Human aggressiveness and reactions towards uncertain decisions

Farhan Bashir [1], Noman Ashraf [2], Ayyaz Yaqoob [1], Abid Rafiq [1], Raza Ul Mustafa [3, *]

[1]Department of Computer Science and Information Technology, University of Sargodha, Sargodha, Pakistan
[2]Department of Computer Science, University of Lahore, Pakpattan Campus, Pakpattan, Pakistan
[3]Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Sahiwal, Pakistan

ARTICLE INFO

ABSTRACT

Big data is a term that defines data sets are so large or complex that traditional data processing applications are inadequate. Data comes with different formats such as Multilanguage, structured, unstructured and emails. Challenges dealing with a huge amount of data includes analysis, search, sharing, transfer, and visualization. Not all the data collected contains useful information, so there's a need to refine this data in order to filter out the useful information. Tweets posted on Twitter are expressed as opinions. These opinions can be used for different purposes such as to take public views on uncertain decisions. These decisions have a direct impact on the user's life such as violations and aggressiveness are common causes. For this purpose, we have collected opinions on some popular decision taken in the past decade from Twitter. We have divided the tweet text into two classes; Anger (negative) and positive. We have proposed a prediction model to predict public opinions towards such decisions. We used Support Vector Machine (SVM), Naïve Bayes (NB) and Logistic Regression (LR) classifier for a text classification task. Furthermore, we have also compared SVM results with NB, LR. The research will help us to predict early behaviors and reactions of people before the big consequences of such decisions. Moreover, the results highlight the feasibility of using social media to predict public opinions.

1. Introduction

The Internet is providing all the services a normal user looking. Health, education, government, and business, all categories of modern life have been covered in the shape of the internet. The Internet provides connectivity between people and information publicly shared globally. Similarly, social media such as Facebook, Twitter, and YouTube are a platform to remain updated with current news and affairs. Through social media, people can share news, opinions and participate in activities being held online. Social Networking Sites (SNS) have been used for expressing opinions on different issues.

Twitter is a popular platform for sharing opinions. Most people preferred sharing their opinions on Twitter because the community on Twitter is more educated and well organized. A study on "Twitter stream" (Nawaz et al., 2017) revealed that despite the high level of noise, a major proportion of Twitter data contain informative content. Various studies have extracted information related to health, education and market trends (Liu and Zhang, 2012). Opinion mining techniques can be used for detection and extraction of subjective information from tweets (Adedoyin-Olowe et al., 2013). A method for detecting human aggressiveness is sentimental analysis (Jiang et al., 2011; Barbosa and Feng, 2010; Torunoğlu et al., 2013; Liu et al., 2012; Bravo-Marquez et al., 2013; Lee et al., 2014). It is used to extract subjective information from a tweet. Among various social platforms, Twitter was chosen for this study due to the following reasons (Mustafa et al., 2017a).

- Twitter is a platform where high profile individuals feel comfortable to share their views on any topic.
- It has a real-time conversation with all the breaking news shared on publicly for people reviews.
- The Twitter API for collecting data is easier to use than those of other social media such as Facebook, LinkedIn, and Tumbler.

In this work, we have collected user's tweets using the twitter streaming Application Programming Interface (API) (Twitter.com). We have classified tweets into two categories such as positive and negative. Then we have proposed a hypothesis model for future considerations. We used machine learning most popular algorithms for text classifications such as SVM, NB, and LR (Sebastiani, 2002). We have performed training and testing in WEKA (Hall et al., 2009). And finally, we have compared the efficiency of these algorithms to find the best among them in the proposed dataset. We believed that this research would be helpful for politicians, industries and for those people who need the quick and early response to such type of decisions.

The rest of the article is organized as follows: In Section 2, related work is discussed. The approach that we used to detect human anger in tweet text is listed in Section 3. Evaluation of the proposed approach and results are discussed in Section 4. Finally, a conclusion is drawn in Section 5.

2. Literature review

Machine learning, data mining and Natural Language Processing (NLP) all used together for the classifications of text documents widely. These three techniques also used to discover patterns from the electronic documents. Text mining is used to discover useful hidden information from the documents and deals with the operations like retrieval, classification (supervised, unsupervised and semi-supervised) and summarization (Khan et al., 2011).

