Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 113 (2017) 65–72

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017)

Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks

Marouane Birjali[a,*], Abderrahim Beni-Hssane[a], Mohammed Erritali[b]

[a]LAROSERI Laboratory, Department of Computer Sciences, University of Chouaib Doukkali, Faculty of Sciences, El Jadida, Morocco
[b]TIAD Laboratory, University of Sultan Moulay Slimane, Faculty of Sciences and Technologies, Béni Mellal, Morocco

Abstract

Sentiment analysis is one of the new challenges appeared in automatic language processing with the advent of social networks. Taking advantage of the amount of information is now available, research and industry have sought ways to automatically analyze sentiments and user opinions expressed in social networks. In this paper, we place ourselves in a difficult context, on the sentiments that could thinking of suicide. In particular, we propose to address the lack of terminological resources related to suicide by a method of constructing a vocabulary associated with suicide. We then propose, for a better analysis, to investigate Weka as a tool of data mining based on machine learning algorithms that can extract useful information from Twitter data collected by Twitter4J. Therefore, an algorithm of computing semantic analysis between tweets in training set and tweets in data set based on WordNet is proposed. Experimental results demonstrate that our method based on machine learning algorithms and semantic sentiment analysis can extract predictions of suicidal ideation using Twitter Data. In addition, this work verify the effectiveness of performance in term of accuracy and precision on semantic sentiment analysis that could thinking of suicide.

Peer-review under responsibility of the Conference Program Chairs.

Keywords: Sentiment Analysis; Machine Learning; Suicide; Social Networks; Tweets; Semantic Sentiment Analysis.

1. Introduction

Nowadays, social networks have changed the way that people express their opinions and points of view[1]. This opportunity is given through textual publications, online discussion sites, product evaluation websites etc. People rely

* Corresponding author.
E-mail address: birjali.marouane@gmail.com

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2017.08.290

heavily on this user-generated content. Social networks offer considerable amount of content generated by the user, it is important content for analysis and offer more services adapted to the needs of users[1,2]. In recent years, the majority of developments in the field of information and opinion exchange have launched the research work for the analysis of feelings expressed on these social networks, presented in the literature as *sentiment analysis*[3].

The analysis of sentiments uses, among other things, the detection of opinions on social networks, clarifying consumer behavior, recommending products and explaining the outcome of the elections. It consists of searching for evaluative texts on the Internet such as criticisms, recommendations and analyzing the feelings expressed therein in an automatic or manual way, in order to understand public opinion[3]. Millions of people are using Twitter and ranked as one of the most visited sites with the average of 58 million tweets per day[2].

However, social networks such as Facebook, Twitter and Google+ are increasingly associated with many social phenomena such as harassment, intimidation, depression[4] or even suicide[5]. This is why it is essential to try to detect potential victims as soon as possible in order to reinforce the prevention of suicide using social networks[5]. In fact, we can cite the case of two American rappers Capital Steez[7] and Freddy E[6] who receive death in the course of commenting live their actions on their social networks accounts.

This research focuses on suicide. Because suicide is simply one of the 20 leading causes of death in the world[8]. Already the pronunciation of suicide word should not be taken with simplicity and lightly. It can be the last cry of someone's help, and yet if the signs and clues are recognized at the beginning, lives could be saved. Suicide is preventable, it is an act of those who have not been able to accomplish others and the prevention of suicide should be the responsibility of everyone.

Recent research studies have demonstrated that social networks can be used to extract feelings of depression[4], which makes it easier for us to prevent suicide online[5]. The content addressed and the phrases used by depression, discouragement and suicides are well known and available[4]. In case of example, these users of social networks are often victims of cyberbullying or sexual harassment. In the case of Twitter, it would set up a real time monitoring in relation to several risk factors[2,9].

On the other hand, sentiment analysis could lead to several challenges like semantic sentiment analysis which is evaluate and implement a new semantic similarity metric to determine the affective content of a word in different dimensions[10]. We have proposed a distance metric based on WordNet to determine the semantic orientation of social networks data[11]. The purpose of this paper is to propose a method of predicting suicidal ideas, to predict suicidal acts and ideas using data collected from social. In this work, we use Weka as a data mining tool to extract all useful information for the classification of this data according to the machine learning algorithms implemented in Weka. Therefore, we present our algorithm to calculate the semantic similarity between the tweets collected from Twitter in the training set based on a semantic analysis resource using WordNet.

