Global Journal of Computer Sciences: Theory and Research

Volume 7, Issue 3, (2017) 153-159

Discovering future of the social trends using social media tools

Omer Sevinc*, Ondokuz Mayis University Computer Programming Department, Vezirkopru MYO Taskale Mah., Universite Cad. No :2 Vezirkopru-SAMSUN.

Iman Askerbeyli, Ankara University Computer Engineering Department, Golbası 50.yıl Yerleskesi Bahcelievler Mh, I Blok Golbasi/ANKARA

Serdar Mehmet Guzel, Ankara University Computer Engineering Department, Golbasi, 50.yıl Yerleskesi Bahcelievler Mh, I Blok Golbasi/ANKARA

Suggested Citation:

Sevinc, O., Askerbeyli, I. & Guzel, S.M. (2017). Discovering future of the social trends using social media tools. *Global Journal of Computer Sciences: Theory and Research. 7*(3),153-159.

Received July 13, 2017; revised October 21, 2017; accepted December 1, 2017.

Selection and peer review under responsibility of Prof. Dr. Dogan Ibrahim, Near East University, North Cyprus.

Abstract

Social media has been widely used in our daily lives, which, in essence, can be considered as a magic box, providing great insights about world trend topics. It is a fact that inferences gained from social media platforms such as Twitter, Faceboook or etc. can be employed in a variety of different fields. Computer science technologies involving data mining, natural language processing (NLP), text mining and machine learning are recently utilized for social media analysis. A comprehensive analysis of social web can discover the trends of the public on any field. For instance, it may help to understand political tendencies, cultural or global believes etc. Twitter is one of the most dominant and popular social media tools, which also provides huge amount of data. Accordingly, this study proposes a new methodology, employing Twitter data, to infer some meaningful information to remarks prominent trend topics successfully. Experimental results verify the feasibility of the proposed approach.

Keywords: Social web mining, Tweeter analysis, machine learning, text mining, natural language processing.

*ADDRESS FOR CORRESPONDENCE: **Sevinc Omer,** Ondokuz Mayis University Computer Programming Department, Vezirkopru MYO Taskale Mah. Universite Cad. No :2 Vezirkopru-SAMSUN.

E-mail address: sevinc.omer@gmail.com

1. Introduction

Social media environment is a prominent era to put the feelings and ideas bravely. Somehow people put their characters and make comments easier than their normal lives because they can mention their ideas by using social media materials like re-tweets. However billions of people use social media all around the world. According to recent statistical studies, the number of social network users is expressed in billions and increasing day by day [1]. Scholars, advertisers and political activists see massive online social networks as a representation of social interactions that can be used to study the propagation of ideas, social bond dynamics and viral marketing, among others [2]. Consequently, this makes a huge library to infer valuable information all around the world on many different fields from politics to science. Under the concepts of opinion mining, sentimental analysis and clustering inferences can be fetched [3]. Analyzes gained from social media mining may ease to predict about nature events, health, and politics etc. In this study, Twitter is utilized to handle trend topics around the world or on any specific subject. It is the most popular text-messaging service to share ideas, which also allows users to analyse messages using specific tools. One way to describe Twitter is as a micro-blogging service that allows people to communicate with short, 140-character messages that roughly correspond to thoughts or ideas. In that regard, you could think of Twitter as being akin to a free, high-speed, global text-messaging service. In other words, it's a glorified piece of valuable infrastructure that enables rapid and easy communication [4]. Twitter does not require people to be friends but they just need to share similar topics under hash tags. It has a decentralized structure it is not a friends graphs but an interest graph, which provides better possibilities for data mining realm. Twitter's decentralized structure makes hash tags as labels and people come together around these labels independently from each others. Accordingly, in this study, it is preferred to employ Twitter shares as web source to analyze and make inferences by using statistical and machine learning methods.

