

# Entity-based Opinion Mining from Spanish Tweets

Fabián Paniagua-Reyes, José A. Reyes-Ortiz and Maricela Bravo

*Department of Systems, Metropolitan Autonomous University, Azcapotzalco, Mexico City, Mexico*

**Keywords:** Opinion Mining, Natural Language Processing, Social Networks.

**Abstract:** Networking service has grown in the last years and therefore, users generate large amounts of data about entities, where they can express opinions about them. This paper presents an approach for opinion mining based on entities, which belong to banks, musicians and automobiles. Our approach uses machine learning techniques in order to classify Spanish tweets into three categories positives, negatives and neutral. A Support Vector Machine (SVM) and the bag of word model is used to obtain the corresponding class given a tweet. Our experimentation shows promising results and they validate that entity-based opinion mining is achievable.

## 1 INTRODUCTION

Twitter has become one of the most used social networking services nowadays. Thousands of users express their opinion about a named entity, such as a product, service or interesting person. Such opinions can be charged with a positive, negative or neutral polarity. They are communicated within 140 characters, called tweets and they are allowed by the Twitter social network. With regard to opinion tweets, millions of text are generated daily that should be used to make decisions about where to direct an event or action in order to improve a service, product or the image of a famous person.

In addition, opinion texts can be very useful for both public and private organizations, since they provide fresh data about an entity, i.e. data generated almost instantly. Therefore, the decisions taken have a database that considers the instantaneous polarity produced by users on an entity of interest.

Manually processing of all generated information about opinions is impossible, costly and time-consuming. However, it is possible to process data in order to obtain relevant information thanks to data mining techniques.

But, the fact of being expressions of opinions in free text, these can be written in any language. It causes the complexity of processing to increase. In addition, considering that there is a lack of linguistic approaches for opinion mining in the Spanish language, the problem described above is increased. We have detected a need to have approaches for text

analysis in Spanish.

Therefore, this paper focuses on providing a linguistic approach for mining opinions using data from the social network Twitter, to reduce the lack of it for the Spanish language. These data are short text, known as tweets. In this paper, we use a machine learning approach in order to classify texts. Support Vector Machine (SVM) as supervised classification algorithm is used to predict the corresponding class of each tweet into three classes positive, negative or neutral. The main aim is to detect the polarity contained in a text message to a specific entity, such as automotive brand, bank, artists/musicians.

## 2 RELATED WORK

Socher model emphasizes the need for improved Natural Language Processing (NLP). Authors implement Latent Dirichlet Allocation (LDA), autoencoder improved by adding to the objective function topic information. The system generates a distribution 1-dimensional over topics. Classification task is done over lexicons, PMI n-grams lexicons (Pointwise mutual information), negation detection and elongated words. Such system has improved results, which are reported in (Ren et al., 2016).

Contextual semantics are emphasized in (Saif et al., 2016), authors use the SentiCircle which is used for learning directions of the sentiment as the horizontal axis represents a neutral sentiment, the

vertical axis separates positive and very positive above and below horizontal axis negative and very negative sentiment. SentiCircle tasks include term indexing, term-content vector generation (a representation of the term in relation to a previously given value to the sentiment and a degree of correlation to the context), contextual features generation. Also capable of using the negated value of SentiCircle when finding negations. The system is used for entity analysis and tweet level sentiment detection. Results show high F1-measure as 85.45.

Two approaches for polarity analysis for building a framework are mentioned in (Lima et al., 2015), even though, authors have implemented a hybrid architecture as results show improvement. The key elements of the proposed framework are lexicon-based along machine learning based for polarity analysis task, automatic generation of the training set for machine-learning. Short text classification by contextual verification, entity detection among other techniques to reduce false positive detection.

In (Trinh et al., 2016) proposes a system based on, building a sub-divided emotional dictionary in nouns, verbs, adjectives and adverbs, and a method for emotional analysis in English and adaptive with Vietnamese. Core processes include removing foreign words, Stopwords, icons, post-tagging phase. A key step is the emotional analysis evaluation as whether a text is an emotion or non-emotion based. Emotional Dictionary is based in SO-CAL since has demonstrated best results for topic experimentation. Authors claim very good results, average 90.4% in precision.

A corpus obtained by indirect emotional query search with emoticons (a set of characters related to emotion) such as “:)” or “:(” is the start point in (Martinez-Cámara et al., 2015). Pre-process steps include removing new line, opposite emoticons, no clear sentiment emoticons i.e. “:p”, repeated tweets, repeated letters and laugh normalization. Different approaches have experimented with vector modeling

Term Frequency, Term Frequency-Inverse Document Frequency, Term Occurrences and Binary Term Frequency. Results demonstrate superior metric results.

