which takes data and applies probabilistic classification. We applied a Machine Learning (ML) approach as it has been applied in several works, achieving great results for sentiment classification. The Machine learning methods often rely on supervised classification approaches. This approach is based on using a collection of data to train the classifiers. Among the machine learning techniques commonly used in sentiment polarity classification we find Support Vector Machine (SVM) [26, 27], Naive Bayes (NB) [28, 29], and Maximum Entropy (MaxEnt) [30].

3.2 Front-End Web Application

SentiTrend provides a web interface where users can carry out the following tasks: (1) view the recent trends in real-time, (2) view tweets about a selected trend, as well as, the percentage and total of positive, negative, and neutral tweets.

Next, Fig. 2 shows the SentiTrend Web application.



SentiTrend

Trending topics

#OrgulloLGTB
#GraciasCasillas
Toro de la Vega
#NegociacionesMVT
#ResacaElectoralARV
Esperanza Aguirre
#NoAlTerrorismo
Renault
Pedros y Pablos
#ElPrincipito

Tweets relacionados: #GraciasCasillas

Show 10 entries

Tweet	Polaridad
Blanco Universo: Se despide el mejor portero de la historia. Nos hizo campeones de todo @SeFutbol @CasillasWorld #GraciasCasillas	Positivo
David Cebey: #GraciasCasillas Por irte, topo q eres 1 topo, pesetero, aun cobrar del Madrid m cago en la vida misma... #CasillasAbuelo#AdiosCasillasadios	Negativo
FCB Onfire #M10: El mejor portero que mis ojos vieron, el mejor portero de la Historia, el Unico madridista q Admiro #GraciasCasillas	Positivo
AxeeL: Has sido el mejor portero que jamás hemos tenido y hemos disfrutado contigo durante muchos años. #GraciasCasillas	Positivo
Álvaro.: Porque podré decir que vi jugar al mejor portero de la historia del fútbol español. #GraciasCasillas	Positivo
Mánex: Gracias por todo lo que nos diste. Yo vi jugar al mejor portero del mundo, eres una leyenda. #GraciasCasillas	Positivo
Toquero: Por ser un ejemplo dentro y fuera del campo. Ídolo, mito, leyenda. El mejor de la historia. #GraciasCasillas	Positivo
SiempreCasillas: #GraciasCasillas Por haber sido el mejor portero de la historia, pero a la vez el mas humilde y solidario. El nº 1 hasta en el banquillo.	Positivo
Noentemo: #graciascasillas por tu humildad, por ser el capitán, por hacernos creer, por tu clase dentro y fuera del campo	Positivo
Don Corleone: Contigo al mando empezamos a soñar en 2008. Merecías una digna despedida con #Esp Gracias porterazo #GraciasCasillas	Positivo

Fig. 2. SentiTrend Web application.

4 Evaluation and Results

In order to evaluate the effectiveness of the system for sentiment classification and trend detection, we have used three evaluation measurements: precision, recall and F-measure. Recall (5) is the proportion of factual positive cases that were correctly predicted as such. On the other hand, precision (6) represents the proportion of predicted positive cases that are actually positive. Finally, F-measure (7) is the harmonic mean of precision and recall [31].

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

where TP is the number of true positives and FN is the number of false negatives.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where TP is the number of true positives and FP is the number of false positives.

$$\mathrm{F1} = 2 * \frac{\mathrm{Precision}{*}\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{7}$$

The experiments carried out in this study are described in detail below.

4.1 Trend Detection

In order to evaluate the effectiveness of the system for trend detection, several experiments were carried out. The experiments involve obtaining results for different time intervals with our system (SentiTrend), Twitter API, and Trends24, and then, carry out

Table 1. Trend detection results obtained by SentiTrend.