2. Previous Studies

There many studies on Twitter mining because it gives us insight what is going on around the world and more that what is going to happen next. Hence Twitter data should be analyzed with many aspects. For instance together with tweets, it can be considered which mentioned the same people, replies and re-tweets. However if topic derivation is done through a two-step matrix factorization process it can be conducted using a number of experiments on several Twitter datasets to reveal both the individual and integrated effects of the various features being considered [5]. In another study, it is focused on that when an earthquake occurs, people make many Twitter posts related to the earthquake, which enables detection of earthquake occurrence promptly, simply by observing the tweets. So accordingly the real-time interaction of events such as earthquakes in Twitter is investigated and an algorithm is proposed to monitor tweets and to detect a target event. To detect a target event, a classifier of tweets is devised based on features such as the keywords in a tweet, the number of words, and their context. Subsequently, they produce a probabilistic spatiotemporal model for the target event that can find the centre and the trajectory of the event location. [6]. Alternatively, Twitter data can be used for understanding diffusing of an illness or where a virus emergences. In a comprehensive study author present a more general approach that discovers many different ailments as well as can learn symptom and treatments obtained from tweets. To create structured information from the data, they develop a new topic model that organizes health terms into ailments, including associated symptoms and treatments [7]. Another study makes predictions over stock markets depending on societies' mood by analysing the Twitter texts. The study claim that the economics tells us those emotions can profoundly affect individual behaviour and decision-making. This also applies to societies at large, i.e. can societies experience mood states that affect their collective decision making. By extension the public mood is correlated or even predictive of economic indicators. The study investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. They analyse the text

content of daily Twitter feeds by two mood tracking tools, namely "Opinion-Finder" that measures positive vs. negative mood and "Google-Profile" of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). The results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others [8]. Another study subject on Twitter data is sentiment analysis.

Micro blogging has recently become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life every day. Therefore, micro blogging web-sites are rich sources of data for opinion mining and sentiment analysis. Because micro blogging has appeared relatively recently, there are a few research works that were devoted to this topic. In this paper, author focuses on employing Twitter, the most popular micro blogging platform, for the task of sentiment analysis. It is showed that how to automatically collect a corpus for sentiment analysis and opinion mining purposes. We perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, a sentiment classifier is built that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that proposed techniques are efficient and perform better than previously proposed methods. In this research, it is worked with English sentences; however, it is claimed that the proposed technique can be used with any other language [9].

3. Methods

In this study five different machine learning algorithms are used to classify the data obtained from tweets to decide whether the reviews are positive or negative. One of the method is naive bayes used in similar studies [10]; which can successfully make text classification especially with correct parameter optimizations [11]. Navie Bayes is a Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent given class, that is $P(X \mid C) = \prod_{i=1}^{n} P(X_i \mid C)$, where $X = (X_1 \ldots\ldots\ldots X_n)$ is a feature vector and is a class. Despite this unrealistic assumption, the resulting classifier known as naive Bayes is remarkably successful in practice, often competing with much more sophisticated techniques Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management [12]

Support Vector Machine SVM; is a learning machine used as a tool for data classification, function approximation, etc, due to its generalization ability and has found success in many applications [13]. Feature of SVM is that it minimizes and upper bound of generalization error through maximizing the margin between separating hyper plane and dataset. SVM has an extra advantage of automatic model selection in the sense that both the optimal number and locations of the basis functions are automatically obtained during training. The performance of SVM largely depends on the kernel [14]

Random Forest classifier; consists of a combination of tree classifiers where each classifier is generated using a random vector sampled independently from the input vector, and each tree casts a unit vote for the most popular class to classify an input vector [15].

Decision tree classifier; recursively partition the instance space using hyper planes that are orthogonal to axes. The model is built from a root node which represents an attribute and the instance space split is based on function of attribute values (split values are chosen differently for different algorithms), most frequently using its values. Then each new sub-space of the data is split into new sub-spaces iteratively until an end criterion is met and the terminal nodes (leaf nodes) are each assigned a class label that represents the classification outcome (the class of all or majority of the instances contained in the sub-space) [16].