There have been many efforts regarding text classifications in the past. Chodey and Hu (2016) have analyzed large data from clinics and try to find the clinical disorders. Saini and Kohli (2016) have used machine learning techniques for analysis of social network E-Health data. Fernandes and D'Souza (2016) have analyzed product value using sentimental analysis publicly given on Twitter. Both have worked to solve the problem of reading millions of reviews by a single user for a particular product, they have developed a model using reviews posted which gives product classification in term of positive, negative and neutral reviews.

In the same context, Barnaghi et al. (2016) used Twitter sentiments to predict event winner. They used Bayesian Logistic Regression (BLR). They manually labeled tweets into two categories positive and negative. A model proposed by them can be used to predict the winner of any event using sentiments. Kashyap et al. (2016) have worked on music lyrics to categorize the mood of individuals. They have used different text mining and data mining approaches to deal with such a problem. They have considered music associations, melody choice and music proposal as a feature to demonstrate the data. It is beneficial for predicting more accurate understanding of the music mood in the mood mapping process. Similarly, many studies have been found to investigate the online business trends using social data. Online business and larger company's worldwide used user feedback which has been given on social sites for the improvement of product and business need with the passage of time. The amount of text and information shared on Twitter in the form of tweets have valid information, and it can be used to track the progress of the product. They have categorized the data into categories such as positive and negative and used machine learning clustering algorithms to do so. They have found that the data available online can be used for the process of information extraction and it is beneficial for the companies to track the progress of their product and handy for future considerations (Žunić et al., 2016).

Rathore and Ilavarasan (2017) have collected Twitter data for the prediction of Pizza success after its launch. This type of methodologies can be used to predict the behavior of any user for a particular product. Rathore and his company have used R and NodeXL for analyzing tweets collected from Twitter. Furthermore, they have used different text mining, NLP and network analysis techniques to predict user behavior. Any company or food delivering company can use this sort of information for success and failure of the product. However, nobody has worked to analyze the behavior of an uncertain decision and their impact on human life before. This work proposed a methodology to analyze the pattern of human behaviors towards uncertain decisions. For that, we trained and evaluated three classifiers: SVM, NB, and LR. Besides, Section 4 presents a comparative analysis of classifiers which can be handy in the future to select an appropriate classifier for this task. Our proposed methodology saves time and cost for such huge public feedback posted daily on social networks.

3. Methodology

The solution we suggest involves Twitter data. Tweets collected with Twitter API. Our methodology consists of two steps: training and testing phases. Feature representation, tweets collection and classifier training come in training phase, while the testing phase has four phases tweets collection for testing, feature representation, hypothesis prediction and evaluation. The first two tasks (i.e., tweets collection and feature representation) are shared between the training and testing phase. Classifiers such as SVM, NB, and LR used in training and hypothesis. We have used WEKA tool for training and testing of the proposed methodology. Firstly, we divided the data sets into two parts, training data and secondly testing data. Our proposed solution emphasis on the uncertain decision, for that we have collected tweets from the well-known twitter trends which have been held in the past year. Hashtagify is a free platform which can provide the most used trends on any topic (Hashtagify.com). Decisions we considered for the collection of proposed methodology have been listed

in Table 1. We stored all the data in the SQL database. We collected 5000 tweets for each topic.

Table 1: Trends used to collect tweets from Twitter

#SaudiWomenCanDrive	#PanamaVerdict
#Brexit	#NoteBandi
#NoBanNoWall	#Muslimban

3.1. Data preprocessing

Pre-processing reshapes the data into the desired form. The data we collected is not purified for the process of classification, for this, we have applied data processing methodologies to transform the data into meaningful features. It involves tokenization (or featuring), feature weighting and data cleaning (removal of irrelevant features). Once the data is collected, the Uniform Resource Locator (URL) from the tweets were removed. Tweets only with an image or with a link but there was no textual information were also removed. Stop words also do not give any information about the topic and just create noise in the data so using stop word-list, they were also removed from the data. When data is pre-processed, it helps in saving classifier time while classifying. Collected tweets are further pre-processed with following steps:

3.2. Tokenization

Tokenization deals with the breaking of long text strings into substrings which may include phrases and words collectively known as tokens. Among two ways of tokenization (phrase and word tokenization), word-level tokenization is considered as more effective due to statistical significance. In this process, the sentence for instance " the financial crises created the conditions for the worst period of wage growth " was broken into tokens " *the, financial, crises, created, the, conditions, for, the, worst, period, of, wage, growth"*. The algorithms which are used to tokenize a sentence separate the tokens with whitespace and some are based on built-in dictionary. Text can be tokenized in two ways, by words (often called a bag of words) or phrases.