Test	SentiTrend-Twitter			SentiTrend-Trends24		
	P	R	F	P	R	F
1	0.7	0.7	0.7	0.7	0.7	0.7
2	0.8	0.8	0.8	0.8	0.8	0.8
3	0.8	0.8	0.8	0.6	0.6	0.6
4	0.7	0.7	0.7	0.7	0.7	0.7
5	0.9	0.9	0.9	0.9	0.9	0.9
6	0.6	0.6	0.6	0.6	0.6	0.6
7	0.8	0.8	0.8	0.7	0.7	0.7
8	0.9	0.9	0.9	0.8	0.8	0.8
9	0.8	0.8	0.8	0.8	0.8	0.8
10	0.9	0.9	0.9	0.9	0.9	0.9
11	0.7	0.7	0.7	0.7	0.7	0.7
12	0.9	0.9	0.9	0.7	0.7	0.7
13	0.9	0.9	0.9	0.7	0.7	0.7
14	0.8	0.8	0.8	0.8	0.8	0.8
15	0.9	0.9	0.9	0.9	0.9	0.9
16	0.9	0.9	0.9	0.9	0.9	0.9
17	0.8	0.8	0.8	0.8	0.8	0.8
18	0.8	0.8	0.8	0.8	0.8	0.8
19	0.7	0.7	0.7	0.6	0.6	0.6
20	0.7	0.7	0.7	0.7	0.7	0.7
AVG	**0.8**	**0.8**	**0.8**	**0.755**	**0.755**	**0.755**

a comparison between the results obtained by the aforementioned tools. The experiments were evaluated using precision, recall, and F-measure metric. Aiming to calculate the corresponding scores, the following facts are considered.

- True positives: the items that were identified as trending topic by SentiTrend, Twitter API and Trends24.
- False positives: the items identified as trending topic by SentiTrend that were not identified as trending topic by Twitter API and Trends24.
- False negatives: the items identified as trending topics by Twitter API and Trends24 that were not identified as trending topics by SentiTrend.

Table 1 shows precision (P), recall (R), and F-measure (F) results obtained by SentiTrend and Twitter API tools and SentiTrend and Trends24 tools for different time intervals.

As can be seen in Table 1, SentiTrend obtains good results for trend detection with an average precision, recall, and F-measure of 80.0 % with regards to Twitter API, and 75.5 % with respect to Trends24. In fact, the best result (precision, recall, and F-measure of 90 %) was obtained by several tests for both SentiTrend-Twitter and SentiTrend-Trends24. Also, the results show that SentiTrend obtains more matches of trending topics with Twitter than with Trends24.

Table 2. Sentiment classification results

	Precision	Recall	F-measure
1	0.800	0.800	0.799
2	0.800	0.800	0.799
3	0.817	0.817	0.817
4	0.813	0.814	0.813
5	0.810	0.810	0.810
6	0.801	0.800	0.800
7	0.807	0.807	0.806
8	0.830	0.830	0.830
9	0.774	0.773	0.773
10	0.789	0.780	0.782
11	0.832	0.803	0.813
12	0.817	0.817	0.817
13	0.807	0.807	0.807
14	0.816	0.817	0.816
15	0.794	0.777	0.780
16	0.846	0.843	0.844
17	0.800	0.800	0.800
18	0.844	0.843	0.843
19	0.766	0.763	0.764
20	0.833	0.830	0.830
AVG	**0.810**	**0.807**	**0.807**

4.2 Sentiment Classification

The experiments of sentiment analysis were carried out on a set of tweets related to a trending topic. For this purpose, the trending topic with the highest score obtained by SentiTrend for each of the twenty case studies presented in the previous section (see Sect. 4.1) was selected. Then, 300 tweets related to the trending topic were collected. Each of them was classified as positive, negative, or neutral by both, an expert group on sentiment analysis and the SentiTrend system.

Finally, a comparison of the results obtained by the aforementioned methods was carried out through precision, recall, and F-measure metrics. The evaluation results are shown in Table 2.

As can be seen in Table 2, the system provides encouraging results for sentiment classification of tweets in the Spanish language, with average Precision, Recall and F-measure values of 81 %, 80.7 %, and 80.7 %, respectively.

4.3 Discussion

General results show that the system successfully performs trend detection in Twitter and polarity detection.

With respect to trend detection, experiments show that the method is effective. However, much remains to be done about this topic. For example, we believe that analyzing fake content on twitter would be an interesting factor.

Regarding sentiment analysis, the system provides encouraging results. As mentioned above, we have used the Stanford Classifier to perform MaxEnt (Maximum Entropy) classification. The results obtained for the MaxEnt are very good. These results can be justified by the analysis presented in [32], where the authors mention that MaxEnt has been successfully employed for natural language processing tasks since the main advantages of MaxEnt are its robustness and statistic efficiency. However, it would be interesting to carry out several experiments with other classifiers such as SVM, BayesNet by using some machine learning tools, such as Weka [33] and RapidMiner [34], aiming to compare the results provided by several algorithms.

