DOI 10.29042/2018-4071-4076

Helix Vol. 8(5): 4071- 4076

Estimating User's Social Behaviour by Analysing Online Tweet Pattern

[*1]Miss.Shabistan Ruhi, [2]Dr.Tausif Diwan

[1, 2] Department of Computer Science & Engineering, Shri Ramdeobaba college of Engineering & Management, Nagpur, India.

**Email: sharfuddinsr@rknec.edu*

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract

Social network usage has been increasing among common people day by day, which has lead to an outburst of social data. This social data is a reflection of the user's behaviour and can be used to predict the patterns with which the user will interact with the social community at large. In this work we propose a novel technique for prediction of user behaviour on the twitter platform which combines natural languages processing (NLP) ,session based analysis, sentiment analysis and behaviour pattern analysis. In this work first we perform NLP based pre-processing which clean all the tweets, for this we used POS tagging which remove all stop words, whitespaces etc. and extract variations of nouns , adjectives and adverbs, these module compares the tweets, and converts them into positive and negative polarity via fuzzy comparison ,and finally a session based engine is used to divide the tweet polarities into sessions and analyse the session-by-session polarity pattern to predict the next tweet behaviour of the user .Our work shows more than 80% accuracy when tested on real time data sets, and is significantly lightweight in terms of processing speed than standard systems.

Keywords: Polarity, Sentiment Analysis, Social Networks, Sessions Based Analysis, User Behaviour Prediction.

Introduction

Social pattern analysis has come up recently as an excellent field of study for observing user behaviour both online and in real life. Experts suggest that social network websites like Facebook and Twitter are one of the best tools to observe and in some cases predict the user behaviour by using pattern analysis and data mining. Now-a-days twitter has become most popular social networking site where users send and read posts of up to 140 characters and the posts are either image based or text based. On social networking sites, users share and post lots of knowledge and useful information which is most beneficial for another user in many ways.

User's action on social media generates behavioural data which show user tastes, involvement, opinions and relationships. Social behaviour of user means the user's opinion, tweeting or commenting positively or negatively and what kind of language (or word) they used on social media.

On social networking site there are different types of people where they exchange their opinions, some user has positive opinion and some have negative opinion . Muhammad Al-Quraishi et al. proposed a work to analyse and anomalous behaviours and proposed an integrated social media content analysis that leverages three levels of features, i.e. user generated content, social graph connections and user profile activities. For this also collected a large number of user profile files from twitter and YouTube and this system is based on multiple layers(four layers) which are mutually related and every module having direct communication with every other module. For this experiment, trained five supervised algorithms and used one computer for the experiment and assessed several classifiers i.e. decision tree J48,classification by regression (ClsReg),a support vector machine(SVM) ,random forest (RF)and OIR and trained the classifier model using both normal user and malicious user and the classifier perform well for the tested dataset, the F-measure for the RF and OIR algorithm were the same i.e. 95%,94% for J48,95% for ClsReg and the SVM attained a rate of only 84%[1].

In our work, we consider only twitter data or twitter user profile and we will be analysing the social behaviour of the users from their tweets and evaluate either the user is malicious or not and predict whether the user will be tweeting positively or negatively based on his/her history. For this work we used corpus based (or rule based model) and also implement by using support vector machine (SVM) algorithm.

The remaining paper is organized as follows: Section 2 describes the various techniques for analysis of social network patterns, Section 3 describe our technique used for analysis, section 4 results and comparison with other standard techniques and finally section 5 describe the conclusion by suggesting some future work which can be done by researchers reading this text.

Literature Review

The task of cutting or splitting a tweet into meaningful segment is called tweet segmentation**.**

For example: I will visit my parents in October.
This example split in 5 segments as shown below;

(I will) | (visit) | (my parents) | (in) | (October)

