Harvesting Opinions in Twitter for Sentiment Analysis

Juan Guevara[13], Joana Costa[12], Jorge Arroba[34], Catarina Silva[12]
[1]School of Technology and Management, Polytechnic Institute of Leiria, Portugal
[2]Center for Informatics and Systems of the University of Coimbra, Portugal
[3]Universidad Central del Ecuador, Ecuador
[4]University of Alicante, Spain
2162315@my.ipleiria.pt,{catarina, joana.costa}@ipleiria.pt, jarroba@uce.edu.ec

Abstract — **Sentiment analysis is a very popular technique for social network analysis. One of the most popular social networks for microblogging that has a great growth is Twitter, which allows people to express their opinions using short, simple sentences. These texts are generated daily, and for this reason it is common for people to want to know which are the trending topics and their drifts. In this paper we propose to deploy a mobile app that provides information focusing on areas, such as, Politics, Social, Tourism, and Marketing using a statistical lexicon approach. The application shows the polarity of each theme as positive, negative, or neutral.**

Keywords – Sentiment Analysis; Twitter; Lexicon.

I. INTRODUCTION

In recent years, the growing advances in information and communication technologies (ICT) has brought with it a wave of new applications on the Internet. The ones that have had the greatest impact on society have been social networks and microblogs such as Facebook, and Twitter, where billions of users share content and often give their opinion related to a topic in particular [1]. This has attracted the attention of private companies and public institutions that want to take advantage of this amount of users and information, to carry out market studies in different areas [2]. This has given rise to the development of a subarea in text mining, called sentiment analysis.

Sentiment analysis focuses on analyzing people's opinions, feelings, attitudes, and emotions with respect to a specific product, organization, service, movie, individual, politics, and other topics [3]. This type of analysis can be used by commercial and public organizations that, by monitoring social networks, carry out market studies taking advantage of users' comments related to particular products/policies/etc. The results of such studies are to help the organization to receive feedback and thus improve the quality of its products or services [4].

To analyze the opinions in the texts it is very common to rate them with a polarity and thus be able to classify the words or phrases among three categories: positive, negative or neutral [2], [5].

The techniques used for the sentiment classification can be divided into two groups: machine learning approaches and lexicon-based approaches [6], [7]. Although the precision of lexicon-based approaches sometimes cannot provide an accuracy comparable to machine learning approaches, the main advantage of lexicon-based approaches is that it is not required to have a large amount of data for training to be easily used in many domains. On the other hand, machine learning approaches usually require a representative training set and high computational costs. Consequently, if the goal is to develop mobile applications for sentiment analysis, lexicon-based approaches are appropriate, due to the limitations of mobile device processing and the energy consumption associated with such processing in mobile devices.

In this paper we present a mobile app which uses a statistical lexicon-based approach to sentiment analysis for Twitter comments. The app performs an analysis on positive and negative words in each tweet using a dictionary related to the specific domain, resulting in the proposal of a statistical method to qualify the polarity of the tweet, which it can be positive, negative, or neutral.

The rest of the paper is structured as follows. Section II presents the background on sentiment analysis on Twitter. Section III explores related work to Twitter data analysis for sentiment polarity. In Section IV we present the proposed approach including the defined architecture and, in Section V, the results are detailed. Finally, Section VI presents the conclusions and future lines of research.

II. BACKGROUND

This section presents fundamental concepts about sentimental analysis in Twitter.

A. *Sentiment Analysis*

Sentiment analysis can be considered a classification process as illustrated in Fig. 1 [1]. The class labels constitute the polarity of each text, which can be positive, negative or neutral. Additionally, sentiment analysis can identify the emotional state of a word, such as anger, sadness, happiness, etc.

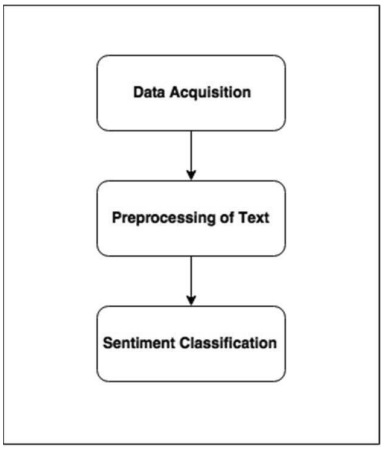

Figure 1. Sentiment Analysis process.

1) *Data acquisition*

Data acquisition is an extremely important phase, since without a well-defined and representative dataset the rest of the process is useless. Today there are several web sites and API that we can use to obtained users' tweets. In Table 1 we can see some examples of public datasets and their characteristics.

