



Twitter Analysis Based on Social Bots using Learning Automata with URL Features in Twitter Network

A. Prabavathy ^{#1}, G. Koteeswari ^{#2}

#1 M.Sc, Department of Computer Science, Kamban College of Arts and Science for Women, Tiruvannamalai-606603.

#2 Assistant professor, Department of Computer Science, Kamban College of Arts and Science for Women, Tiruvannamalai-606603.

ABSTRACT

Twitter is one type of social media that is often used. Users use Twitter to convey their tweet to the general public. The number of Twitter users has reached 330 million people worldwide. Besides that, in Twitter there are tweets that can be sentiments. Twitter is one of the most popular online social networking sites where users communicate and interact on various topics. In this paper, Deep CNN Algorithm is used to analysis the sentiments in twitter data. The sentiments are analyzed and labelled as positive and negative labels. These sentiments are visualized as bar graph. This improves the accuracy of the classification.

1.INTRODUCTION:

Malicious social bot is a software program that pretends to be a real user in online social network. Moreover, malicious social bots perform several malicious attacks, such as spread social spam content, generate fake identities, manipulate online ratings, and perform phishing attacks. In Twitter, when a participant (user) wants to share a tweet containing URL(s) with the neighboring participants (i.e., followers or followees), the participant adapts URL shortened service in order to reduce the length of URL (because a tweet is restricted up to 140 characters). Moreover, a malicious social bot may post shortened phishing URLs in the tweet. They generate fake tweets and automate their social relationships either by pretending like a follower or by creating multiple fake accounts with malicious activities. Moreover, malicious social bots post shortened malicious URLs in the tweet in order to redirect the requests of online social networking participants to some malicious servers. Hence, distinguishing malicious social bots from legitimate users is one of the most important tasks in the Twitter network. Several approaches have been proposed to detect spam in the Twitter network. These approaches are based on tweet-content features, social relationship features, and user profile features. However, the malicious social bots can manipulate profile features, such as hashtag ratio, follower ratio, URL ratio, and the number of retweets. The malicious social bots can also manipulate tweet-content features, such as sentimental words, emoticons, and most frequent words used in the tweets, by manipulating the content of each tweet. The social relationship-based features are highly robust because the malicious social bots cannot easily manipulate the social interactions of users in the Twitter network. However, extracting social relationshipbased features consumes a huge amount of time due to the massive volume of social network graph. Therefore, identifying the malicious social bots from the legitimate participants is a challenging task in the Twitter network. The existing malicious URL detection approaches are based on DNS information and lexical properties of URLs. The malicious social bots use URL redirections in order to avoid detection.

2.LITERATURE SURVEY

[1]. **Matthew Toews**This paper proposes an inference method well-suited to large sets of medical images. The method is based upon a framework where distinctive 3D scale-invariant features are indexed efficiently to identify approximate nearest-neighbor (NN) feature matches in $O(\log N)$ computational complexity in the number of images N . It thus scales well to large data sets, in contrast to methods based on pair-wise image registration or feature matching requiring $O(N)$ complexity. Our theoretical contribution is a density estimator based on a generative model that generalizes kernel density estimation and K-nearest neighbor (KNN) methods.

[2]. **Ping Zhong**In wireless rechargeable sensor networks (WRSNs), there is a way to use mobile vehicles to charge node and collect data. It is a rational pattern to use two types of vehicles, one is for energy charging, and the other is for data collecting. These two types of vehicles, data collection vehicles (DCVs) and wireless charging vehicles (WCVs), are employed to achieve high efficiency in both data gathering and energy consumption. To handle the complex scheduling problem of multiple vehicles in large-scale networks, a twice-partition algorithm based on center points is proposed to divide the network into several parts. In addition, an anchor selection algorithm based on the tradeoff between neighbor amount and residual energy, named AS-NAE, is proposed to collect the zonal data. It can reduce the data transmission delay and the energy consumption for DCVs' movement in the zonal. Besides, we design an optimization function to achieve maximum data throughput by adjusting data rate and link rate of each node.

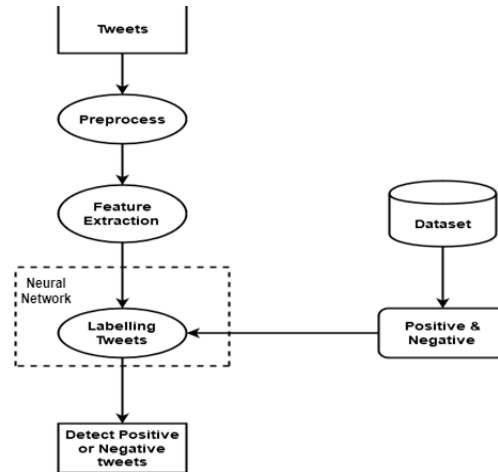
[3]. **Yuting Xu**With the continuous development and progress of healthcare monitoring system, medical diagnosis for human health plays a particularly critical role, which can help doctors make correct choices and effective treatment plans. However, effective feature extraction is very important for the analysis of functional Magnetic Resonance Imaging (fMRI) data, the traditional feature-based dictionary learning algorithm ignores the relationship between atoms and the input samples, and the small sample data is prone to over-fitting. In this paper, we propose a new weighting mechanism, which effectively considers the relationship between the atom and the input sample; Meanwhile, the cross-validation method performed well on obtaining additional validation sets but proved to be over-fitting on small datasets in the traditional dictionary learning algorithm. Therefore, ℓ_2 -norm and norm regularization constraint is adopted to avoid over-fitting, achieve the limitations of the model space, and improve the generalization ability of the model. In order to extract features better, this paper uses the cosine similarity method to select good feature subsets, which effectively improves further the generalization ability and enhances the feature extraction accuracy. The results show that the improved dictionary classification algorithm has better performance in terms of accuracy, sensitivity, and specificity, it also demonstrates that the proposed algorithm has an effective classification about mobile multimedia medical diseases, which can provide better guidance for the diagnosis of later diseases, so as to promote the rapid development of medical feature extraction.