The authors (Ren et al., 2016) improve in results for state-of-art sentiment classification system at SemEval 2013 different models are proposed. Single-prototype and multiple-prototype for word embedding are used for every model which are Neural Model, Topic-Enriched Word Embeddings (TEWE) Topic, Sentiment Word and Sentiment Information in Word Embeddings (SSWE) and TEWE combined with SSWE (TSWE). A

convolutional neural network is used for sentiment classification model. Results demonstrate improvement in multi-prototypes word embeddings with SSWE and TSWE, F-measure is 86.10 and 86.83 respectively.

The authors in (Terrana et al., 2014) report good results without the use of third-party technology. For a corpus obtained from Twitter Social Network, a different perspective for obtaining the polarity by the words contained in a tweet with emoticons, positive for the text with the following character combination “:)”, “:D” or negative with “:(“, “;(“ and their variations. As a result, the lexicon is capable of mapping and enriching informal expressions with slang or grammatical error. The word polarity calculation is obtained by counting the difference in positive and negative occurrences over the sum of positive and negative occurrences. Tweet polarity is given by the averaging the polarity scores of its words.

In (Severyn and Moschitti, 2015) propose a framework for sentiment analysis based on deep learning scheme using a convolutional neural network. Authors claim their results are compared with the top positions in SemEval 2015. The architecture for this framework starts with a Sentence Matrix with all words from the tweets represented by distributional vectors. Followed by a Convolutional Map which objective is to extract word patterns found in the training instances. The system needs to learn the limits to make decisions by implementing an activation function which can be logistic or hyperbolic tangent. Then pooling is needed from the activation function helping to reduce the representation. Another step is calculating the probability of distribution over the labels using Softmax. The final step is sentiment analysis by feeding the network with additional inputs indicating the target phrase in a tweet. Best accuracy result obtained is 84.79.

The presented approach in this work is similar to other works still not the same as in (Sidorov et al., 2012) where authors do not propose an entity-based approach resulting in different results. However, it can be used to support the presented approach here.

Another independent approach to realize classification in of different words according to a lexicon (a group of selected words) that best represent emotions to help determine the polarity of the sentiment which has shown good results in this task. This work can be used to support the presented approach.

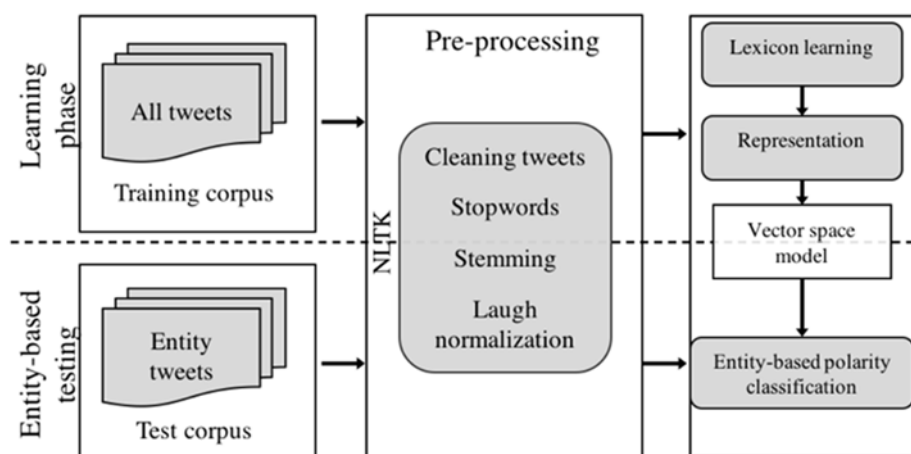


Figure 1: System architecture for entity-centred sentiment analysis.

### 3 SENTIMENT ANALYSIS APPROACH

In this section is presented the components of the framework for entity-based for sentiment analysis for tweets in Spanish these components are shown in Figure 1, which include the tasks such as pre-processing of tweets, feature extraction, lexicon obtaining, feature weighting by frequency of the occurrences in the tweets of the lexicon and finally, the classification algorithms used for the three topics, such as automobiles, banks and musicians.

#### 3.1 Pre-processing of Text

The corpus must be converted to a common structure by following the next steps, this work helps to improve the results in sentiment analysis phase i.e. removing unwanted characters, accents or foreign language, information that does not add value, and punctuation.

