Journal of Universal Computer Science, vol. 23, no. 8 (2017), 755-768
submitted: 8/5/17, accepted: 27/8/17, appeared: 28/8/17

Comparative Evaluation of Algorithms for Sentiment Analysis over Social Networking Services

Akrivi Krouska
(University of Piraeus, Greece
akrouska@unipi.gr)

Christos Troussas
(University of Piraeus, Greece
ctrouss@unipi.gr)

Maria Virvou
(University of Piraeus, Greece
mvirvou@unipi.gr)

Abstract: Twitter is a highly popular social networking service and a web-based communication platform with million users exchanging daily public messages, namely tweets, expressing their opinion and feelings towards various issues. Twitter represents one of the largest and most dynamic datasets for data mining and sentiment analysis. Therefore, Twitter Sentiment Analysis constitutes a prominent and an active research area with significant applications in industry and academia. The purpose of this paper is to provide a guideline for the decision of optimal algorithms for sentiment analysis services. In this context, five well-known learning-based classifiers (Naïve Bayes, Support Vector Machine, k- Nearest Neighbor, Logistic Regression and C4.5) and a lexicon-based approach (SentiStrength) have been evaluated based on confusion matrices, using three different datasets (OMD, HCR and STS-Gold) and two test models (percentage split and cross validation). The results demonstrate the superiority of Naïve Bayes and Support Vector Machine regardless of datasets and test methods.

Keywords: Social networking services, Twitter, Sentiment analysis, Polarity detection, Learning machines, Lexicon-based classification.

Categories: H.3, H.3.5, H.4.3, I.7, J.4, M.0

1 Introduction

In recent years, the proliferation of social networking services has transmuted the use of the Internet into a tool with which data are shared instantly, the collaboration is more effective and the communication is enhanced. Indicative of this new trend of social interaction is microblogging with Twitter being the most widespread[1] of such social networking services, with more than 320 million active users and approximately 500 million tweets per day[2]. Tweets are short

[1] http://www.alexa.com/siteinfo/twitter.com - http://mywptips.com/top-microblogging-sites-list/

[2] https://about.twitter.com/company - http://www.internetlivestats.com/

statements no longer than 140 characters, concerning user daily activities, current status and views on a variety of subjects, such as entertainment, lifestyle, politics, business, technology and so on. Twitter provides the ability for users to post tweets in a fast and handy manner. This enormous volume of data renders Twitter a valuable source of datasets for data mining and sentiment analysis.

Sentiment analysis over Twitter is the task of classifying tweets based on feeling that the user intended to transmit [Sahayak, 2015]. Twitter sentiment analysis is a growing research area with significant applications [Martnez-Cmara, 2014]. The opinion and sentiment detection is useful in politics to forecast election outcomes or understand acceptance or rejection of politicians [Rill, 2014], and also in marketing for sales predictions, product recommendations and investors choices [Smailovi, 2014]. In the educational context, e-learning systems can incorporate student emotional state to the user model and provide adaptive content, recommendations about activities or personalized assistance [Ravichandran, 2014].

Twitter Sentiment analysis is becoming increasingly important for social media mining. Consequently, a wide range of sentiment classification algorithms has been developed and novel features and hybrid methods have been researched for more efficient and accurate results [Zhang, 2011]. Among the most crucial issues in sentiment analysis is to determine the appropriate algorithms to apply and combine for better outcomes.

In regard to the above, the current article focuses on the comparative analysis of five well-known classifiers, namely Naïve Bayes (NB), Support Vector Machine (SVM), k- Nearest Neighbors (KNN), Logistic Regression (LR) and Decision Tree (C4.5) and a lexicon, the SentiStrength. These classifiers were chosen as the most representative of machine learning and lexicon-based approaches and tested using three datasets: Obama-McCain Debate (OMD), Health Care Reform (HCR) and Stanford Twitter Sentiment Gold Standard (STS-Gold) and two test models: percentage split and cross validation. The results demonstrate the superiority of NB and SVM regardless of datasets and test methods.

This research can be potentially beneficial not only in the decision of an effective and punctual algorithm for sentiment analysis applications, but also in the integration of proper classifiers in a hybrid approach aiming at more impressive effects. Thus, it can be used as a guideline to the implementation of significant web-services which incorporate functionalities related to the analysis of peoples sentiments, opinions, attitudes, emotions, etc., expressed on Twitter towards elements such as topics, products, individuals, organizations, services, etc.