KNN is one of the most widely used lazy learning approaches. Given a set of n training examples, upon receiving a new instance to predict, the kNN classifier will identify k nearest neighbouring training examples of the new instance and then assign the class label holding by the most number of neighbours to the new instance [17]. KNN is a typical supervised algorithm. Reducing or eliminating

statistical redundancy between the components of high-dimensional vector data enables a lower-dimensional representation without significant loss of information. Recognizing the limitations of principal component analysis (PCA), researchers in the statistics and neural network communities have developed nonlinear extensions of PCA [18]

The other method is commonly used in web page and text classifications is the support vector machine SVM which is used in similar studies [19] and mostly gives the best results especially in bigger data sets [20]. In this study SVM gives the highest F1 score. The other method that is used in this study is the random forest approach which gives optimal results on classification and regression [21,12]. KNN and decision trees algorithms are also commonly used methods which are used in text classifications [23,24,25 and 26].

In this study as illustrated in Fig. 1, the tweets are handled from tech review corpus which has tweet id's, topics and labels. Labels are one of the two choices of negative (0), positive (1). After tweets are collected with tweet id's the pre-processing applied to the all text and features are extracted. According to the relevant words the bag of words are created to apply to all tweets. Then the tweets are separated into train and test data out of 1000 tweets. In the pre-processing stage regular expression and Python "nltk" library is used. In one case the maximum feature dimensionality reduction is applied to all methods and in the other case PCA is additionally applied for dimensionality reduction to train and test data set to prevent from being sparse matrix of data set. In both cases performance result are handled and compared. 20% of the tweets are selected randomly for test set and rest of the tweets are selected as training set. The work flow of our study can be seen on Figure 1. In this study 5 main methods in machine learning applied to the data set and models are trained. The first model naive bayes applied and then the other main method are applied sequentially and performance results are handled to evaluate which model best fit and gives the better results.

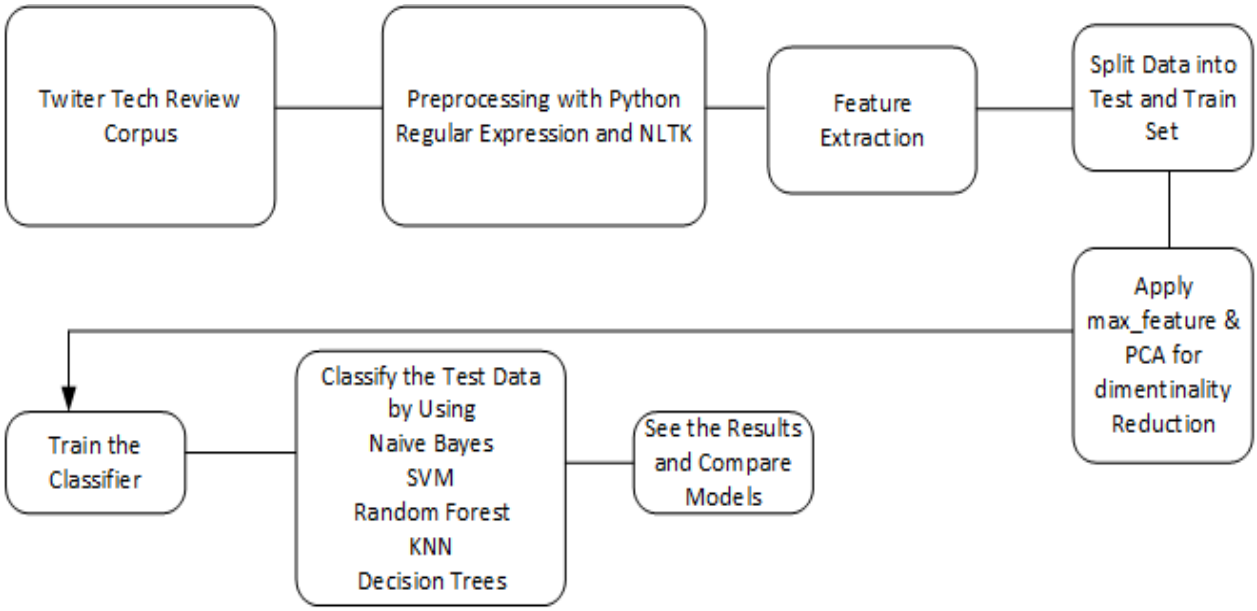

Figure 1. The workflow schema of our study.