The steps for pre-processing are

- Cleaning tweets
- Removing URLs, mentions (@) and entities (#)
- Stemming
- Removing stopwords
- Laugh normalization

##### 3.1.1 Cleaning Tweets

The first step to obtain a lexicon from the texts of tweets is cleaning the text, a method of separating phrases into words (tokens) and deleting special characters that are not defined in Unicode Transformation List (UTF-8).

##### 3.1.2 Removing URLs, Mentions (@) and Entities (#)

The following step is to remove URL, links to websites, mentioning users using “@”, named entities i.e. #shakira #bmw #volvo #ferrari among others. The objective is to support the algorithms work more accurately by avoiding classification task in such useless text.

##### 3.1.3 Stemming

Improving the text includes, for every word in the lexicon obtained until this point, a process called Stemming helps to reduce the words to its “root”, meaning, it is done by removing suffixes or word variations giving a common word for all its variations as a result. For this task to be applied it is needed the Porter algorithm (Porter, 1980). In Table 1, contains an example, words in Spanish from the tweets, of this process to obtain the roots using the SnowballStemmer from NLTK.3.1.4

Table 1: Examples of words stemming process.

| Word      | Stem    |
|-----------|---------|
| gustar    | gust    |
| europa    | europ   |
| cuenta    | cuent   |
| hermoso   | hermos  |
| ilusion   | ilusion |
| presumido | presumi |
| musica    | music   |
| orgullo   | orgull  |
| proximo   | proxim  |

### 3.1.4 Removing Stopwords

This step removes words which do not add meaning or value to the text (tweet), these words are called (stopwords), thus they are helpless to the opinion classification task. This list of words contains articles, prepositions, non-functional verbs among others. Other words that are considered nonfunctional are words with a length of two characters. Negation words “not”/”no” and affirmative words “yes”/”sí” which length is two remain intact in the tweets because they are functional to define the intended polarity in a tweet.

### 3.1.5 Laugh Normalization

People tend to express several feelings with the expression of laugh and they do so by the repetition of patterns. Even tough, for avoiding redundancy in the style to express laughter, during pre-processing, the normalization of laugh is considered a step, with the objective of helping to classify. By using regular expressions which are a set of rules that are applied to get the different possible combinations of the pattern used for laughter and replace with a common expression, an example of these is in Table 2 where the symbol (+) means one or more occurrences.

Table 2: Laugh normalization.

| Pattern | Phrase   | Normalized laugh |
|---------|----------|------------------|
| (ja)+   | ja       | jaja             |
| (je)+   | jeje     | jaja             |
| (jo)+   | jojojo   | jaja             |
| (ji)+   | jijijiji | jaja             |
| (#)?lol | #lol     | jaja             |

## 3.2 Lexicon Learning (Entity Extraction)

The lexicon or dictionary is built in this phase, after applying the pre-processing, as non-repeated lexical units from the tweet corpus. This lexicon is the set of features to be classified and it is called a bag of words.

With the task of entity extraction, a lexicon normalized and reduced is obtained, which will be used to represent every instance (tweet) by the weighting. The value of a normalized tweet representation of the text in number is that every word has a value depending the importance of it according to the tweet and the corpus. The representation in numbers allows the next step in the process, running the algorithm, to do the tweet classification in positive, negative or neutral. A total of 5936 words are obtained from the text (corpus of

tweets) that shapes the lexicon vocabulary.

## 3.3 Vector Space Model Representation

There are different approaches to obtain the importance or weighting of the vocabulary from a short text. This vocabulary is represented by the vector space model with the model bags of words (BoW) (Sebastiani, 2002), which consist of a set of texts and the vocabulary of terms (entities). Every tweet is represented as a vector  $V_j = (v_{1j}, v_{2j}, v_{3j})$ , where every component  $v_{ij}$  represents the importance that produces this feature  $i$ , word from the lexicon, in the tweet  $j$  relating the words in the tweet with all the corpus.

For the weighting of words, meaning, determine the importance of the term in a tweet, it is used the term frequency of occurrence for a term in a set of texts in the domain of an entity (TF-IDF).