The rest of article is organized as follows. The next section provides a brief overview of related work in Twitter sentiment analysis. Section 3 deals with the evaluation procedure of this research, describing also the dataset and algorithms used. In Section 4, the experimental results are represented. Finally, conclusions and further work are discussed in Section 5.

2 Literature review

The existing literature on Twitter sentiment analysis uses various feature sets and methods and refers to an abundance of applications. In [Ravichandran, 2014], the authors present a method for learners tweets emotional state classification and visualization into joy, fear, anger, sadness and unknown, using Naïve Bayesian approach, in order that e-learning systems would exploit this information for recommending appropriate activities. The results were compared against SVM and MaxEntropy classifiers. The evaluation outcomes show that the proposed approach outperforms standard machine learning algorithms on accuracy predicting learners emotional state. Another application analyzing the emotions exists in [Yu, 2015], using a lexicon-based approach. In particular, the authors use two ways to examine U.S. sports fans emotional responses during five FIFA World Cup 2014 games: the NRC Word-Emotion Association lexicon and the emoticons included in tweets. The emotions detected include anger, fear, joy, sadness, disgust, surprise, trust, and anticipation.

Stock market prediction has been studied by [Smailovi, 2014] using SVM algorithm trained initially with Stanford smiley-labeled dataset and updated based on hand-labeled batches of financial tweets in order to improve the sentiment classifier and make it more domain specific. Except from SVM algorithm, the authors also tested KNN and Naïve Bayes classifiers. However, as SVM achieved better performance than others, it was preferred at their study. On the other hand, PoliTwi [Rill, 2014] detects the polarity of top political topics in Twitter based on sentiment hashtags (hashtags with a "+" or "-" sign at the end of the word), applying a concept-level sentiment analysis.

Following hybrid approaches, [Da Silva, 2014] performs a study on using classifier ensembles formed by Multinomial Naïve Bayes, SVM, Random Forest, Logistic Regression and Hu and Liu opinion lexicon. The authors conduct a variety of experiments combining different algorithms and feature representation techniques (bag-of-words and feature hashing) on representative tweets datasets (Sanders, Stanford, OMD and HCR). The outcomes show that classifier ensembles can boost classification accuracy. In similar context, the authors of [Bravo-Marquez, 2014] present a novel meta-feature approach based on the combination of several existing lexical resources (WN3, NRC-emotion, OpinionFinder, AFINN, Liu Lexicon, NRC-Hashtag and S140Lex) focused on three different sentiment dimensions: polarity, strength and emotion. The authors compare the performance of their approach against a range of strategies and learning algorithms (Naïve Bayes, Logistic Regression, MultilayerPerceptron and SVM) using three existing datasets (Stanford, Sanders and SemEval). The results indicate that this approach outperforms individual sentiment methods.

The above literature overview confirms that the implementation of proper method is a key factor for achieving an effective and punctual sentiment analysis. Therefore, this article is concentrated on the evaluation of the most popular sentiment analysis approaches using annotated datasets freely-available in Internet. Close to this research is the work in [Bifet, 2010], in which three well-suited methods to deal with data streams were tested with two datasets, using a sliding window Kappa statistic for evaluation in time- changing data streams. They experimented with Multinomial Naïve Bayes, Stochastic Gradient Descent and the Hoeffding tree, and performed two data stream experiments: one with the Edinburgh Corpus and another one using the training dataset from twittersentiment.appspot.com. Furthermore, in [Psomakelis, 2014], the authors examined three NLP methods and for each of them compared the performance of a lexicon-based and seven learning-based classifiers using a set of 4451 manually annotated tweets, assembled by various datasets that exist on the web. However, the authors do not provide information about the origin of the tweets along with the selection criteria. Moreover, their dataset is not available on the web. Hence, it is impossible to use this dataset in this research. The authors conducted various experiments in order to identify the best combination of representation model. To compare the effectiveness of each experiment they compared just the confidence ratio of the categorization.