They proposed a novel frame work for tweet segmentation which splits a tweet into a sequence of consecutive n-grams and to recognize the named entity using Hybridseg framework and evaluate 5 variations of two NER(Named Entity Recognition) method i.e. random walk-based (RW-based) and POSbasedNER,namely$GlobalSeg_{Rw}$,$HybridSeg_{RW}$,$HybridSeg_{POS}$,$GlobalSeg_{POS}$ and $Unigram_{POS}$ where GlobalSeg denotes $HybridSeg_{Web}$ and HybridSeg denotes $HybridSeg_{Iter}$.The results show that segmentation is better for NER and $Unigram_{POS}$ is a worst performer and $HybridSeg_{POS}$ is much better and achieved the best NER result than the $HybridSeg_{RW}$ and demonstrate that segment-based NER methods achieve much better accuracy than the word-based [2].In our proposed work we used POS tagging and chunking for segmentation and NER and these named entities are further used for sentiment analysis.
The named entity may be the name of person ,place and so on ,[3],[4],[5],and [6] also proposed a Hybridseg framework for tweet segmentation and named entity recognition(NER).In [5] also clustering algorithm is used for dividing tweet stream into clusters of crime ,politics, religious, sports etc. Tweet classification is the task of dividing the tweet in a particular region or category ,tweet classification improves the accuracy and efficiency of tweets and the tweets are divided in specific region by applying data mining algorithm[6]'[7].The main purpose of [7] are classification of tweets by using data mining and clustering algorithm and also SVM is used for classification to removing the noisy tweet and identify the spam words, it also provide the current event detection .In our proposed work we used SVM classifier for sentiment analysis and detection of malicious user on twitter. In [8] ,presents a method for constructing effective classifiers for malicious application.
Nowadays, the popularity of social media is increased, most of the people consume news from social media instead of traditional media and on social media most of the time people spread rumors or fake news. In [9] explore the fake news problem and introduced a basic concept of fake news in traditional media and social media.
The excessive use of OSNs causes a great increase in anomalies .In OSNs anomalies can signify irregular and illegal behavior, the anomalies can be identified as malicious user, sexual predators and online fraudsters, for detecting these types of anomalies twitter dataset is used for analyzing the user behavior and analyzed the tweets whether it is an anomalous or not [10].As compared to [9] and [10] we are focused on malicious user detection which used bad language (abusing word) and negative tweeting.
Proposed a method that identifies the spam tweets without knowing the previous background of user and detect the spam on twitter by analysis of language based on language tool, they mainly focus on analysis of tweet not on the user account [11] , and we are focused on the social behavior of users from their tweets and evaluate the user is malicious or not and predict whether the user tweeting positively or negatively by knowing the previous background of user.
Identified the malicious content and the short URLs and protect the user from unauthorized activities and provide a security to twitter user, besides the user gets some alert mails [12]. Presents a formal formulation of malicious URL detection using machine learning, also discussed practical issues in system design and open research challenges[13].A generic tweet segmentation framework named Facebook Rigorous Application Evaluator(FRAppE),which detect the misuses in cyberbulling ,this app automatic forward warning before uploading of critical and harmful messages[14].FRAppE can only detect malicious applications on social media, and VIPS algorithm and Knuth-Morris-Pratt algorithm detects malicious posts and shared malicious URL's and also blocked the malicious posts and URL's. By using a similarity approach in text analysis detects suspicious posts in social network [15].
A supervised learning algorithm is used to detect the malicious URL or the attackers and consume very less time as compare to other existing system [16].A semi supervised spam detection (S^3D) framework detect spam tweets on real time basis and updates the models periodically in batch mode [17].To discover tweet spam and non-tweet spam used a new Fuzzy k means clustering based method which is dissimilar from the normal spam detection method and it combine similar user tweet trending topic [18].

Proposed Work

Our proposed work consists of the following blocks;

- Fetching of real time tweets for users
- NLP based pre-processing
- Corpus based sentiment analysis
- Session based behaviour pattern analysis
- Prediction of user behaviour

In the first module, we use a Tweet fetching API provided by twitter themselves in order to fetch the latest tweets for the users. These tweets are stored in

text files which can be further used for processing and pattern analysis. In our proposed work, a service is developed which fetches the tweets on a twice-a-day basis and keeps them ready for further processing. This service is kept running in the background, and is only stopped when the analysis for a particular user is completed.

The tweets fetched from this module are then given for language processing to the NLP module. The NLP module first tags the sentence into parts of speech, such as nouns, pronouns, verbs, adverbs, adjectives, conjunction, interjection and others. Out of these parts of speech tags, we only extract all variations of nouns, verbs and adjectives and discard the words which belong to any other tag. This helps to reduce all the stop words from the tweets, and thus helps in speeding up the process of comparison for polarity evaluation; we process only those words which are useful, thereby improving the overall accuracy of the system.

For example, if the input tweet is, "(*Going to the shore for some fishing, hoping to catch a big trout)*", then the total words in the tweet are 13, after NLP, the tweet text is converted to, *"going shore fishing hoping catch big trout"*. The processed tweet has 7 words, which are almost half as that of the original tweet, thus it takes less delay to process the text obtained post NLP, and as the text only contains action words, thus the comparison with ontology's is more accurate.