The term frequency (TF) consist of the number of times a term ( $t$ ) from the vocabulary appears in a tweet ( $V$ ), see Equation 1, and the inverse frequency (IDF) determines when a term is common in the set of tweets, see Equation 2. This information is used to calculate the value of TF-IDF using Equation (3). For every term  $v_i$  in a tweet  $V$  the value is normalized in the equation 4.

$$TF(t_i, V_j) = f(t_i, V_j) \quad (1)$$

$$IDF(t_i, V_j) = \log\left(\frac{1 + |V|}{1 + |v \in V : t_i \in v|}\right) + 1 \quad (2)$$

$$w_{ij} = TF(t_i, V_j) \times IDF(t_i, V_j) \quad (3)$$

A normalization phase is carried out from matrix obtained by applying the Equation (4).

$$W_{norm} = \frac{w_{ij}}{\sqrt{\sum_{i=0}^n |w_{ij}|^2}} \quad (4)$$

where  $n$  means total number of tweets and  $j$  represents each tweet.

## 3.4 Entity-based Polarity Classification

This work relies on weighting the features based on the importance of terms in a tweet focused on entities. Specifically, the weighting of terms concerning the classification of tweets is centered in three entities which are brand (automobile), bank, and artists/musicians.

Polarity classification is based on the vectors produced by the weighting of terms in relation to

every tweet. The terms are weighted by implementing TF-IDF algorithm explained in the previous section and then the entity-based tweets polarity classification process is executed as supervised learning by evaluating the features. These vectors are the feed for this next task in classification which has been executed with the SVM algorithm to analyze the behavior and to be able to decide for each tweet its polarity from the corpus.

The objective of this phase is to build a supervised learning classifier of tweets capable of predicting the polarity from three possible categories positive, negative or neutral. To make this possible it is necessary to divide the dataset into two groups, one for training and the remaining for testing.

As for how it has been mentioned before, the classifier is centered in entities, meaning, it obtains just one vocabulary for the set of tweets.

The task of opinion classification based on entities it is done by the algorithm Support Vector Machines (SVM) (Chang and Lin, 2011) that builds hyperplanes in an n-dimensional space from the training tweets, these hyperplanes are used to predict the class for new tweets. It is done by "plotting" each feature from the training data and then realizes a classification task were in a two dimension space a straight line divides in to n-number or groups, for the system presented it will be 3 classes, positive, neutral, and negative. The SVM algorithm iterates trying to find the set of features that best represent a class. Next level of complexity is adding n-dimension to the model so the best way to classify is by finding hyperplanes to find the features that adjust to the representation of every class.

The idea is to evaluate sentiment classification by the execution of the algorithm SVM, since it was identified as the best option compared with C4.5, K-NN and Naïve Bayes in previous investigation work (Reyes-Ortiz et al., 2017), evaluating the weighting of the terms (TF-IDF) to find the best solution in terms of precision. To run the algorithm, the WEKA framework was used (Garner, 1995) Using the default configuration of the program (10 cross-fold validation and 66% of the dataset is used for training and the rest 34% for testing).

## 4 DATASET

We use a dataset provided in RepLab (Amigó et al., 2013) for the specific task called "reputation polarity", whose purpose is to decide whether the content of a message (Tweet) in Spanish has positive

or negative implications for the reputation of entities, such as automotive brand, financial institution, educational institution or person famous in music. Such entities are grouped into four topics musicians, banks, universities and car brands. The dataset is manually tagged into three labels mentioned by human experts in linguistic.

We rely on that an entity-based opinion mining can provide promising clues to determine the entity's reputation. Thus, we focus on three entities for each topic in order to obtain a polarity from their tweets, by analyzing texts from Twitter and to determine whether it has negative or neutral positive implications. We have decided that the universities domain entities should be left out because they are not balanced with respect to the entities of the other domains.

From the selected data set, 6965 effective texts were obtained for all entities, for which it was possible to obtain their Twitter content and, in addition, they were manually classified with their label or category for polarity (Positive, Negative, Neutral) by human experts as indicated in (Amigó et al., 2013). This data set represents an excellent frame of reference for the evaluation of opinion mining algorithms.

For evaluation, we run the experiments with a specific configuration, due to our paper focuses on an entity-based opinion mining. Therefore, we separate the corresponding tweets to the entity under evaluation from the final data set. Table 3 shows the number of effective tweets for each entity to leave out in the corresponding evaluations.

Table 3: Effective tweets for each entity.

| Entity         | Number of tweets |
|----------------|------------------|
| #Bankia        | 716              |
| #BBVA          | 623              |
| #Santander     | 158              |
| #Ferrari       | 317              |
| #Volkswagen    | 193              |
| #Yamaha        | 150              |
| #Shakira       | 622              |
| #JustinBieber  | 274              |
| #JenniferLopez | 327              |

The rest of tweets are used to train our classifier model in each running experiment. For example, to evaluate #BBVA entity, we use 6342 tweets for training model and 623 tweets for testing the classification task.