For calculating the performance criteria values the accuracy, precision recall and f1 scores are calculated over true positive TP, true negative TN, false positive FP, and false negative FN prediction values of the review sentiments. The formulizations are as written below.

accuracy = (TP + TN) / (TP + TN + FP + FN) (1)

precision = TP / (TP + FP) (2)

recall = TP / (TP + FN) (3)

f1_score = 2 * precision * recall / (precision + recall) (4)

4. Results & Findings

Here for the evaluation purpose accuracy, precision, recall and f1 score values are calculated. According to the result that can be seen on Table-1 and Table-2 the predictions are made in two different cases. In the first case the maximum feature dimensionality reduction is applied and performance results are calculated and then in the second case the additionally the PCA dimensionality reduction is also applied to data set and performance values are calculated from scratch. In the first case the best results are handled with naïve bayes method as can be seen on Table-1. On the other hand when dimensionality reduction is applied the best results are handled with SVM method which also has the best F1 score. So it can be concluded that the sentiment analysis can be made over at least % 71 success prediction rates. This shows that the methods can be applied for sentiment analysis to tweets and reviews or similar text content and good results can be handled to predict future trends through these indicator results. The results demonstrate that dimensionality reduction had impact more on naive bayes and SVM than the other methods. Similar results are handled with and without PCA for random forest, decision trees and kNN because they are not heavily depended on dimensionality. As random forest itself already performs a fair regularization without assuming linearity, it is not necessarily an advantage on this method and similar for the others.

Table 1. Scores of the applied machine learning techniques Only with max_Features

Methods	Legends			
	Accuracy	*Precision*	*Recall*	*F1 Scores*
Naïve Bayes	0.73	0.684210526316	0.883495145631	0.771186440678
SVM	0.72	0.752688172043	0.679611650485	0.714285714286
Random Forest	0.685	0.77027027027	0.553398058252	0.64406779661
KNN	0.61	0.676056338028	0.466019417476	0.551724137931
Decision Trees	0.71	0.747252747253	0.660194174757	0.701030927835

The results are shown when five different methods are applied to same data set in different cases. PCA is an effective dimentionality reduction technique. On the other hand the linear maximum feature is also a prominent dimentionality reduction technique [27] that give efficient results.

Table 2. Scores of the applied machine learning techniques additionally using PCA

Methods	Legends			
	Accuracy	*Precision*	*Recall*	*F1 Scores*
Naïve Bayes	0.505	0.54	0.262135922	0.352941176471
SVM	0.72	0.752688172043	0.679611650485	0.714285714286
Random Forest	0.685	0.77027027027	0.553398058252	0.64406779661
KNN	0.61	0.676056338028	0.466019417476	0.551724137931
Decision Trees	0.71	0.747252747253	0.660194174757	0.701030927835

5. Conclusion & Future Works

As a conclusion it can be said that the small text data like tweets can be analyzed with preprocessing techniques and the feelings can be caught by machine learning methods with more then %71 success. The results which include feeling that positive or negative can be an indicator for

future plans and actions. Consequently, some further evaluation and prediction can be performed based on these results. This study demonstrates that naïve bayes and SVM methods work more successful and give better results than the other methods applied on the data set. The decision trees were another successful method that suit for the text classification of this review data set. However an important part in this study is the dimensionality reduction that makes the methods work fast and accurately. In the study the linear maximum features and PCA is used to be able to get good results but beside that some other techniques like linear discriminant analysis (LDA), self-organized map (SOM) and feature embeddings can also be applied to see their effects on the model. This study gives good classification results with sentiment analysis and works very efficiently. Although the neural network can be applied by implementing deep structured learning to give better results. Deep learning works with many hidden layers which give the power of understanding the myth behind input data so it can conclude more accurate results. So with different data sets much dimensionality reduction can be tested and then prominent machine learning models can be applied by combining or the new strong learning model that deep learning neural networks can be applied to make a more successful model. As a result in this study many prominent machine learning algorithms applied to the reviews to make a sentimental analysis and good results are handled. This study shows us that text contents like tweets can be analyzed with pre-processing and machine learning techniques to make future predictions on trends. The model works successfully and gives good accuracy results.