For our approach we have constructed the dataset using the Twitter API to obtain the users' tweets. These tweets are defined in topics such as: Politics, Economics, Tourism and Marketing.

2) *Preprocessing of text*

The preprocessing phase allows to reduce noise in data, reduce its dimension and perform suitable selection of features. There are some techniques that we can use:

- Remove all punctuations, symbols, numbers [8].
- Remove stopwords.
- Remove all URLs (e.g. www.xyz.com), hashtags (e.g. #topic), targets (@username)
- Stemming: this technique allows to eliminate the endings or beginnings of words and detect their root form [9].
- Lowercasing: it is a technique to lower-case all words. By doing so, many words are merged and the dimensionality of document collection is reduced [10].

TABLE I. PUBLICLY AVAILABLE DATASETS

Source	Type	Description
Stanford University, students graduate of Computation and Science[1]	Opinions about brands and products.	Training dataset contain 1600000 sentences, tag with 0 = negative, 2 = neutral, 4 = positive and test dataset contains 497 sentences.
Michigan University[2]	Opinions	Training dataset contain 7086 sentences, tag with 1 (positive) or 0 (negative) and test dataset contain 33052 sentences.
SemVal 2015[3]	Opinions	Opinions
Stanford dataset[4]	Movie reviews	50,000 movie reviews

B. *Aproaches for Sentiment Analysis*

There are two main techniques for sentiment analysis in Twitter: lexicon-based and machine learning approaches. They both use techniques to classify text or tweets into classes. They can be divided into:

1) *Supervised Learning*

Supervised learning depends on training documents that contain labels. Usually these labels are obtained by manually labelling by a supervisor, which means that the documents or tweets are cataloged or labeled according to the supervisor previous knowledge and experience [11], [12]. In supervised learning, the training dataset will be used to build the model, and the testing dataset will be used for labeling prediction.

Several supervised machine learning techniques have been formulated to classify the tweets into classes, among them we have Support Vector Machine (SVM) and Naive Bayes (NV), which have achieved success in sentiment analysis [1].

2) *Unsupervised Learning*

This approach does not have a supervisor or data label. It just has input data and aims to find the regularities in the input. There is a space where we can see patterns more frequently and also we can analyze its robustness [13].

3) *Lexicon-based approach*

This technique works on an assumption that the collective polarity of a document or tweet is the sum of polarities either individual words or phrases [14].

In Fig 2. we present a lexicon-based model.

[1] http://help.sentiment140.com/for-students/
[2] https://www.kaggle.com/c/si650winter11
[3] http://alt.qcri.org/semeval2015/task12/
[4] http://ai.stanford.edu/~amaas/data/sentiment/

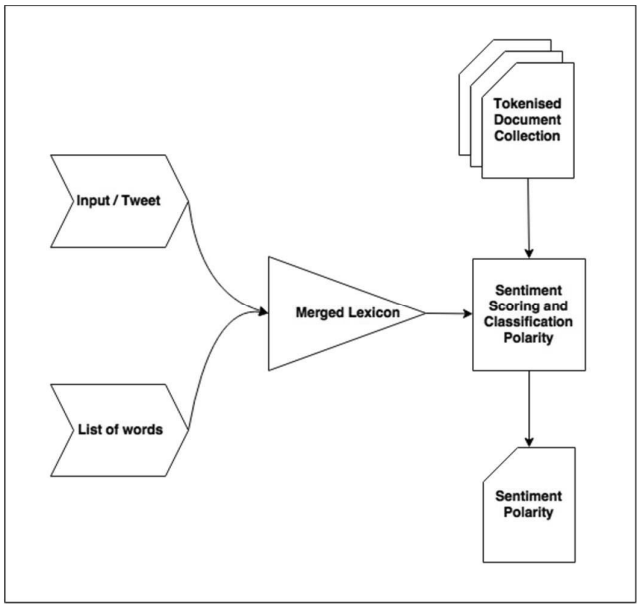

Figure 2. Lexicon-Based Model

There are two sub classifications for this approach, as follows:

3.1) Dictionary-based approach

It is based on terms (seeds) that are usually collected and manually annotated. This set of words grows by searching the synonyms and antonyms out of a dictionary, e.g., one can use WordNet as dictionary of words.

3.2) Corpus-based approach

This approach has the objective of providing dictionaries related to a specific domain. These dictionaries are generated through the terms (seeds) and grows with words related either for statistical or semantic techniques:

a) Statistical approach

Statistical methods are used to determine the polarity of text. Through counting of words, if there are more positive words than negative word in the text, the polarity will be positive. In the other case, if there are more negative words than positive words, the polarity will be negative, and finally if there are (more or less) the same positive and negative words, the polarity will be neutral.

b) Semantic approach.