## 5 ENTITY-BASED EXPERIMENTATION AND RESULTS

Experimentation consists on to execute a classifier algorithm to predict the corresponding class for each tweet for entities.

Each experiment is carried out by removing the tweets of the entity to be evaluated and then, the rest of the tweets are used for as training corpus, from which the vocabulary is learned. Also, we weighed the terms of lexicon obtained by the bag of word model and finally, the SVM classifier is employed to determine the correct polarity of each tweet to be tested. Nine entities are evaluated, three for each topic from our corpus.

An evaluation of all experiments is performed using the well-known metrics Precision (P), Recall (R) and F-measure, which have been widely used in any task of textual classification. These metrics compare the results of the classifier to be evaluated with the external confidence values (previously classified tweets), using the following values a) True Positive (TP) is the number of correct predictions of the classifier that correspond to the external judgment of confidence (previously classified tweets); True Negative (TN) is the number of correct predictions from the classifier that does not correspond to the external judgment of confidence; False Positive (FP) corresponds to the number of incorrect classifier predictions that correspond to the external judgment of confidence; and finally, False Negative (FN) is the number of incorrect predictions of the classifier that do not correspond to the external judgment of confidence.

Under these criteria, Precision (P) is used to evaluate the algorithms in terms of positive prediction values, which is defined, in Equation (5).

$$P = \frac{TP}{TP + FP} \quad (5)$$

Also, Recall (R) is used to express externally the correct correspondences rate with the previously classified tweets with high confidence (Equation 6).

$$R = \frac{TP}{TP + FN} \quad (6)$$

Finally, F-measure that represents the harmonic mean between Precision and Recall, which is based on obtaining a weighted unique value between them (Equation 7).

$$F - measure = 2 * \frac{P * R}{P + R} \quad (7)$$

All experiments use word weighting using the TF-IDF and SVM algorithm as the classifier. On the other hand, we show the results based on entities because it is the essence of this paper. In addition, we expose and analyze the results in each entity of interest. So that, Table 4 shows results of polarity identification for bank entities in terms of Precision, Recall and F-measure average with the purpose of summarizing results.

Table 4: Classification results of bank entities.

| Entity     | Precision | Recall | F-measure |
|------------|-----------|--------|-----------|
| #Bankia    | 0.66      | 0.67   | 0.67      |
| #BBVA      | 0.71      | 0.72   | 0.71      |
| #Santander | 0.57      | 0.54   | 0.55      |

Table 5 presents the summary results for music entities, specifically, musicians like Shakira, Justin Bieber and Jennifer Lopez are used for identifying their reputational polarity from Spanish tweets.

Table 5: Classification results of music entities.

| Entity         | Precision | Recall | F-measure |
|----------------|-----------|--------|-----------|
| #Shakira       | 0.72      | 0.74   | 0.73      |
| #JustinBieber  | 0.52      | 0.56   | 0.54      |
| #JenniferLopez | 0.63      | 0.63   | 0.62      |

Three entities of car brands are selected from corpus. Ferrari, Volkswagen and Yamaha are monitored and results are shown in Table 6.

Table 6: Classification results of automobile entities.

| Entity      | Precision | Recall | F-measure |
|-------------|-----------|--------|-----------|
| #Ferrari    | 0.56      | 0.58   | 0.56      |
| #Volkswagen | 0.60      | 0.56   | 0.57      |
| #Yamaha     | 0.77      | 0.77   | 0.76      |

Based on Tables 4, 5 and 6, we have noticed that the entities “#BBVA”, “Shakira” and “Yamaha” have show the best results in their respective topics. The entity “Yamaha” has shown the best results in the whole corpus, achieving an approximate 76% effectiveness of the classifier algorithm to determine the polarity of its tweets.

## 6 CONCLUSIONS

This paper has presented an approach for opinion mining, using the networking service called Twitter for obtaining data about entities. Our approach uses a machine learning technique in order to predict

polarity of tweets by analyzing their texts. The main idea is to determine whether words of a tweet has negative, positive or neutral implications for the entity in question.

The main contributions of this papers are a) the extraction of a lexicon from the corpus, which is useful to classify texts; b) the entity-based classifier using lexicon for opinion mining from Spanish tweets; and c) the experimentation carried out confirms that entity-based opinion mining can provide promising clues to determine the entity's reputation.