3.3. Feature weighting

A standard function to compute the weights is TF-IDF. TF-IDF scheme is based on two parts: TF and IDF. TF stands for term frequency which is used to counts the represented terms/tokens in a document. It can give a complete measure of term occurrence. IDF stands for inverse document frequency of a term in a collection of documents (Salton and Buckley, 1988).

3.4. Sentiment classification

Once we applied the pre-processing, we have data in a suitable format to apply classification algorithm on it. We have categorized the data into two classes. A data with false words labeled Negative and data with positive words labeled as Positive. A sample of tweets rows have been listed in Table 2. Different algorithms are available in this domain that can be used to train the classification task. Different experimental studies have been directed to analyze these methods for text categorization. As the result of these experiments, SVM, LR, and NB are observed to be very effective algorithms (Sebastiani, 2002).

Table 2: Tweets rows with labeled positive and negative

Positive Samples	Negative Samples
nawaz produced two letters of mr hamad bin jassim bin jaber althani as his defense in panama case [#PanamaVerdict]	can we forget the bundles of fake documents presented in supreme court during panama case [#PanamaVerdict]
british passports will be navy blue after brexit says home office interesting [#Brexit]	the financial crises created the conditions for the worst period of wage growth [#Brexit]
no muslim ban just ban entry from some muslim countries [#Muslimban]	#MuslimBan is not freedom of religion Deciding what a woman can do with her body and her life is not freedom at all [#Muslimban]
Kapil Sibal at Congress presser: On April 9, I showed you video related to Notebandi, showing field assistant Rahul Ratrekhar converting old currency into new notes. Later,Cabinet Secretariat released statement that he was expelled in 2017 on suspicion of converting old currency [#NoteBandi]	Bjp is likely to show sunny leone ji CD so that people will forget about GST andamp; NOTEBANDI [#NoteBandi]
its good news for Saudi women that they can drive,they can contest in municipal election and attend in to social media [#SaudiWomenCanDrive]	Imagine all the crashes now this is going to happen. And all that bad parking. #SaudiWomenDriving [#SaudiWomenCanDrive]

3.5. WEKA

To perform the desired task, we used WEKA. WEKA is an open source free software which has been used for various machine learning problems using data. It contains features which can be used for classifications, preprocessing, clustering, visualization, association rules, etc. Machine learning is nothing without giving artificial intelligence to your data. Machine learning methods are very similar to data mining algorithms. WEKA have a collection of machine learning algorithms which are applied on data to extract desired results from it. Standard 10-fold cross validation is employed to evaluate the performance of classifiers. Cross-validation technique is used for model validation, and it also evaluates the generalization of independent data set over statistical results that are provided by the model. In 10 cross-fold validation, the dataset is partitioned (randomly) into 10 sub-datasets. Out of 10, 1 sub-dataset is selected as a validation set for model testing and remaining 9 sub-datasets are used for model training. This process is repeated 10 times in total where each sub-dataset is

used exactly once as the validation set. Single estimation of the result is obtained by taking the average of 10 results (Choi et al., 2010). Three evaluation measures (precision, recall and f-measure) are used to evaluate the performance of the classifier. The reason to choose these three measures is that they can evaluate a category wise prediction of the classifier. The mathematical definition of these measures concerning positive class is defined as follows:

$$Recall\ (R) = \frac{no\ of\ CPP}{no\ of\ PE} \tag{1}$$

$$Precision = \frac{no\ of\ CPP}{no\ of\ PP} \tag{2}$$

$$F - measure = \frac{2*P*R}{P+R} \tag{3}$$

In Eqs. 1 and 2, CPP, PE, and PP stand for correct positive prediction, positive examples and positive predictions, respectively.

4. Results and discussion

The results of sentiment classification using SVM classification are given in Table 3. Precision, recall, and f-measure are approximately 89%, 90%, and 88% respectively. The accuracy of training and testing of validating of data is between a ranges of 88-90%. SVM has showed high precision and recall on #PanamaVerdict, #SaudiWomenCanDrive, and #NoBanNoWall.