The paper is organized as follows: Section 2 presents related works. Proposed method of Suicidal acts analysis is in section 3, while presenting the collection of data from social networks based on the vocabulary associated with suicide and presenting the proposed algorithm. Section 4 gives experimental results, conclusion and future works are given in Section 5.

2. Related Works

Social networks have attracted the attention of researchers, which attempts to understand and analyze, among others, the structure of interconnection and the interaction of users in social networks. Internet users tend to express their opinions and feelings and talk about their lives and activities of everyday life via Twitter[3,9,11]. The application of machine learning methods for the identification of suicide has increased in recent years. The LIWC2007 (Linguistic Inquiry and Word Count Version 2007) application judges evaluate emotional and cognitive words and expressions in written and oral sentences of individuals. Noting that the use of LIWC2007 could identify and trend an emotional post[12]. The work of Ramirez-Esparza, et al.[13] focused on the idea of language markers and on the discussion of feeling depression by collecting information from both already depressed people and undepressed forums while using Bulletin Board Systems (BBS). They also explain that depressed people who have written in English are more likely to report medical problems.

Sentiment analysis is treated as a task of natural language processing at several levels of granularity. There has been a significant amount of research on feeling analysis, rule-based approaches, from bag-of-words to machine learning algorithms. From a classification task at the level of the document in Turney[14]. It is treated at the level of the sentence in Hu and Liu[15] and more recently at the sentence level Wilson et al.[16]. Among the social networks most used as Twitter, its users publish tweets on real time and opinions on any topic. Two main areas of mining opinion research work either at the level of the document[17]. The two methods of classification at the document level and at the sentence level are generally based on the identification of opinion words or sentences.

In this work, we compute the semantic similarity of a tweet collected from Twitter and training data. We use several machine learning algorithms to build our work. In this context, we focus on pessimistic, bad thoughts and thinking through suicide. We evaluate our method and present the results to predict the semantic orientations of Twitter data.

3. Methodology of Suicidal acts analysis

The methodological needs of this work can be divided into four main parts: the first is the needs related to the manual construction of the vocabulary associated with suicide theme, the second part is about the collection of Twitter data, and then the needs related to automatic classification using machine learning algorithms implemented in Weka[18], the last is the requirements for a semantic analysis of these sentiments to improve our results. The proposed architecture of our method of suicide detection based on social network, machine learning and semantic analysis, is shown in Figure 1.

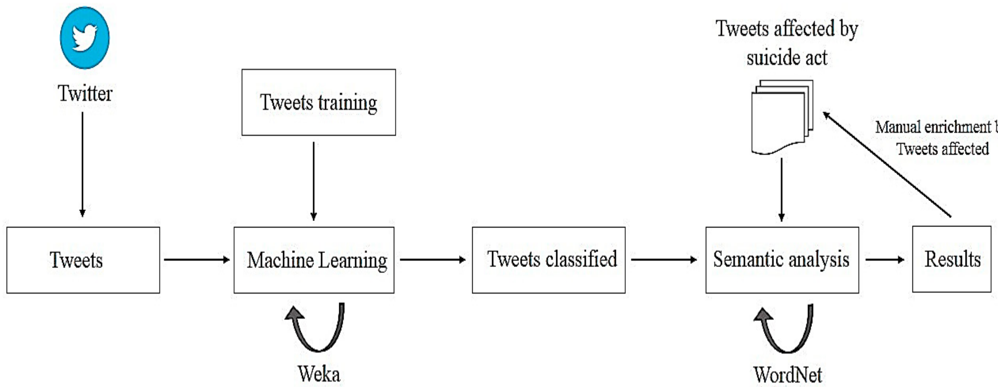

Fig. 1. Architecture of our methodology work of suicide detection.

3.1. Vocabulary associated with Suicide act

In this subsection, we begin by defining a manual vocabulary associated with the various themes related to suicide (fear, depression, harassment, etc.). Subsequently, this vocabulary is divided into several categories and sub-categories in order to easily determine the degree of threat of each tweet. For example, in case of harassment via Twitter, it is highly necessary to identify the recipient concerned by these tweets because they are the ones who can act. On the other hand, because most textual posts on Twitter are published in English, it was preferable to define the vocabulary in English.

3.2. Data collection

The collection of tweets from Twitter is an essential part in our work. In this sense, Twitter4J API is used to collect tweets as shown in Figure 2. Twitter4J is a Java library for programming applications related to Twitter. In addition, it is an integrated Java library with all services related to Twitter[19]. In this work, we collect the tweets using the search word defined in the part of the manual vocabulary. The treatment mechanism is used to collect sequential data from Twitter and the extraction is performed several times for more tweets. The token and access keys obtained, are necessary for the extraction of the tweets in real time from the Twitter page.