This approach gives sentiment values based on different principles to find similarity between words. The principle gives similar sentiment values to semantically close words.

III. RELATED WORK

In this section we present works related with sentiment analysis in Twitter using lexicon-based and also supervised learning. These works use public datasets or APIs for data acquisition.

In last years sentiment analysis has received many attention. In spite of that, there are still challenges in the analysis of microblogging. At the same time, the amount of text generated daily have require the use of simples or fast methods to text analysis. Here we find lexicon-based and machine learning approaches to sentiment analysis [15].

- Lexicon-based.

The lexicon-based approach exploits the dictionaries of positive and negatives words found in the text to calculate the polarity or score. For example, SentiWordNet is a lexical resource that support sentiment analysis applications which provides an annotation or numeric values (positive, negative, neutral) [16]. A remarkable feature of SentiWordNet is that could be built automatically. Unlike, Affective Norms for English Words(ANEW) [17], which is a lexicon with affective norms for English words that must be built manually. In this way, Nielsen created a lexicon named AFINN inspired from ANEW but add contemporary phrases used on social networks [18].

On the other hand, the dictionaries are too restricted in size of words. To tackle this issue, in [19] a lexicon is semi-automatically built from web documents without a web-derived lexicon as WordNet. As a result, the lexicon is not dependent on words of class.

- Machine Learning

Machine learning approach uses different methods to sentiment analysis. In general, the accuracy using machine learning is better than lexicon-based approaches. Inside machine learning methods applied to sentiment analysis, we have Naïve Bayes(NV), Support Vector machine (SVM), Decision Trees, among others [20].

In [8] the comparison between supervised and unsupervised methods is mentioned and the obtained results indicate that both SVM and NV are very precise, unlike lexical methods that are not very effective. Within this research, the use of bigrams is highlighted as a method with better accuracy. In [21] supervised methods are mentioned to have better performance over lexical ones, that are unsupervised. Although it should be considered that supervised methods require a large amount of data or a corpus to perform an appropriate classification. Within the supervised methods they detailed that SVM presents a high accuracy over other methods.

In this work we have used a lexicon-based method using a public dictionary of word in Spanish language to sentiment analysis in Twitter.

IV. PROPOSAL

A. *General Description*

The aim of this work is to show the develop of an app which use sentiment analysis to evaluate the polarity of tweets. The application takes advantage of being a mobile app to present the results on screen in a compact and friendly way.

The mobile app was developed in Android Operating System, and allows users to know the polarity (positive, negative or neutral) of five topics previously defined, such as, “Ecuador”,

"Moreno", "Correa", "Coca Cola Ecuador" and "Plaza Foch". These topics are related to Politics, Economics, Tourism and Marketing. The polarity of a particular topic is defined by the total number of tweets classified as positive, negative or neutral.

The information used to feed the mobile app is provided by the social network Twitter, because it is a relevant source of information, with up-to-date data.

B. *Application Architecture*

The architecture of this system uses Model View Controller (MVC), which is a software architecture concept considered as an architectural pattern in software engineering.

The web service communicates with Twitter using the API, that is available to subsequently get the tweets related to each of the topic. These tweets are then stored in a database. The intermediate layer communicates with the android application through a REST web service which returns the information that is requested by the application layer. Fig 3 shows the architecture of the proposed application.