Using rule based model

These NLP processed tweets are given to a corpus based sentiment analysis module. This module compares the tweets, and converts them into positive and negative polarity via fuzzy comparison. We prepared corpuses of positive words, negative words, inversion words and extra sentiment words. For each of the NLP processed tweets we perform the following;

Algorithm 1: Corpus based sentiment analysis

1. Check If the input text is present in the positive ontology,
2. If yes,
3. Increase the positive score
4. Check if the input text is present in the extra sentiment words along with the positive word,
5. If yes,
6. Increment the positive score
7. Check if the word has an inversion word,
8. If yes,
9. Make positive score = 0, and increment the negative score by 2
10. Repeat the same steps for negative ontology

After the checking is completed, we compare the positive word score with the negative word score. If the positive word score is more than the negative word score, then the polarity of the tweet is positive, otherwise the polarity of the tweet is negative.

The polarities of each of the tweets are stored in an array, and this array is further processed using session analysis. The session analysis algorithm works according to the following rules;

Algorithm 2: Session based behaviour pattern analysis

1. Divide the input polarities into sessions
2. Each session contains more than 1 tweet (For example 10 tweets per session), and all sessions are evenly distributed except the last one
3. For each session, evaluate the mean polarity (M_P)
4. $M_P = sum\ of\ polarity\ of\ each\ tweet / No.\ of\ tweets$
5. If MP > = 3
6. Session marked as positive session
7. Else
8. Session marked as negative session
9. Count no. Of positive session and no. of negative session
10. Calculate the positive probability (Pos_P) and negative probability (Neg_P)

$Pos_P = no.\ of\ positive\ session / no.\ of\ positive\ session + no.\ of\ negative\ session$

$Neg_P = no.\ of\ negative\ session / no.\ of\ positive\ session + no.\ of\ negative\ session$

11. If $Pos_P > Neg_P$
12. User should tweet positively
13. Else
14. User should tweet negatively

The sessions scores are then stacked into an array, and that array is given to the pattern analysis block. The pattern analysis block checks the continuity between the sessions of positive and negative polarities. If for a person, the session scores are continuously improving, then the probability of the person sending a positive tweet increases, and vice versa. But, if the scores are positive initially and fall down for some sessions, and are again positive for the next sessions, then there is a low probability that the person will tweet positively, and thus is marked as a mildly malicious user. But, if a person's tweets are continuously negative, and are positive only for some abrupt sessions, then the person is marked as a malicious user, and can be blocked from social media websites. This method is tested on real time tweet records, and the results are evaluated in the next section.

Using support vector machine (SVM) algorithm

The same work is done by using support vector machine (SVM) algorithm ,which is supervised machine learning algorithm .First we trained the 1000 of tweets as positive and negative .Positive tweet marked by value 1 and the negative tweet marked by value 2.For testing, fetch real time tweet of any user

and divide the tweets into sessions and calculate the mean polarity of each session .The mean polarity of each session is evaluated, where positive tweets are given a polarity value of 5, and negative tweets are given a polarity of 1. If the mean polarity for a particular session is more than or equal to 3, then that session is marked as a positive session, else it is marked as a negative session. The SVM algorithm gives best result for a small amount of data.

Results and Discussion

In this work our main task is to analyse the social behaviour of the user and detect the malicious user and predict whether the user will be tweeting positively or negatively based on his/her previous tweets. We evaluate this by using two methods, namely i) Rule based model and ii) Support vector machine algorithm and try to achieve high accuracy.

Using rule based model

We tested the tweet patterns of our Honorable Prime Minister, Shri Narendra Modiji from his twitter handles; narendramodi, narendramodi177 and PMOIndiaModi.

Each of the tweets is assigned a polarity score by using algorithm 1 demonstrated in section 3. The tweet analysis is divided into sessions, and the session analysis can be observed from the following pie chart (Figure 1).

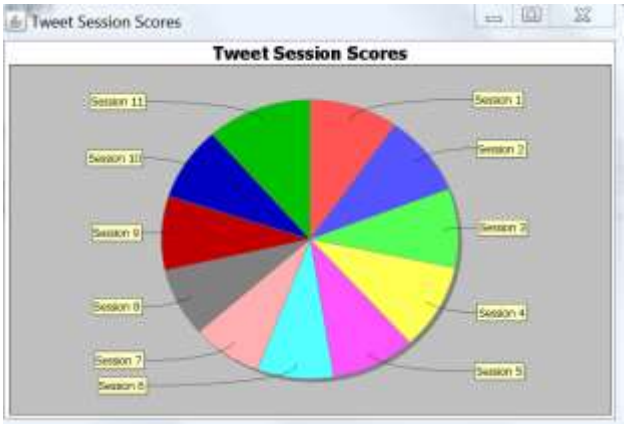

Fig. 1: Session Based Analysis

As we can see, the session and their scores are an indication of the positivity of the user's tweet, and is further pattern analyzed in order to obtain the behavior pattern(by using algorithm 2 demonstrated in section 3) of the user under study. In this case, there are 11 sessions formed, and each session has nearly 10 tweets. The session wise analysis for this case indicates that the person/user under test will be tweeting something positive, and is a valid and genuine user. The following session analysis Figure 2 demonstrates this particular fact.