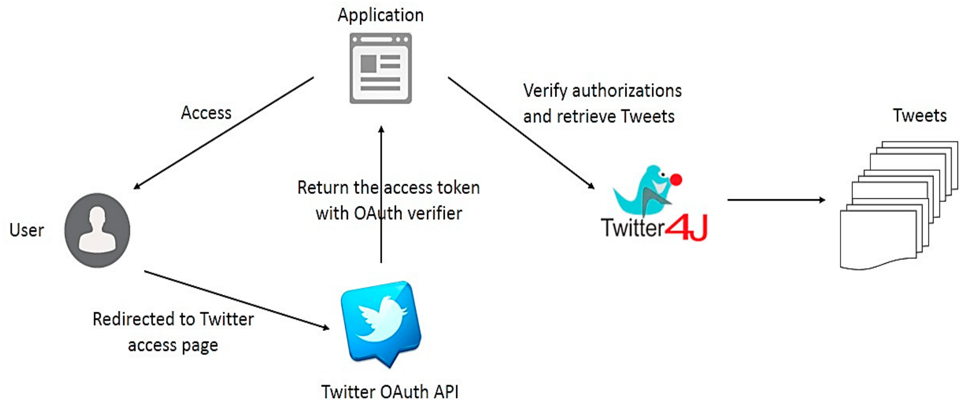

Fig. 2. Architecture for Tweet collection from Twitter using Twitter4J.

3.3. Machine learning classification

Machine learning is a set of classification algorithms that automates the construction of classification algorithms using a set of data called the training set. In this work, the training tweets is a set of tweets already affected by the act of suicide and built manually, it groups input vectors and their corresponding classes. From this training set, a classification model is defined which allows to classify the input characteristics vectors in the corresponding classes. Then, the use of machine learning algorithms for the automatic detection of strongly linked suicidal tweets was developed by supervised or unsupervised learning[20,21]. Supervised learning methods are performed by studying the characteristics that possibly possessed by tweets in particular suicidal classes. This method usually named after classification approach. This classification divides all existing tweets into tweets. While using training tweets and our vocabulary associated with suicide theme, the proposed method produces a performance analysis model to determine in which tweets have a high risk that the authors of these tweets pass the act of suicide or tweets suspect safely, is saying without the author of the tweet will do it.

Several machine learning methods using Support Vector Machines (SVM), Maximum Entropy and Naive Bayes are used for classification[22]. A number of features can be used for ranking, such as term presence, term frequency, negation, n-grams and part of speech[23]. The use of these main features can be to extract sentences, the semantic value of words and tweets. For the case of our work, we propose to study the textual content of tweets, especially associated with user profiles to accurately classify user profiles, so we do not consider the tweet of a particular user. On the other hand, we have targeted our tweets associated with vocabulary to consider the particular user.

3.4. Semantic analysis measure based on WordNet

After applying different machine learning algorithms, we will use semantic analysis. Our method for semantic analysis will based on WordNet database, where each term is associated with others[11]. This database is of English words that are interconnected. If two words are close to each other, they are semantically similar. Specifically, we are able to determine similarity as synonymous. We map the terms and examine their relationship in the ontology. The main task is to use the stored documents that contain terms and then check the similarity with the words that the user uses in their sentences. Thus, it is useful to show the polarity of sense to users. For example in the sentence "I feel very happy", the term " happy " is compared with the feature vector stored synonyms. Suppose 2 terms; "Happy" and "satisfied" tend to be very similar to the word "happy." Now, after the semantic analysis, "happy" replaces "happy" that gives a positive polarity.

The study[24] provides an approach to the social integration of data and information using semantic technology and knowledge representation using ontology. The method uses an algorithm groups words from their similarity between them and chooses the root words of the densest groups and semantics research used to improve system performance[25]. The authors of the work[26] are proposed a technique for finding experts in social media websites by analyzing social graphs and user interactions post and member groups. They constructed a model from social graphs in order to reflect the interests of users based on the implicit and explicit social data from Facebook[26]. The reason for using a semantic analysis in this subsection is to establish the semantic meaning of the new tweets set in relation to training tweets by using the semantic meaning of the elements (words) of each tweet.