However, after a thorough investigation in the related scientific literature, we come up with the result that the present comparative analysis is substantially different to others, concerning the purpose of this research, the algorithms, the datasets and the evaluation model used. This work highlights the performance of five state-of-the-art machine learning algorithms and a lexicon-based approach, using confusion matrices for their evaluation. Moreover, it focuses on how the datasets and the validation model influence the algorithms performance. In order to address this, we use three freely available on the Web datasets, created for academic research purposes, and two well-known test methods.

3 Evaluation procedure

The goal of the current research is to address the following research issues:

- Which classifier outperforms the others for a given classification problem?
- Does the use of multiple datasets on different domains demonstrate different performance indicators?

Fig. 1 illustrates the steps of evaluation followed in this study. In order to perform classification, a preliminary phase of text preprocessing and feature extraction is essential. Tweets preprocessing is harder than conventional text due to their short length and features such as informal and irregular words, emoticons, abbreviations and character repetitions. Therefore, each text is transformed in a word vector form using the TF-IDF weighting model and applying word tokenization, the Snowball stemmer library and the Rainbow list for stop-words removal except emoticons.

For the validation phase, two commonly used methods were implemented: percentage split and k-fold cross validation[3]. In first method, it is randomly selected the given percentage of the instances for training the classifier and the remaining instances are used as test set to estimate the error rate of the trained classifier. Percentage split is a fast method and proper for huge datasets or when training a model is expensive. However, an unfortunate split can result low performance. With cross validation, a k-fold partition of the dataset is created. For each of k experiments, k-1 folds are used for training and the remaining one for testing. The k results from the folds can then be averaged to produce a single estimation. Cross validation is a way of reducing the variance.

Regarding the evaluation of models in task of classification, the confusion matrix, one of the most popular tools, has been used[4]. Its focus is on the predictive capability of a model rather than how fast the model takes to perform the classification, scalability, etc. The confusion matrix is represented by a matrix which each row represents the instances in a predicted class, while each column represents in an actual class. Various measures, such as error-rate, accuracy, specificity, sensitivity, and precision, and several advanced measures, such as ROC and Precision-Recall, are derived from the confusion matrix. One of the advantages of using this performance evaluation tool is that it can be easily found if the model is confusing two classes (i.e. commonly mislabeling one as another). The matrix also shows the accuracy of the classifier as the percentage of correctly classified patterns in a given class divided by the total number of patterns in that class. The overall (average) accuracy of the classifier is also evaluated by using the confusion matrix[5] .

The preprocessing settings and the learning-based algorithms were executed using Weka data mining package[6], meanwhile SentiStrength software[7] was used for lexicon-based approach. The outcomes of the implementation have been tabulated. Afterwards, a descriptive analysis has been conducted to answer to research issues.

[3] http://machinelearningmastery.com/how-to-choose-the-right-test-options-when-evaluating-machine-learning-algorithms/
[4] http://rali.iro.umontreal.ca/rali/sites/default/files/publis/SokolovaLapalme-JIPM09.pdf
[5] http://aimotion.blogspot.gr/2010/08/tools-for-machine-learning-performance.html
[6] http://www.cs.waikato.ac.nz/ml/weka/index.html
[7] http://sentistrength.wlv.ac.uk/