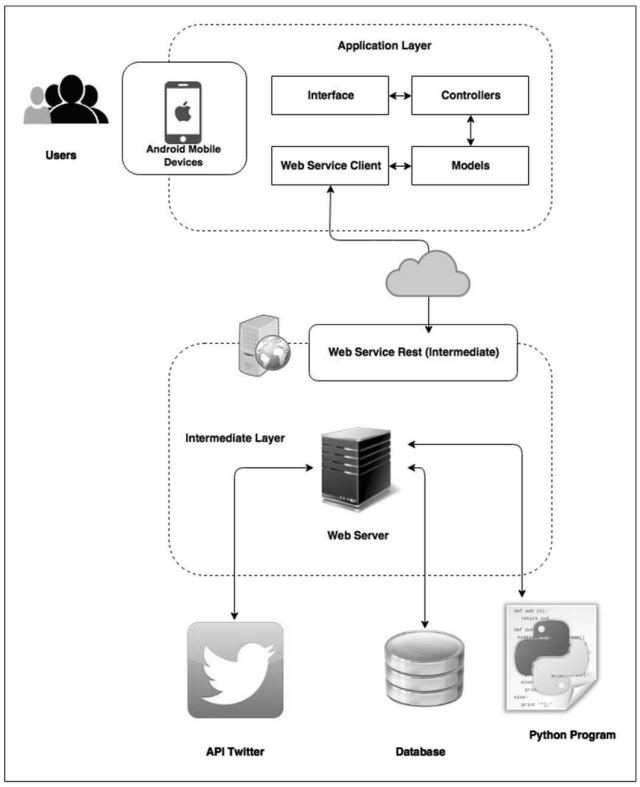

Figure 3. Architecture of proposed application

C. *Usage Overview*

To understand the functionality of the application, the operation of some elements is detailed in Fig 4.

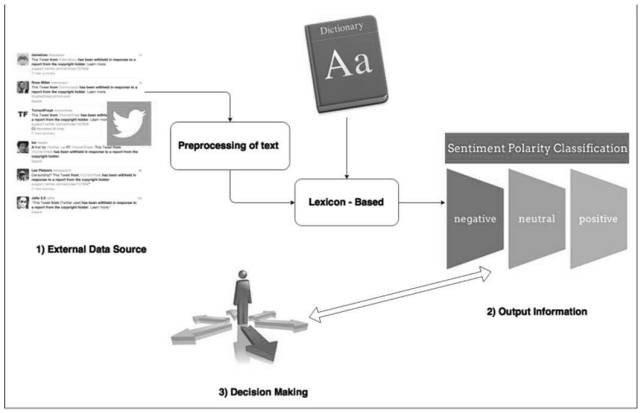

Figure 4. Elements of functionality.

1) *External Data Source*

To extraction of tweets was done using the Twitter API, using python code. The method used is to search for a collection of relevant tweets matching a specified query. In this case keywords were specified for each topic.

2) *Preprocessing of text*

In this component we have used some techniques such as:

- Remove www and http inside of tweet.
- Remove @username contents in the tweet.
- Lower-case all words in the tweet.

3) *Dictionary*

Dictionary is a set of words that have a polarity as positive or negative. The dictionary has a total of 4276 words in Spanish, of which 1555 are positive and 2721 negative[5].

4) *Lexicon-Base*

This component is an intermediate layer which uses a dictionary of words and tweets, and then it executes a lexicon method to polarity classification of each tweet: if the tweet has more positive words then the tweet will be positive, if the tweet has more negative words then tweet will be negative, otherwise the tweet is neutral.

The following Fig 5 shows an example to classification of tweet using lexicon method.

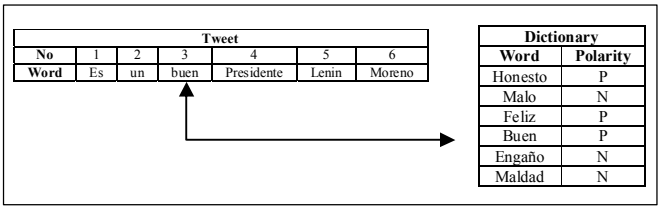

Figure 5. Classification of tweet using lexicon method.

[5] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

- Number of positive words in the tweet : 1
- Number of negative words in the tweet : 0
- Number of neutral words in the tweet : 0

Therefore, the tweet is cataloged as positive, since it has more positive words.

5) Output Information

The information output shows both the total number of related tweets, according to the topic, along with the corresponding polarity.

V. EXPERIMENTAL SETUP AND TESTS

The aim of this section is to show the performed tests of the developed app. Five tests were performed. Each test uses a different topic, seeking to cover different social problems in each case. The topics selected include Ecuador, Moreno, Plaza Foch, Coca Cola Ecuador and Moreno as shows in Fig 6. During the performance tests of the mobile app, the following results were obtained.

- **First Topic: Ecuador**

The results of polarity for this topic were positive 62,14%, neutral 12,04 % and negative 25,80%. Fig 7 on the left shows the results of Ecuador with their polarity, indicating that this topic has great acceptance in the data set taken.

Figure 6. Main screen showing the five topics.

Figure 7. Polarity consultation with topics Ecuador and Moreno.

- **Second Topic: Moreno**

This topic got 880 tweets and the results of polarity were positive 18,97%, neutral 29,20% and negative 51,81%. In Fig 7 on the right the results of Moreno with its polarities are presented. Similar to the previous test, the difference in percentages is notorious. But in this case it was not positive but rather negative.