The aim of this subsection is to implement the proposed algorithm of semantic analysis of tweets that can advance our research. The principal contribution is to propose an algorithm for computing the semantic similarity between training tweets and the new test tweets collected by Twitter4J, using WordNet as an external network semantic resource. In addition, this contribution will based on Leacock and Chodorow approache[27,28]. This approach based on the combination of the method of informational content and the counting of the arcs method. In fact, the semantic measure proposed by Leacock and Chodorow is based on the shortest length between two syntaxes of WordNet. This measure is defined by the formula:

$$Sim_{lc}(A,B) = -\log\left(\frac{cd(a,b)}{2*L}\right) \tag{1}$$

L is the longest length, which separates the concept root, of ontology, of the concept more in bottom. We indicate that (A, B) is the shortest length that separates A of B.

To compute the semantic similarity between words of new tweets set and words of the tweets training set, we apply the following algorithm:

```
InputData : tweetSet, training_tweets
RemoveStopWord (tweetSet) and RemovePunctuation (tweetSet)
For each element ∈ (tweetSet, termOftweet)
Write(tweetSet, termOftweet)
End for
List(IndQuery) = indexing (training_tweets)
SemSim = 0; A = 0; B = 0
For each term ∈ List(termOftweet)
N = wordcount(term)
For each elem ∈ List(IndQuery)
R = wordcount(elem)
A ← A + N × R × Sim(term, elem)
B ← B + N × R
End for
End for
SemSim ← A / B
Return (tweet, SemSim)
```

The proposed semantic analysis measure is presented by the following formula:

$$Sim_{lc}(a,b) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} a_i * b_j * Sim(i,j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} a_i * b_j} \tag{2}$$

i: represents the terms of the training tweets of b.
j: represents the terms of the new tweet a.

qi: is the frequency occurrence of the term i in training tweets q.
dj: is the frequency occurrence of the term j in new tweet a.
Sim(i, j): is the semantic similarity measure between the two terms of new tweets i and training tweets j.

4. Experimental results and Analysis

For sentiment analysis related on the suicidal acts, we use our algorithm to compute the semantic similarity between the new collected tweets and the training tweets already affected by the act of suicide, because this approach requires an important semantic similarity score based on WordNet as an external network semantic resource.

The experiment is performed using the Weka tool. The choice of this tool is suitable for the fact that it is widely adapted and used in the field of data mining and machine learning, and easy to manage and use. It is simply compatible with the data format we have chosen. The following figure shows the results of Weka's classification of tweets. The percentage of suspicious tweets at risk of suicide and tweets suspect to risk compared to suspect tweets.

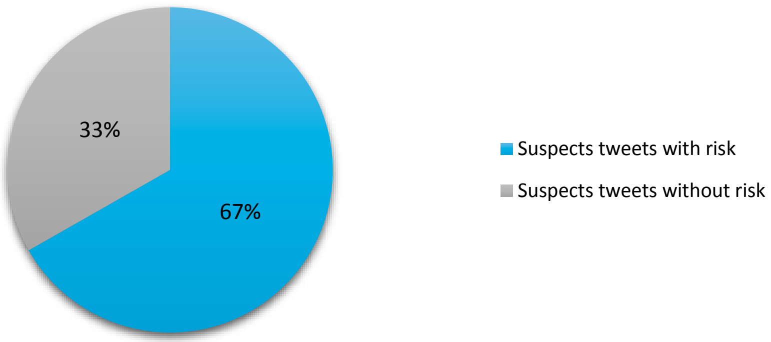

Fig. 3. Suspects tweets with and without risk of suicidal acts.

As we have already mentioned, we examine the classification using Weka. In this tool, we find all the corresponding machine learning algorithms classified in a tab dedicated to classification algorithms. After loading the new tweets and training tweets in Weka, we apply the different algorithms on this data. The Weka tool provides options for evaluating the performance of the model proposed in this experiment, we consider only two. First, the classification algorithms are applied to the set of data. Then we work on cross-validation. The results below shows the evaluation of the analysis tests associated with the suicide act of the machine learning algorithm. While using TwitterAPI, only the tweets related to our vocabulary are collected. A dataset is built using 892 tweets. The dataset is divided into training set tweets and a set of tests. The following tables present the different classification algorithms and the results obtained in terms of precision:

Table 1. Cross-validation of evaluations on classifiers for suspected tweets with risk of suicide.

Algorithms	IB1	J48	CART	SMO	Naive Bayes
Precision	77%	81.2%	83.1%	89.5%	87.50%
Recall	82.2%	84.2%	88.5%	89.11%	78.8%
F-score	79.5%	82.6%	85.7%	89.3%	82.9%

Table 2. Cross-validation of evaluations on classifiers for suspected tweets without risk of suicide.