Table 3: Results obtained using SVM classifier

Class	Precision	Recall	F-Measure
#Brexit	0.865	0.883	0.863
#PanamaVerdict	0.922	0.925	0.906
#NoteBandi	0.900	0.899	0.889
#SaudiWomenCanDrive	0.911	0.900	0.901
#MuslimBan	0.850	0.873	0.843
#NoBanNoWall	0.916	0.924	0.909

The results of sentiment classification using NB classification are given in Table 4. Precision, recall, and f-measure are approximately 85%. The accuracy of training and testing of validating of data is between a ranges of 85-86%.

Table 4: Results obtained using NB classifier

Class	Precision	Recall	F-Measure
#Brexit	0.817	0.821	0.819
#PanamaVerdict	0.866	0.856	0.862
#NoteBandi	0.899	0.889	0.879
#SaudiWomenCanDrive	0.866	0.867	0.865
#MuslimBan	0.798	0.824	0.809
#NoBanNoWall	0.875	0.880	0.878

The results of sentiment classification using LR are given in Table 5. Precision, recall, and f-measure are approximately 85%, 84%, and 85% respectively. The accuracy of training and testing of validating of data is between a ranges of 84-85%. High precision and recall is achieved for #NoBanNoWall and #PanamaVerdict. To compare the accuracy of the three classifiers, paired t-test (corrected) is performed in WEKA. Statistical paired t-test compares two datasets in which observations in one dataset can be paired with other datasets

observations. The main objective of this test is to investigate the statistical evidence that the mean difference from paired observations from two datasets on a particular outcome is significantly different from zero (Mustafa et al., 2017a).

Table 5: Results obtained using LR classifier

Class	Precision	Recall	F-Measure
#Brexit	0.814	0.763	0.784
#PanamaVerdict	0.881	0.872	0.876
#NoteBandi	0.860	0.822	0.844
#SaudiWomenCanDrive	0.857	0.850	0.855
#MuslimBan	0.855	0.854	0.856
#NoBanNoWall	0.890	0.891	0.891

Obtained results indicate that some differences in the accuracy of SVM, LR, and NB exist, as shown in Table 6. However, differences inaccuracies of classifiers are not statistically significant. SVM, NB and LR have been also used in various other classifications purposes. However, they have shown different accuracies on Urdu tweets. High precision is achieved for SVM and LR (91.9% and 92.7% respectively). On the other hand, high recall and f-measure was achieved for NB (Mustafa et al., 2017b).

Table 6: Accuracy comparison of three classifier

Class	Precision	Recall	F-Measure
SVM	0.894	0.900	0.885
NB	0.853	0.856	0.852
LR	0.855	0.842	0.851

5. Conclusion

Twitter is one of the most important social sharing platforms. Tweets posted on Twitter are expressed as opinions. These opinions can be used for different purposes such as to take public views on uncertain decisions. These decisions have a direct impact on users life such as violations and aggressiveness are common causes. We have collected tweets of such decisions and labeled the tweets into two categories such as anger (negative) and positive. We have used classifier algorithms such as SVM, NB, and LR for building models. We have also compared SVM results with NB, LR. The research is useful for predicting early behaviors and reactions of people before the big consequences of such decisions. In the future, we interested to build a tool which can work as a recommender system to classify tweets automatically into two categories such as anger and positive.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Adedoyin-Olowe M, Gaber MM, and Stahl F (2013). A survey of data mining techniques for social media analysis. Available online at:
https://arxiv.org/abs/1312.4617

Barbosa L and Feng J (2010). Robust sentiment detection on twitter from biased and noisy data. In the 23rd International Conference on Computational Linguistics: Posters, Association for Computational Linguistics, Beijing, China: 36-44.