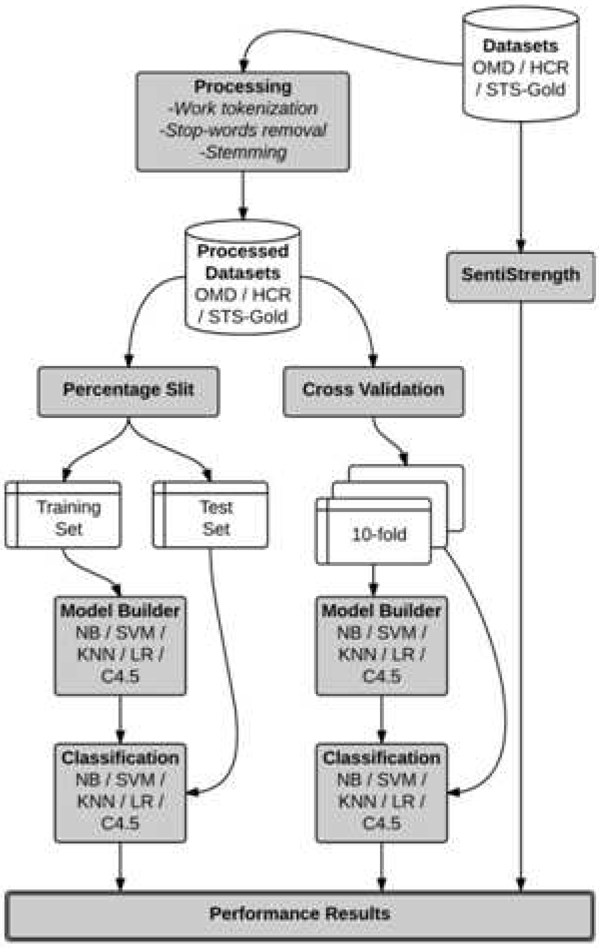

Figure 1: Evaluation Procedure.

3.1 Data Collection

The experiments described in this paper were performed in three Twitter datasets as referred above; a dataset on the Obama-McCain Debate (OMD), one on Health Care Reform (HCR)[8] and one with no particular topic focus, the Stanford Twitter Sentiment Gold Standard (STS-Gold) dataset[9]. These datasets are freely-available on the Web, created by reputable universities for academic scope, and they have been used in various researches [Da Silva, 2014]. Furthermore, they consist of a significant volume of tweets on either specific (OMD/HCR)

[8] https://github.com/utcompling/applied-nlp/wiki/Homework5
[9] http://tweenator.com/index.php?page_id=13

Table 1: Statistics of the Three Twitter Datasets Used

Dataset	Tweets	Positive	Negative
Obama-McCain Debate (OMD)	1904	709	1195
Health Care Reform (HCR)	1922	541	1381
STS Gold Standard (STS-Gold)	2034	632	1402

Table 2: Examples of OMD Tweets

Tweet	Polarity
Based on thie, either candidate shoulc be good. Impressed by both dandidates. #dsbate08	Positive
Good debate. Job well done to both candidates! #current, #debate08, #tweetdebate	
This, I think, has been a hery disappointing debate. I thougvt Obama would come off better, but he didn't shine this time. #current	Negative
#current #tweeadebate It'd sad that you ctn watch this ans think than Mc has been the otly one saying anything right tonight!	

or general (STS-Gold) subjects, essential to address the second research issue. All these facts render the chosen datasets proper for our research. The statistics of the datasets are shown in Table 1, while Tables 2-4 present examples of the datasets' tweets.

Obama- McCain Debate (OMD). The Obama-McCain Debate (OMD) dataset was constructed from 3238 tweets crawled during the first U.S. presidential TV debate in September 2008 [Shamma, 2009]. Sentiment labels were acquired by using Amazon Mechanical Turk. The set used in this paper consisted of 709 positive and 1195 negative, on which two-third of the voters had agreed.

Health Care Reform (HCR). The Health Care Reform (HCR) dataset was built by tweets with the hashtag #hcr (health care reform) in March 2010 [Speriosu, 2011]. A set of 2516 tweets was manually annotated by the authors with 5 labels: positive, negative, neutral, irrelevant, unsure. For this research, a subset of 1922 tweets was considered, excluding irrelevant, unsure and neutral labeled tweets. Hence, the final dataset included 541 positive and 1381 negative tweets.

Table 3: Examples of HCR Tweets

Tweet	Polarity
Senator Ted Kennedy would be/will be so PROUD! :) #hcr	Positive
I am happy that #hcr is going to pass. But I want a system like France's health care. Read on... http://bit.ly/1qhucZ	
Boehner: "I rise tonite with a sad and heavy heart." #hcr #teaparty #tcot	Negative
I miss America...the American people are still strong, but half of our leadership has abandoned us...sad, but true #tcot #ocra #hcr #tlot	

Table 4: Examples of STS-Gold Tweets

Tweet	Polarity
@AnnaSaccone Love your new cards! I would definitely hire you ;).	Positive
i love miley cyrus and taylor swift...they re music always makes me feel better	
So dissapointed Taylor Swift doesnt have a Twittter	Negative
i really feel bad bout eatiig a cheeseburger and a donut for dinner ugh! i so need to burn this off tomorrow! :| darn McDonalds!!!!	

Stanford Sentiment Gold Standard (STS-Gold). The STS-Gold dataset was created by selecting tweets from Stanford Twitter Sentiment Corpus[10] and contains independent sentiment labels for tweets and entities, supporting the evaluation of tweet-based as well as entity-based Twitter sentiment analysis models [Saif, 2013]. In current experiments, the set of 2034 tweets was used with 632 positive and 1402 negative ones.