Algorithms	IB1	J48	CART	SMO	Naive Bayes
Precision	63%	75.4%	66.7%	70%	61.00%
Recall	50.5%	72.4%	58.7%	51.4%	76.1%
F-score	55.8%	63.8%	62.4%	59.3%	67.8%

Figure 4 presents the clustering instances, the incorrectly instances are reported by squares and the correctly instances by crosses. By changing the color dimension to other attributes, we can see their distribution within each of the clusters. Moreover, Figure 5 shows the confusion matrix indicates the misclassified tweets. In that figure, the tiles are tweets misclassified and those of cross-classification.

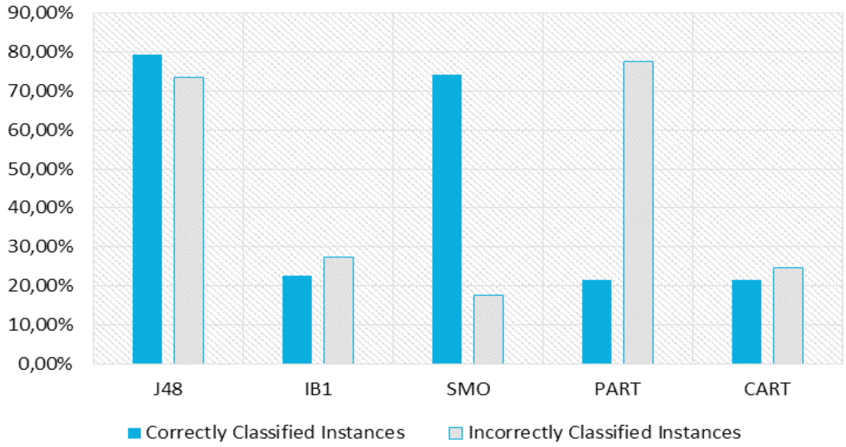

Fig. 4. Performance of different classifiers.

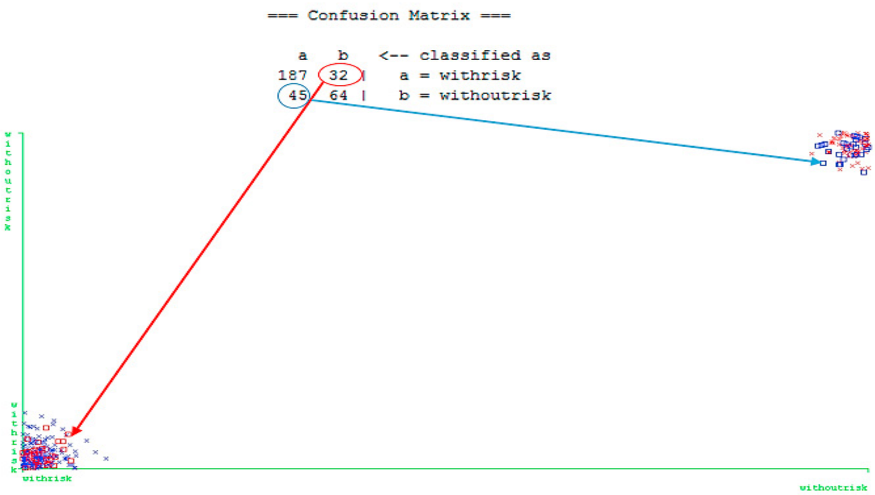

Fig. 5. Visualization errors classifier.

5. Conclusion and Future Work

As part of this work, we present our potentially method based on different machine learning for using the social network Twitter as a preventive force in the fight against suicide. In addition, our work can analyze semantically the Twitter data based on WordNet. In our future work, we plan to further improve and refine our techniques in order to enhance the accuracy of our method. Thereafter, we involved to test multilingual WordNet for tweets and to orient this work in a big data environment.