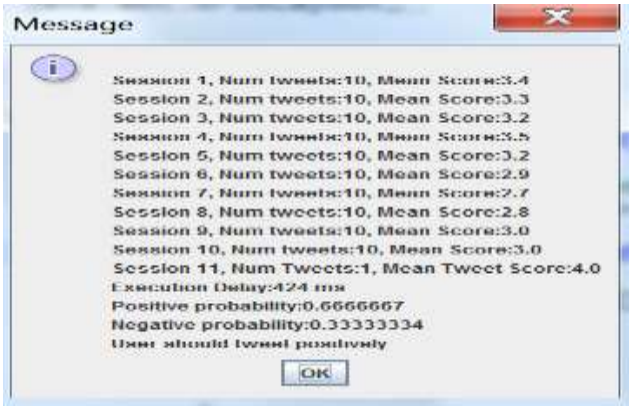

Fig. 2: Prediction Analysis

By using the rule based model we achieve 85% accuracy. Thus, our approach performs much better when the mean accuracy is compared across all the classification based techniques, and can be used for large real time data sets.

Using support vector machine (SVM) algorithm

We trained thousands of tweets as positive tweet and negative tweet first and test tweet for different user one by one and assign a polarity score to each tweet as shown in Figure 4.

Fig. 3: Tweet Analysis (Using SVM)

Session based analysis can be observed from the following pie chart (Figure 5).

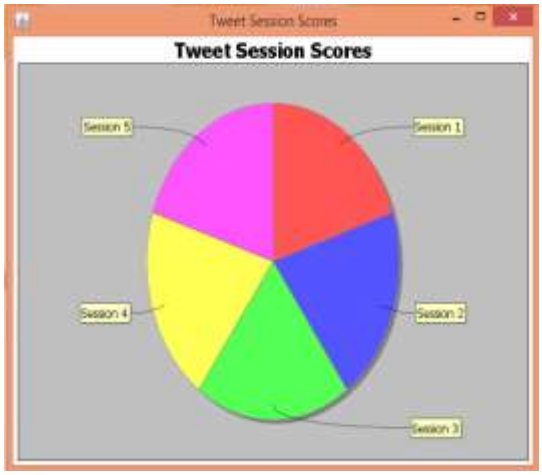

Fig.4: Session Based Analysis (Using SVM)

As we observed Figure 5, there are 5 sessions formed and each session has nearly 5 tweets. The session wise analysis for this case indicates that the person/user under test will be tweeting something positive, and is a valid and genuine user. The following session analysis Figure 6 demonstrates this particular fact.

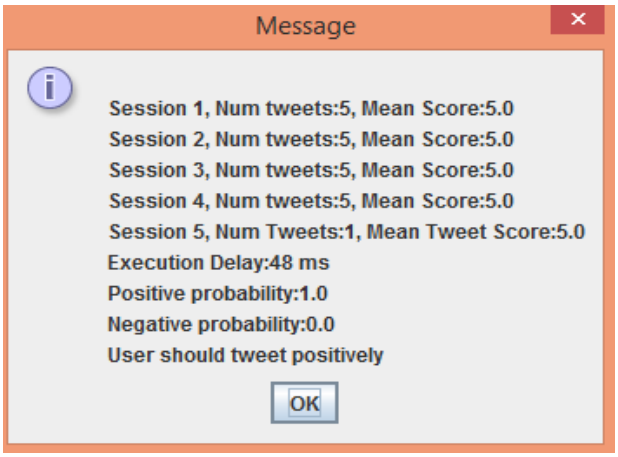

Fig. 5: Prediction Analysis (Using SVM)

Accuracy	True Positive	False Positive
94%	0.995	0.06

Table 1: Accuracy of SVM (Proposed Work)

In [11], they proposed a method to identify the spam tweets without knowing the previous history of user by using SVM classifier and achieve 93% accuracy, as compared to this our work detect malicious user by studying the previous history of user and record 94% accuracy .The [18] demonstrate that extreme learning machine (EML) classification methods perform better than the SVM with the accuracy of 94% .In [1] also SVM classifier attained a rate of only 84% as discussed in section 1.