In this paper, we focus on Support Vector Machine algorithm to classify tweets, which performance the best results for the "Yamaha" entity, achieving an approximate 76% effectiveness of the classifier algorithm to determine the polarity type.

It is important to emphasize that this paper has made a relevant contribution to the problem of lack of linguistic approaches in Spanish texts since our technique uses texts extracted from Twitter in Spanish. In addition, opinion mining approach is provided as a tool to make possible the analysis of reputational polarity for entities in Spanish.

As a future work, experimentation can be realized with other semantic models of representation, such as word embedding or distributional models of semantic representation. In addition, the evaluated approach for opinion mining can help to monitor online reputation of interest entity.

## ACKNOWLEDGEMENTS

This work was partly supported by SEP-PRODEP. The authors would like to thank the Autonomous Metropolitan University Azcapotzalco and SNI-CONACyT.

## REFERENCES

- Aha, D. W., Kibler, D. & Albert, M. K. 1991. *Instance-based learning algorithms*. Machine Learning, 6, 37-66.
- Amigó, E., DE Albornóz, J. C., Chungur, I., Corujo, A., Gonzalo, J., Martin, T., Spina, D.: Overview of replab 2013: *Evaluating online reputation monitoring systems*. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. CLEF 2013. Lecture Notes in Computer Science, vol 8138. Springer.
- Chang, C.-C. & LIN, C.-J. 2011. LIBSVM: *A library for support vector machines*. ACM Trans. Intell. Syst. Technol., 2, 1-27.
- Garner, S.R. 1995. *Weka: The Waikato environment for knowledge analysis*. In: Proc. of the New Zealand Computer Science Research Students Conference, 57-64.
- John, G. H. & Langley, P. 1995. *Estimating continuous distributions in Bayesian classifiers*. *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Montré#233;al, Qu#233;, Canada: Morgan Kaufmann Publishers Inc.
- Lima, A. C. E. S., DE Castro, L. N. & Corchado, J. M. 2015. *A polarity analysis framework for Twitter messages*. Applied Mathematics and Computation, 270, 756-767.
- Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. & Mitkov, R. 2015. *Polarity classification for Spanish tweets using the COST corpus*. Journal of Information Science, 41, 263-272.
- Porter, M. F. 1980. *An algorithm for suffix stripping*. Program, 14, 130-137.
- Ren, Y., Wang, R. & Ji, D. 2016. *A topic-enhanced word embedding for Twitter sentiment classification*. Information Sciences, 369, 188-198.
- Ren, Y., Zhang, Y., Zhang, M. & Ji, D. 2016. *Improving Twitter sentiment classification using topic-enriched multi-prototype word embeddings*. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona: AAAI Press.
- Reyes-Ortiz, J. A., Paniagua-Reyes, J. F., Sanchez, L: *Minería de opiniones centrada en tópicos usando textos cortos en español*. Research in Computer Science. (to be published, 2017).
- Saif, H., He, Y., Fernandez, M. & Alani, H. 2016. *Contextual semantics for sentiment analysis of Twitter*. *Information Processing & Management*, 52, 5-19.
- Salzberg, S. L. 1994. C4.5: *Programs for Machine Learning by J. Ross Quinlan*. Morgan Kaufmann Publishers, Inc., 1993. Machine Learning, 16, 235-240.
- Sebastiani, F. 2002. *Machine learning in automated text categorization*. ACM Comput. Surv., 34, 1-47.
- Severyn, A. & Moschitti, A. 2015. *Twitter Sentiment Analysis with Deep Convolutional Neural Networks*. Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile: ACM.
- Sidorov, G., Miranda-Jimenez, S. Viveros, F. & Gordon, J. 2012. *Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets*. Mexican International Conference on Advances in Intelligence (MICAI) Volumen Part I. Pages 1-14.
- Terrana, D., Augelio, A., Pilato, G. 2014. *Automatic Unsupervised Polarity Detection on a Twitter Data Stream*. Semantic Computing (ICSC), 2014 IEEE International Conference on. IEEE.
- Trinh, S., Nguyen, L., Vo, M. & Do, P. 2016. *Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language*. In: Król, D., Madeyski, L. &

Nguyen, N. T. (eds.) Recent Developments in Intelligent Information and Database Systems. Cham: Springer International Publishing.  
1997. *Readings in information retrieval*: Morgan Kaufmann Publishers Inc.