References

1. Po-Wei Liang, Bi-Ru Dai. Opinion Mining on Social Media Data. IEEE 14th International Conference on Mobile Data Management, pp. 91–96, 2013.
2. R. Li, K. H. Lei,R. Khadiwala, Chang. TEDAS: A Twitter-based Event Detection and Analysis System. ICDE, pp.1273-1276, 2012 IEEE 28th International Conference on Data Engineering, 2012.
3. Gonzalez-Marron D., Mejia-Guzman D., Enciso-Gonzalez A. (2017) Exploiting Data of the Twitter Social Network Using Sentiment Analysis. In: Sucar E., Mayora O., Munoz de Cote E. (eds) Applications for Future Internet. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 179. Springer, Cham
4. Birjali M., Beni-Hssane A., Erritali M. (2017) A Method Proposed for Estimating Depressed Feeling Tendencies of Social Media Users Utilizing Their Data. In: Abraham A., Haqiq A., Alimi A., Mezzour G., Rokbani N., Muda A. (eds) Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016). HIS 2016. Advances in Intelligent Systems and Computing, vol 552. Springer, Cham
5. Kasturi D. V., Nurhafizah T. Suicide detection system based on Twitter. Science and Information Conference 2014, pp. 785-788, August 27-29, London, UK
6. https ://twitter.com/Freddy_E
7. https ://twitter.com/CapitalSTEEZ_
8. Zhao, D. & Rosson, M.B., (n.d), Retrieved from http://research .ihost.com/cscw08-socialnetworkinginorgs/papers/zhao_cs cw08_workshop.pdf
9. Dey P., Sinha A., Roy S. (2015) Social Network Analysis of Different Parameters Derived from Realtime Profile. In: Natarajan R., Barua G., Patra M.R. (eds) Distributed Computing and Internet Technology. ICDCIT 2015. Lecture Notes in Computer Science, vol 8956. Springer, Cham
10. Yang D., Zheng H., Yan J., Jin Y. (2012) Semantic Social Network Analysis with Text Corpora. In: Tan PN., Chawla S., Ho C.K., Bailey J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2012. Lecture Notes in Computer Science, vol 7301. Springer, Berlin, Heidelberg
11. Rawashdeh A., Rawashdeh M., Díaz I., Ralescu A. (2014) Measures of Semantic Similarity of Nodes in a Social Network. In: Laurent A., Strauss O., Bouchon-Meunier B., Yager R.R. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2014. Communications in Computer and Information Science, vol 443. Springer, Cham
12. James W. Pennebaker, et al. The Development and Psychometric Properties of LIWC2007. The University of Texas at Austin and the University of Auckland, New Zealand, http://www.liwc.net/LIWC2007LanguageManual.pdf
13. Nairan Ramirez-Esparza, Cindy K. Chung, Ewa Kacewicz, and James W. Pennebaker. The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches. Association for the Advancement of Artificial Intelligence (www.aaai.org).
14. P. Turney. Thumbs up or thumbs down Semantic orientation applied to unsupervised classification of reviews. Proceedings of the Association for Computational Linguistics.
15. Hu M. and Liu B. Mining and Summarizing Customer Reviews. KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 168-177.
16. Wilson T., Wiebe J. and Hoffmann P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In the Advanced Research and Development Activity (ARDA).
17. Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. Structural opinion mining for graph-based sentiment representation. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011). 2011.
18. WEKA: A machine learning tool set. Software downloadable from http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html.
19. Twitter4J, Unofficial java library for the Twitter API, http://twitter4j.org/, 2007 Yusuke Yamamoto
20. B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen. Detecting suicidality on Twitter. Internet Interventions, vol. 2, pp. 183-188, 2015.
21. Pang, Bo and Lee, Lillian. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2. 2008.
22. G. Vinodhini and R. Chandrasekaran. Sentiment analysis and opinion mining: A survey. International Journal, vol. 2, no. 6, 2012.
23. Yelena Mejova. Sentiment analysis: An overview. Comprehensive exam paper.
24. Yuan Wang, lie Zhang and Julita Vassileva. A User-Centric Approach for Social Data Integration and Recommendation. Human-Centric Computing (HumanCom), 2010 3rd International Conference, pp. 1- 8, 2010.
25. Maxim Grinev, Maria Grineva, Alexander Boldakov, Leonid Novak, Andrey Syssoev, Dmitry Lizorkin,. Sifting Micro-blogging Stream for Events of User Interest. Proceedings of the third ACM conference on Recommender systems, pp. 385 - 388, 2009.
26. Akram AI-Kouz, Ernesto William De Luca, Sahin Albayrak. Latent Semantic Social Graph Model for Expert Discovery in Facebook. IICS 2011, pp. 128- 138, 2011.
27. Leacock C. and M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet: An Electronic Lexical Database, C. Fellbaum, MIT Press, 1998.
28. Mohammed Erritali, Abderrahim Beni-Hssane, Marouane Birjali, Youness Madani. An Approach of Semantic Similarity Measure between Documents Based on Big Data, IJECE Vol. 6, No. 5, October 2016, pp. 2454-2461.