References

[1] Seker, S.E. (2014). *Sosyal Aglarda Akan Veri Madenciligi*. *YBS Ansiklopedi, 1*(1), 21- 26

[2] Huberman, B. A., Daniel M.R.,& Fang, W. (2008). *Social networks that matter: Twitter under the microscope*.

[3] Seker, S.E. (2015). *Cizge Teorisi (Graph Theory*), *YBS Ansiklopedi, 2*(2), 17-29

[4] Russell, M.A.(2013) *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* O'Reilly Media, Inc.

[5] Nugroho, R. (2016). Using time-sensitive interactions to improve topic derivation in twitter.World Wide Web 1-27.R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Sakaki, T., Makoto,O., & Yutaka,M. (1989). Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science.

[7] Paul, M.J., & Mark, D. A. (2012).Model for mining public health topics from Twitter. *Health, 11* (2012), 16-6.

[8] Bollen,J., Huina,M., & Xiaojun, Z. (2011). Twitter mood predicts the stock market. *Journal of computational science, 2*(1),1-8.

[9] Pak, A., & Patrick,P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc., 10,* 2010.

[10] McCallum, A., & Kamal, N. A.(1998). Comparison of event models for naive bayes text classification." *AAAI-98 workshop on learning for text categorization, 752*, .

[11] Rish, I. (2001). An empirical study of the naive Bayes classifier." *IJCAI 2001 workshop on empirical methods in artificial intelligence, 3*(22), IBM.

[12] Acir, N. A. (2006). Support vector machine classifier algorithm based on a perturbation method and its application to ECG beat recognition systems. *Expert systems with application New York, 31*, 150-158.

[13] Girosi, J.M., & Poqqio, T. (1995). Regularization theory and neural network architectures. *Neural computation Cambridge, 7,* 217-269.

[14] Kim, Sang-Bum, et al. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering, 18*(11), 1457-1466.

[15] BREIMAN, L. (1999). *Random forests—random features.* Technical Report 567, Statistics Department, University of California, Berkeley, ftp://ftp.stat.berkeley.edu/pub/users/ breiman.

[16] Polaka, I., Igor,T., & Arkady,B. (2010). Decision Tree Classifiers in Bioinformatics." Scientific Journal of Riga Technical University. *Computer Sciences, 42*(1), 118-123.

[17] Jiang, Y., and Zhi-Hua,Z. (2004). Editing training data for kNN classifiers with neural network ensemble. *Advances in Neural Networks–ISNN,* 356-361.

[18] Kambhatla, N., & Todd,K.L. (1997). Dimension reduction by local principal component analysis. *Neural computation, 9*(7), 1493-1516.

[19] Yu, H., Jiawei,H., & Kevin Chen-Chuan, C. (2002). PEBL: positive example based learning for web page classification using SVM." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.

[20] Sun, A., Ee-Peng,L.,& Ying,L. (2009). On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems, 48*(1), 191-201.

[21] Liaw, A., & Matthew,W. (2002). Classification and regression by randomForest. *R news 2*(3), 18-22

[22] Ham, J. et al. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing, 43*(3), 492-501.

[23] Shi, K. (2011). An improved KNN text classification algorithm based on density." Cloud Computing and Intelligence Systems (CCIS), 2011 *IEEE International Conference on. IEEE.*

[24] Wajeed, M.A., & Adilakshmi, T.(2011). Different similarity measures for text classification using KNN." Computer and Communication Technology (ICCCT), 2011 *2nd International Conference on. IEEE*.

[25] Gorunescu, F. (2008). Classification and Decision Trees. Data Mining. Springer Berlin Heidelberg, 2011. Vens, Celine, et al. "Decision trees for hierarchical multi-label classification. M*achine Learning, 73*(2), 185-214.

[26] Celine, et al. (2008). Decision trees for hierarchical multi-label classification. *Machine Learning, 73*(2), 185-214.

[27] Roweis, S.T., & Lawrence, K.S. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290, 5500* 2323-232