- **Third Topic: Plaza Foch**

The third got 323 tweets and the results of polarity were positive 29,10%, neutral 43,34% and negative 27,55%. In Fig 8 on the left the results of Plaza Foch with its polarities are presented. In this case, neutral opinions were more that positive and negative opinions.

- **Fourth topic: Coca Cola (Ecuador)**

The Fig 8 on the right shows the polarity were positive 18,97%, neutral 29,20% and negative 51,81%. The results indicated the brand don't have good acceptance in Ecuador.

Figure 8. Polarity consultation with topics Plaza Foch and Coca Cola Ecuador.

- **Fifth topic: Correa**

This topic is related to an ex-president of Ecuador who carries out a political campaign against the current government. In Fig 9 the results for Correa with its polarities are presented.

Figure 9. Polarity consultation with topic Correa.

VI. CONCLUSIONS AND FUTURE WORK

In this work we analyzed the impact of sentiment analysis in some topics such as Politics, Social, Tourism, and Marketing. This work was addressed in Spanish language and with themes from Ecuador. To this goal, lexicon approach was used in addition to a dictionary of positive and negative words, which allowed for the polarity classification of each tweet extracted by twitter API.

The techniques used in the processing of text, such as, remove URLs, usernames and lower case, have allowed the reduction of features or words, as well as processing time. Also the classification of tweets by the lexical method demands a great processing time, which means that more data, more processing time.

As future work, we aim to do a further study of the topic with the greatest impact on society, since in some topics there is not much information related in Twitter. We also thought that both machine learning and lexical methods could be combined to improve sentiment analysis in Twitter.

ACKNOWLEDGMENTS

This work was possible thanks to Senescyt of Ecuador for the financing of research studies at the Polytechnic Institute of Leiria, Portugal.

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.

[2] L. L. Pang B., “Opinion mining and sentiment analysis,” Found. Trends Inf. Retr., 2008.

[3] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, “Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination,” 2015.

[4] B. Liu, “Sentiment Analysis and Opinion Mining,” 2012.

[5] L. L. Pang B., “Sentiment classification using machine learning techniques,” Proc. EMNLP, pp. 79–86, 2002.

[6] Z. L. Chaovalit P., “Movie review mining: a comparison between supervised and unsupervised classification approaches,” Proc. 38th Hawaii Int. Conf. Syst. Sci., p. 112, 2005.

[7] L. Y. . Ye Q., Liu B., “Sentiment Classification for Chinese Reviews: A Comparison between SVM and Semantic Approaches,” Proc. Fourth Int. Conf. Mach. Learn. Cybern., pp. 2341–2346, 2005.

[8] V. A. Kharde and S. S. Sonawane, “Sentiment Analysis of Twitter Data: A Survey of Techniques,” Int. J. Comput. Appl., vol. 139, no. 11, pp. 975–8887, 2016.

[9] R. Duwairi and M. El-Orfali, “A study of the effects of preprocessing strategies on sentiment analysis for Arabic text,” J. Inf. Sci., vol. 40, no. 4, pp. 501–513, 2014.

[10] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, “A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis,” vol. 10450, no. September, 2017.

[11] C. Silva, “Inductive Inference for large scale text classification,” p. 210, 2008.

[12] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009.

[13] A. K. Behera, “Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis,” pp. 128–133, 2017.

[14] H. Krishnan, “Sentiment Analysis of Tweets for Inferring Popularity of Mobile Phones,” Int. J. Comput. Appl., vol. 157, no. 2, pp. 2–4, 2017.

[15] L. Pollacci, A. Sˆırbu, F. Giannotti, and D. Pedreschi, “Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis on Twitter,” vol. 10640, pp. 114–127, 2017.

[16] A. Esuli and F. Sebastiani, “SentiWordNet : A High-Coverage Lexical Resource,” pp. 1–26, 2007.

[17] M. M. Bradley and P. P. J. Lang, “Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings,” Psychology, vol. Technical, no. C-1, p. 0, 1999.

[18] F. Å. Nielsen, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” CEUR Workshop Proc., vol. 718, pp. 93–98, 2011.

[19] L. Velikovich, S. B. Kerry, and H. Ryan, “The viability of web-derived polarity lexicons,” Naacl, no. June, pp. 777–785, 2010.

[20] A. S. Hosseini, “Sentence-level emotion mining based on combination of adaptive Meta-level features and sentence syntactic features,” Eng. Appl. Artif. Intell., vol. 65, pp. 361–374, 2017.

[21] M. VOHRA and J. TERAIYA, “A Comparative Study of Sentiment Analysis Techniques,” Ejournal.Aessangli.in, vol. 17, no. 4, pp. 313–317, 2013.