For accuracy, we have considered the success rate with which our system is able to detect whether the person will be tweeting positively or negatively based on his history. Our propose work is much better than the all other previous work.

Conclusion

Our proposed session based approach outperforms all other classification based approaches, and thus can be used in real time with large datasets. Our session based approach demonstrates a mean performance of more than 80% accuracy, and by using SVM algorithm we achieve accuracy more than 90%, which is a big jump from the conventional classification approaches, and thus is suitable to be integrated into real life social networks like twitter, Facebook or LinkedIn.

Future Work

Researchers can further evaluate the performance of the session based system for other types of applications like ecommerce, and block chain systems. As an example, the session based approach can be applied to ecommerce systems in order to evaluate if a particular product is performing well, or if the demand of a particular product is following what kind of trend. The applications of the session based approach can be limitless.

References

[1] Muhammad Al-Quraishi, Shamim Hossain ,Majed Alrubaian,Sk Md Mizanur Rahman and Atif Alamri,"Leveraging analysis of use behavior to identify malicious activities in large-scale social networks", IEEE transactions on industrial informatics,Vol.14,No.2(2018),pp-1551-3203.

[2] Cheliang Li,Aixin sun,Jianshu Weng and Qi He," Tweet segmentation and its application to named entity recognition", IEEE transactions on knowledge and data engineering,Vol.27,No.2(2015),pp-1041-4347.

[3]Vikas Balasaheb Burgute and A.K.Gupta,"Named entity recognition using tweet segmentation", International research journal of engineering and technology,Vol.4,No.7(2017),pp-2395-0056.

[4]Chetan Chavhan and Ranjeetsingh Suryawanshi,"Summarization of tweets and named entity recognition from tweet segmentation", International conference on automatic control and dynamic optimization techniques,(2016),pp-66-71.

[5] Anjum I.Inamdar,Vishaka V.Shinde,Harshata P.Tohake and Kartiki S.Wahatole ,"A tweet segmentation of HAVK2",International journal of research in advent technology,(2017),pp-2321-9637.
[6] Cheliang Li,Aixin sun,Jianshu Weng and Qi He," Exploiting hybrid contextsforweet",Sigir(Researchgate),(2013),(DOI:10 .1145/2484028.2484044).
[7] Sonam Meshram and Hirendra Hajare,"Tweet segmentation and enhancement of tweets", International journal of science and research,Vol.5,(2016),pp-2319-7064.
[8] Sonam U.Meshram ,Manali R.Raut and Madhavi R.Bichwe,"Tweet segmentation and spam prevention", International journal of engineering science and computing,Vol.7,No.3(2017).
[9] Kai Shu,Amy Sliva and Suhang Wang,"Fake news detection on social media :Adata mining perspective",Cssi,(2017).
[10] Vishal Chauhan, Ajay Pilaniya and Vishesh Middha," Anamalous behaviour detection in social networking", IEEE,(2017).
[11] Sagar Gharge and Manik Chavan,"An integrated approach for malicious tweets detection using NLP", International conference on inventive communication and computational technologies,(2017).
[12] Nupur S. Gawale and Nitin N.Patil,"Implementation of a system to detect malicious URLs for twitter users", International conference on pervasive computing,(2015).
[13] Doyen Sahoo,Chenghao Liu and Steven C.H.Hoi," Malicious URL detection using machine learning: A survey", IEEE,(2017).
[14] M.Ganga and S.Aanjan Kumar," Segmenting and detecting malicious tweets and harmful entity recognition", International journal of innovative research in computer and communication research,Vol.4,No.4(2016)-pp-2320-9801.
[15] Sayali S.Karmode and Vaishali B.Bhagat,"A review: detecting and blocking social media malicious posts", International journal of modern trends in engineering and research,Vol.3,No.11(2016),pp-2349-9754.
[16] K.Divya and R.Kiruba Kumari,"Sentiment analysis based named entity recognition with tweet segmentation", SSRG International journal of communication and media science,Vol.3,No.5(2016),pp-2349-641X.
[17] Surendra Sedhai and Aixin Sun,"Semi-supervised spam detection in twitter stream", CSIR(research gate),(2017).
[18] Saini J.Soman and S.Murugappan,"Detecting malicious tweets in trending topics using clustering and classification", International conference on recent trend in information technology,(2014).