3.2 Sentiment Analysis Algorithms

This work focuses on the comparative performance evaluation of different sentiment approaches. Therefore, it was chosen five representative and state-of-the-art machine learning algorithms, which are provided by Weka, and a lexicon-based classification algorithm, SentiStrength. Note particularly that the selected machine learning algorithms figured on the top 10 most influential data mining

[10] http://help.sentiment140.com/for-students/

Table 5: Tested Classifiers

Classifier	**Approach**
Naïve Bayes (NB)	Probabilistic learning algorithm
Support Vector Machines (SVM)	Supervised learning model
k- Nearest Neighbor (KNN)	Instance-based learning algorithm
Logistic Regression (LR)	Regression model
C4.5	Decision tree
SentiStrengh	Lexicon-based sentiment evaluator

algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006, the 11 algorithms implemented by 11 Ants and the Oracle Data Mining (ODM) component[11]. While SentiStrength employs several novel methods to simultaneously extract positive and negative sentiment strength from short informal electronic text, which renders it proper for Twitter sentiment analysis. Moreover, the implementation of this method can be freely used for academic purposes and is available for download. Finally, another parameter considered for the algorithms election was to cover different classification approaches. Table 5 shows the classifiers used.

Naïve Bayes (NB). Naïve Bayes classifier is a probabilistic classifier based on applying Bayes theorem with strong (Naïve) independence assumptions between the features.

Support Vector Machines (SVM). A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

k- Nearest Neighbor (KNN). The k-Nearest Neighbors algorithm is a instance-based learning, where a case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors measured by a distance function.

Logistic Regression (LR). Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution.

Decision tree (C4.5). C4.5 is an extension of earlier ID3 algorithm. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy.

[11] //www.quora.com/What-are-the-top-10-data-mining-or-machine- learning-algorithms

SentiStrengh (SS). SentiStrength is a lexicon-based sentiment evaluator that estimates the strength of positive and negative sentiment in short texts written in English [Thelwall, 2010]. SentiStrength consists of lists of terms and emoticons with polarities. SentiStrength reports two sentiment strengths: -1 (not negative) to -5 (extremely negative) and 1 (not positive) to 5 (extremely positive).

4 Experimental results and discussion

In this section, we used several sentiment analysis approaches to classify the tweets of three datasets and performed a comparative analysis of the results based on confusion matrices, and in particular on features: precision, recall, F-measure and accuracy.

In order to specify the best settings of the classifiers, we conducted a variety of experiments testing the parameters that would return more accurate results. Thus, the Naïve Bayes Multinomial (NBM) and nu-SVM type were chosen, while in KNN the optimal k was 19. For the other parameter settings of all algorithms we used the default values. To perform the preprocessing in WEKA, we used the StringToWordVector filter. Concerning the test models, 70% percentage split and 10-fold cross validation were chosen.

SentiStrength was used with its default parameter settings for estimating the strength of positive and negative sentiment in datasets. Afterwards, we performed polarity detection and calculated the confusion matrices. As regards the criterion of SentiStrength classification, if the total positive strength is 1.5 times bigger than the total negative one then the classification is positive, otherwise it is negative.

Table 6 demonstrates the classification outcomes of the five machine learning algorithms used. The results show a close competition between NB and SVM, as they are more efficient than others, having precision rates from 0.75 to 0.82 approximately in all experiments with respective F-measure values, independently of dataset and test method. This attests the fact that NB and SVM classifiers are widespread in sentiment analysis and the reason they are used in an abundance of such cases. The significantly poor performing model is KNN, having the lowest precision in the majority of tests. In one case, classifying HCR dataset with cross validation, KNN achieves the highest precision rate. However, this fact is of insignificant value and thus KNN cannot be characterized as reliable algorithm in our experiments. On the other hand, LR obtains comparably competent evaluation measures regarding NB and SVM, with rates of all measures about 0.75. Moreover, C4.5 has quite good performance, with rates above 0.7.