5 Conclusions and Future Work

In this work, we have proposed SentiTrend, a system for trend detection and sentiment analysis. We have also presented the experiments whose objective was to evaluate the proposed approach concerning trend detection and sentiment analysis. Our proposal yielded encouraging results, with an average F-measure of 80.7 % for sentiment analysis, and an average F-measure of 80 % and 75.5 % for trend detection with regards to Twitter API and Trends24, respectively.

In spite of all the advantages and possibilities of the proposed approach, it has several limitations that could be improved in future work. First, our approach is only able to deal with tweets expressed in Spanish, which is a disadvantage owing to the vast amount of information available in other languages. We shall therefore attempt to apply this approach to the English language. Second, our approach is not able to detect irony, sarcasm, and satire. These aspects can play the role of a polarity reverse, with respect to the words used in the tweet. This is one of the most interesting aspects to check in social media for sentiment analysis. We plan to integrate a module to detect irony, sarcasm, and satire. Third, in order to train and validate the sentiment analysis method, we collected a corpus, which was manually labeled by an expert group. However, we plan to use a standard/benchmark corpus in order to evaluate the effectiveness of our method and compare our results with other proposed works. Finally, another disadvantage of our proposal is that it is not able to identify spam users as well as spam tweets. Trending topics are a very effective method for tricking users into visiting malicious or spam websites. Accordingly, the attackers collect information regarding the most popular trending topics and include them in tweets pointing to spam websites. Therefore, we are interested in carry out a study regarding spam propagation through Twitter such as that presented in [35].

Acknowledgments. María del Pilar Salas-Zárate and Mario Andrés Paredes-Valverde are supported by the National Council of Science and Technology (CONACYT), the Public Education Secretary (SEP) and the Mexican government.

References

1. Martinez-Romo, J., Araujo, L.: Detecting malicious tweets in trending topics using a statistical analysis of language. Expert Syst. Appl. **40**(8), 2992–3000 (2013)
2. Atefeh, F., Khreich, W.: A survey of techniques for event detection in Twitter. Comput Intell. **31**(1), 132–164 (2015)
3. Kontopoulos, E., Berberidis, C., Dergiades, T., Bassiliades, N.: Ontology-based sentiment analysis of Twitter posts. Expert Syst. Appl. **40**(10), 4065–4074 (2013)
4. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in Twitter: a closer look. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2, pp. 581–586, Stroudsburg, PA, USA (2011)
5. Paltoglou, G., Thelwall, M.: Twitter, MySpace, Digg: unsupervised sentiment analysis in social media. ACM Trans Intell Syst Technol. **3**(4), 66 (2012)
6. Kaleel, S.B., Abhari, A.: Cluster-discovery of Twitter messages for event detection and trending. J. Comput. Sci. **6**, 47–57 (2015)
7. Benhardus, J., Kalita, J.: Streaming trend detection in Twitter. Int. J. Web Based Communities **9**(1), 122–139 (2013)
8. MacCartney, B.: Stanford Classifer, The Stanford Natural Language Processing Group (2015). http://nlp.stanford.edu/software/classifier.shtml. Accessed 18 May 2015
9. Ghosh, M., Animesh, K.: Unsupervised linguistic approach for sentiment classification from online reviews using SentiWordNet 3.0. Int. J. Eng. Res. Technol. **2**(9), 55–60 (2013)
10. Peñalver-Martinez, I., Garcia-Sanchez, F., Valencia-Garcia, R., Rodríguez-García, M.A., Moreno, V., Fraga, A., Sánchez-Cervantes, J.L.: Feature-based opinion mining through ontologies. Expert Syst. Appl. **41**(13), 5995–6008 (2014)
11. Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A.: A knowledge-based approach for polarity classification in Twitter. J. Assoc. Inf. Sci. Technol. **65**(2), 414–425 (2014)

12. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. Expert Syst. Appl. **36**(3), 6527–6535 (2009)
13. Mohammad, S.M., Zhu, X., Kiritchenko, S., Martin, J.: Sentiment, emotion, purpose, and style in electoral tweets. Inf. Process. Manag. **51**(4), 480–499 (2015)
14. Sidorov, G., et al.: Empirical study of machine learning based approach for opinion mining in tweets. In: Batyrshin, I., González Mendoza, M. (eds.) MICAI 2012, Part I. LNCS, vol. 7629, pp. 1–14. Springer, Heidelberg (2013)
15. Salas-Zárate, M.P., Paredes-Valverde, M.A., Limon-Romero, J., Tlapa, D., Baez-Lopez, Y.: Sentiment classification of Spanish reviews: an approach based on feature selection and machine learning methods. J. UCS **22**(5), 691–708 (2016)
16. Habernal, I., Ptáček, T., Steinberger, J.: Supervised sentiment analysis in Czech social media. Inf. Process. Manag. **50**(5), 693–707 (2014)
17. Balage Filho, P.P., Pardo, T.A., Alusio, S.M.: An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara, pp. 215–219 (2013)
18. Salas-Zárate, M.P., López-López, E., Valencia-García, R., Aussenac-Gilles, N., Almela, Á., Alor-Hernández, G.: A study on LIWC categories for opinion mining in Spanish reviews. J. Inf. Sci. **40**(6), 749–760 (2014)
19. Lu, R., Yang, Q.: Trend analysis of news topics on Twitter. Int. J. Mach. Learn. Comput. **2**(3), 327 (2012)
20. Ding, L., Pang, C., Kew, L.M., Jain, L.C., Howlett, R.J., Weilin, L., Hoon, G.K.: Personalization of trending tweets using like-dislike category model. Procedia Comput. Sci. **60**, 236–245 (2015)
21. Mathioudakis, M., Koudas, N.: TwitterMonitor: trend detection over the TwitterStream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, pp. 1155–1158. ACM, New York (2010)
22. Paredes-Valverde, M.A., Valencia-García, R., Rodríguez-García, M.A., Colomo-Palacios, R., Alor-Hernández, G.: A semantic-based approach for querying linked data using natural language. J. Inf. Sci. (2015) doi:10.1177/0165551515616311
23. Paredes-Valverde, M.A., Rodríguez-García, M.Á., Ruiz-Martínez, A., Valencia-García, R., Alor-Hernández, G.: ONLI: an ontology-based system for querying DBpedia using natural language paradigm. Expert Syst. Appl. **42**(12), 5163–5176 (2015)
24. Agarwal, B., Mittal, N.: Prominent feature extraction for review analysis: an empirical study. J. Exp. Theoret. Artif. Intell. **28**(3), 485–498 (2016)
25. Elshater, Y., Elgazzar, K., Martin, P.: goDiscovery: web service discovery made efficient. In: 2015 IEEE International Conference on Web Services (ICWS), pp. 711–716 (2015)
26. Rushdi Saleh, M., Martín-Valdivia, M.T., Montejo-Ráez, A., Ureña-López, L.A.: Experiments with SVM to classify opinions in different domains. Expert Syst. Appl. **38**(12), 14799–14804 (2011)
27. Moraes, R., Valiati, J.F., Gavião Neto, W.P.: Document-level sentiment classification: an empirical comparison between SVM and ANN. Expert Syst. Appl. **40**(2), 621–633 (2013)
28. Xia, R., Zong, C., Li, S.: Ensemble of feature sets and classification algorithms for sentiment classification. Inf. Sci. **181**(6), 1138–1152 (2011)
29. Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña-López, L.A.: Ranked WordNet graph for sentiment polarity classification in Twitter. Comput. Speech Lang. **28**(1), 93–107 (2014)
30. He, Y., Zhou, D.: Self-training from labeled features for sentiment analysis. Inf. Process. Manag. **47**(4), 606–616 (2011)

31. Salas-Zárate, M.P., Valencia-García, R., Ruiz-Martínez, A., Colomo-Palacios, R.: Feature-based opinion mining in financial news: Aan ontology-driven approach. J. Inf. Sci. (2016). doi:10.1177/0165551516645528
32. Shah, H., Bhandari, P., Mistry, K., Thakor, S., Patel, M., Ahir, K.: Study of named entity recognition for indian languages. Int. J. Inf. **6**(1), 11–25 (2016)
33. Bouckaert, R.R., Frank, E., Hall, M.A., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: WEKA − experiences with a Java open-source project. J. Mach. Learn. Res. **11**, 2533–2541 (2010)
34. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press, Boca Raton (2013)
35. Antonakaki, D., Polakis, I., Athanasopoulos, E., Ioannidis, S., Fragopoulou, P.: Exploiting abused trending topics to identify spam campaigns in Twitter. Soc. Netw. Anal. Min. **6**(1), 1–11 (2016)