Barnaghi P, Ghaffari P, and Breslin JG (2016). Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In the IEEE Second International Conference on Big Data Computing Service and Applications, IEEE, Oxford, UK: 52-57.
https://doi.org/10.1109/BigDataService.2016.36

Bravo-Marquez F, Mendoza M, and Poblete B (2013). Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, Chicago, Illinois: 1-9.
https://doi.org/10.1145/2502069.2502071

Chodey KP and Hu G (2016). Clinical text analysis using machine learning methods. In the 15th International Conference on Computer and Information Science, IEEE, Okayama, Japan: 1-6.
https://doi.org/10.1109/ICIS.2016.7550908

Choi Y, Jung Y, and Myaeng SH (2010). Identifying controversial issues and their sub-topics in news articles. In: Zeng D, Wei D, Chau M, and Wang F (Eds.), Pacific-Asia workshop on intelligence and security informatics: 140-153. Springer, Berlin, Germany.
https://doi.org/10.1007/978-3-642-13601-6_16

Fernandes R and D'Souza R (2016). Analysis of product Twitter data though opinion mining. In the IEEE Annual India Conference, IEEE, Bangalore, India: 1-5.
https://doi.org/10.1109/INDICON.2016.7839025

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, and Witten IH (2009). The WEKA data mining software: An update. ACM SIGKDD Explorations Newsletter, 11(1): 10-18.
https://doi.org/10.1145/1656274.1656278

Jiang L, Yu M, Zhou M, Liu X, and Zhao T (2011). Target-dependent twitter sentiment classification. In the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 1: 151-160.

Kashyap N, Choudhury T, Chaudhary DK, and Lal R (2016). Mood based classification of music by analyzing lyrical data using text mining. In the International Conference on Micro-Electronics and Telecommunication Engineering, IEEE, Ghaziabad, India: 287-292.
https://doi.org/10.1109/ICMETE.2016.65

Khan A, Baharudin B, and Khan K (2011). Sentiment classification from online customer reviews using lexical contextual sentence structure. In the International Conference on Software Engineering and Computer Systems, Springer, Pahang, Malaysia, 317-331.
https://doi.org/10.1007/978-3-642-22170-5_28

Lee H, Han Y, Kim K, and Kim K (2014). Sentiment analysis on online social network using probability model. In the 6th International Conference on Advances in Future Internet, Lisbon, Portugal: 14-19.

Liu B and Zhang L (2012). A survey of opinion mining and sentiment analysis. In: Aggarwal CC and Zhai C (Eds.), Mining text data: 415-463. Springer, Boston, USA.
https://doi.org/10.1007/978-1-4614-3223-4_13

Liu KL, Li WJ, and Guo M (2012). Emoticon smoothed language models for twitter sentiment analysis. In the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, Canada: 1678-1684.

Mustafa RU, Nawaz MS, Farzund J, Lali MI, Shahzad B, and Viger PF (2017a). Early detection of controversial urdu speeches from social media. Data Science and Pattern Recognition, 1(2): 26-42.

Mustafa RU, Nawaz MS, Lali MIU, Zia T, and Mehmood W (2017b). Predicting the cricket match outcome using crowd opinions on social networks: A comparative study of machine learning methods. Malaysian Journal of Computer Science, 30(1): 63-76.
https://doi.org/10.22452/mjcs.vol30no1.5

Nawaz MS, Bilal M, Lali MI, Ul Mustafa R, Aslam W, and Jajja S (2017). Effectiveness of social media data in healthcare communication. Journal of Medical Imaging and Health Informatics, 7(6): 1365-1371.
https://doi.org/10.1166/jmihi.2017.2148

Rathore AK and Ilavarasan PV (2017). Social media analytics for new product development: Case of a pizza. In the International Conference on Advances in Mechanical, Industrial, Automation and Management Systems, IEEE, Allahabad, India: 213-219.
https://doi.org/10.1109/AMIAMS.2017.8069214

Saini S and Kohli S (2016). Machine learning techniques for effective text analysis of social network e-health data. In the 3rd International Conference on Computing for Sustainable Global Development, IEEE, New Delhi, India: 3783-3788.

Salton G and Buckley C (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5): 513-523.
https://doi.org/10.1016/0306-4573(88)90021-0

Sebastiani F (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1): 1-47.
https://doi.org/10.1145/505282.505283

Torunoğlu D, Telseren G, Sağtürk Ö, and Ganiz MC (2013). Wikipedia based semantic smoothing for twitter sentiment classification. In the 2013 IEEE INISTA, IEEE, Albena, Bulgaria: 1-5.
https://doi.org/10.1109/INISTA.2013.6577649

Žunić E, Djedović A, and Đonko D (2016). Application of big data and text mining methods and technologies in modern business analyzing social networks data about traffic tracking. In the XI International Symposium on Telecommunications, IEEE, Sarajevo, Bosnia-Herzegovina: 1-6.
https://doi.org/10.1109/BIHTEL.2016.7775717