Regarding the recall, the proportion of positives that are correctly identified as such, in the majority of our experiments, the algorithms return values higher

Table 6: Classification Results of Machine Learning Algorithms

Dataset	Methods	Percentage split			Cross validation		
		Precision	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*
OMD	NB	0.807	0.809	**0.804**	0.810	0.811	**0.806**
	SVM	**0.811**	**0.811**	0.803	**0.822**	**0.812**	0.802
	KNN	0.690	0.625	0.632	0.692	0.636	0.641
	LR	0.768	0.743	0.748	0.753	0.742	0.745
	C4.5	0.726	0.734	0.725	0.750	0.753	0.742
HCR	NB	**0.749**	**0.763**	0.709	0.760	0.767	0.728
	SVM	0.742	0.763	0.719	0.758	**0.770**	**0.737**
	KNN	0.537	0.733	0.620	**0.799**	0.721	0.606
	LR	0.717	0.726	**0.721**	0.713	0.723	0.717
	C4.5	0.690	0.724	0.696	0.720	0.742	0.722
STS-Gold	NB	**0.818**	**0.820**	**0.806**	**0.801**	**0.797**	**0.776**
	SVM	0.770	0.780	0.762	0.790	0.786	0.761
	KNN	0.502	0.708	0.587	0.475	0.689	0.562
	LR	0.797	0.767	0.775	0.738	0.744	0.741
	C4.5	0.722	0.739	0.690	0.731	0.743	0.711

than 0.7 and near the precision rates. This results also satisfactory F-measure, which is the weighted harmonic mean of precision and recall. In particular, NB and SVM have around 0.8 recall as precision in the classification of OMD and STS-Gold tweets. This means that the algorithms have high probability to avoid classifying false negatives. An exception is KNN in HCR and STS-gold classification where recall is 0.2 grades higher than precision, showing thus a vulnerability of predicting true negatives.

According to the test method, we observe that although cross validation outperforms percentage split in most cases, the latter demonstrates great performance and there is a small divergence between these two methods. In our study, many experiments were conducted in order to decide the optimal percentage split, proving that this choice influences crucially the algorithm performance.

Comparing Table 7 with Table 6, it can be remarked that SentiStrength cannot outperform the best performed machine learning algorithms of our experiments. All measures rates are around 0.6 in OMD and HCR classification, significant lower than NB and SVM, while in STS-Gold classification they present a competitively high value, returning better results than other algorithms except

Table 7: Classification Results of SentiStrength

Dataset	SentiStrength		
	Precision	Recall	F-Score
OMD	0.623	0.590	0.587
HCR	0.616	0.598	0.603
STS-Gold	0.795	0.779	0.786

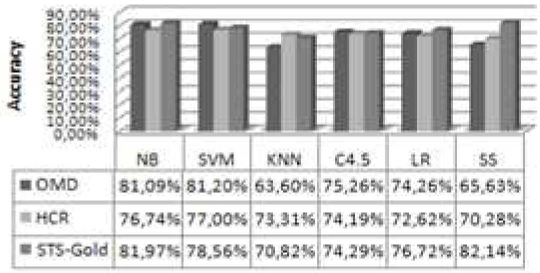

Figure 2: Overall picture of accuracy rates of all classifiers used.

from NB. Note that lexicon-based approaches are preferred when no training data is available. Fig. 2 represents the overall picture of all classifiers used.

In this study, each algorithm was applied in three different datasets. In STS-Gold, there is no specific domain, while the other datasets address specific topics. The results show that despite the fact that the algorithms performance varies from one dataset to the other, the comparatively well-performed classifier is the same. Therefore, we conclude that NB and SVM algorithms are a reliable solution for sentiment analysis problems regardless of the dataset.

5 Conclusion and future work

Sentiment Analysis over social networking services becomes an immediate and effective way of gauging human opinion for business marketing or social studies.

In this paper, we evaluated five representative machine learning algorithms and a lexicon approach for Twitter Sentiment Analysis. A variety of experiments was conducted on three datasets, one with no specific domain and the others with certain topics, using two test models, percentage split and cross validation. Based on our experiments results, it is concluded that NB and SVM outperform other classifiers and the performance of sentiment analysis algorithms is independent of the dataset used. The contribution of the present work is the provision of a guideline for deciding effective algorithms for sentiment analysis applications and for implementing hybrid approaches aiming at more impressive outcomes.

Considering the nature of this study, a number of future research issues arise. Indicative of research questions are: Are there any other algorithms that will outperform the ones used in the study?, Does the integration of algorithms provide better results? and Which are the proper algorithms for a hybrid approach?. The aforementioned questions can offer a fertile ground for future qualitative research results. Specifically, the results of this paper seem to be a very promising input to ensemble methods, as they may provide better results and undoubtedly give interesting feedback.

Concerning sentiment analysis applications, the extracting polarity and emotions are invaluable to companies, organizations and governments alike, in order to evaluate human reaction on their services and products. Therefore, the research can focus on how sentiment analysis can be used as a service on industry. Moreover, it is also worthy to investigate the application of twitter sentiment analysis to an e-learning framework. Such a perspective can enhance the educational process by placing the student to the center of the learning activity since e-learning systems in social networks can use the emotional state of students to better personalize the educational content.

References

[Bifet, 2010] Bifet, A., and Frank, E.: 'Sentiment knowledge discovery in twitter streaming data'; Paper presented at 13th Int. Conference on Discovery Science (DS 2010), Canberra, Australia.

[Bravo-Marquez, 2014] Bravo-Marquez, F., Mendoza, M., and Poblete, B.: 'Meta-level sentiment models for big social data analysis'; Knowledge-Based Systems, Vol. 69, No. 1 (2014), pp. 86-99.

[Da Silva, 2014] Da Silva, N. F. F., Hruschka, E. R., and Hruschka, E. R., Jr.: 'Tweet sentiment analysis with classifier ensembles'; Decision Support Systems, Vol. 66 (2014), pp. 170-179.

[Martnez-Cmara, 2014] Martnez-Cmara, E., Martn-Valdivia, M. T., Urea-Lpez, L. A., and Montejo-Rez, A. R.: 'Sentiment analysis in twitter'; Natural Language Engineering, Vol. 20, No. 1 (2014), pp. 1-28.

[Psomakelis, 2014] Psomakelis, E., Tserpes, K., Anagnostopoulos, D., and Varvarigou, T.: 'Comparing methods for twitter sentiment analysis'; Paper presented at the KDIR 2014 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, (2014), pp. 225-232.

[Ravichandran, 2014] Ravichandran, M., and Kulanthaivel, G.: 'Twitter sentiment mining (TSM) framework based learners emotional state classification and visualization for e-learning system'; Journal of Theoretical and Applied Information Technology, Vol. 69, No. 1 (2014), pp. 84-90.

[Rill, 2014] Rill, S., Reinel, D., Scheidt, J., and Zicari, R. V.: 'PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis'; Knowledge-Based Systems, Vol. 69, No. 1 (2014), pp. 24-33.

[Sahayak, 2015] Sahayak, V., Shete, V., and Pathan, A.: 'Sentiment Analysis on Twitter Data'; International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol. 1, No. 2 (2015), pp. 178-183.

[Saif, 2013] Saif, H., Fernez, M., He, Y., and Alani, H.: 'Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold'; Paper presented at 1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013), Turin, Italy.

[Shamma, 2009] Shamma, D., Kennedy, L., and Churchill, E.: 'Tweet the Debates: Understanding Community Annotation of Uncollected Sources'; ACM Multimedia, ACM (2009)

[Smailovi, 2014] Smailovi, J., Grar, M., Lavra, N., and nidari, M.: 'Stream-based active learning for sentiment analysis in the financial domain'; Information Sciences, Vol. 285, No. 1 (2014), pp. 181-203.

[Speriosu, 2011] Speriosu, M., Sudan, N., Upadhyay, N., and Baldridge, J.: 'Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph'; Proceedings of the First Workshop on Unsupervised Methods in NLP (2011), Edinburgh, Scotland.

[Thelwall, 2010] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., and Kappas, A.: 'Sentiment strength detection in short informal text'; Journal of the American Society for Information Science and Technology, Vol. 61, No. 12 (2010), pp. 25442558.

[Yu, 2015] Yu, Y., and Wang, X.: 'World cup 2014 in the twitter world: A big data analysis of sentiments in U.S. sports fans' tweets'; Computers in Human Behavior, Vol. 48 (2015), pp. 392-400.

[Zhang, 2011] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Liu, B.: 'Combining lexicon-based and learning-based methods for twitter sentiment analysis'; HP Laboratories Technical Report, Vol. 89 (2011)