[4]. **Vince Mariano** Sentiment analysis of in the domain of micro-blogging is a relatively new research topic so there is still a lot of room for further research in this area. Decent amount of related prior work has been done on sentiment analysis of user reviews [x], documents, web blogs/articles and general phrase level sentiment analysis [16]. These differ from twitter mainly because of the limit of 140 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is very expensive. Some work has been done on unsupervised (e.g., [11] and [13]) and semi-supervised (e.g., [3] and [10]) approaches, and there is a lot of room of improvement. Various researchers testing new features and classification techniques often just compare their results to base-line performance. There is a need of proper and formal comparisons between these results arrived through different features and classification techniques in order to select the best features and most efficient classification techniques for particular applications.

[5]. **Pranali Borele**With the rise of social networking epoch and its growth, Internet has become a promising platform for online learning, exchanging ideas and sharing opinions. Social media contain huge amount of the sentiment data in the form of tweets, blogs, and updates on the status, posts, etc. In this paper, the most popular micro blogging platform twitter is used. Twitter sentiment analysis is an application of sentiment analysis on data from Twitter (tweets), to extract user's opinions and sentiments. The main goal is to explore how text analysis techniques can be used to dig into some of the data in a series of posts focusing on different trends of tweets languages, tweets volumes on twitter. Experimental evaluations show that the proposed machine learning classifiers are efficient and performs better in terms of accuracy. The proposed algorithm is implemented in python.

3. PROPOSED SYSTEM:

Twitter sentiment analysis provides the organizations with the ability to surveying public emotion. In this paper, we introduce the Deep CNN algorithm to analysis the twitter sentiments. The twitter data are fetched using twitter api file. The twitter data are analysed and labelled as positive and negative labelling. Deep CNN algorithm performs both training and testing process to label the sentiments.

SYSTEM ARCHITECTURE:**4.METHODOLOGY:****Data Acquisition:**

Data in the form of raw tweets is acquired by using the python library “tweestream” which provides a package for simple twitter streaming API [26]. This API allows two modes of accessing tweets: SampleStream and FilterStream. SampleStream simply delivers a small, random sample of all the tweets streaming at a real time. FilterStream delivers tweet which match a certain criteria. It can filter the delivered tweets according to three criteria:

Specific keyword(s) to track/search for in the tweets

Specific Twitter user(s) according to their user-id’s • Tweets originating from specific location(s) (only for geo-tagged tweets). A programmer can specify any single one of these filtering criteria or a multiple combination of these. But for our purpose we have no such restriction and will thus stick to the SampleStream mode.

Retrieve words:

Twitter is not just an extended source of news. The twitter data can be using the consumer key, consumer secret, access token, access token secret. Twitter allows the usage of their API via an oauth2 authorization framework.

Learning phase:

The learning data is used to make sure the machine recognizes patterns in the data, the cross-validation data is used to ensure better accuracy and efficiency of the algorithm used to learn from data, and the test data is used to see how well the machine can predict new answers based on its learning.

Clustering phase:

Within Deep Learning, a Convolutional Neural Network or CNN is a type of artificial neuralnetwork, which is widely used for image/object recognition and classification.

Prediction module:

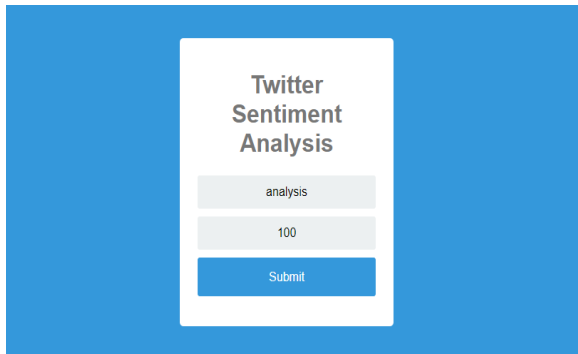
When the reasons behind a model’s outcomes are as important as the outcomes themselves, Prediction Explanations can uncover the factors that most contribute to those outcomes. This shows the output of the algorithm from the testing case.

5CONCLUSION

In this work, it is found that the classifying the twitter data using the sentiment analysis techniques. This labels the tweets as positive and negative labelling. The classification model is designed in this research work for the hot news detection. The clustering is the labelling of positive and negative tweets using deep CNN algorithm. The performance of the proposed model is analyzed in terms of accuracy, precision and recall.

6.RESULTS

Fig:1 Twitter Analysis Tweets



The image shows a web interface for Twitter Sentiment Analysis. It features a blue background with a white central box containing the text "Twitter Sentiment Analysis". Below the text are three input fields: the first contains "analysis", the second contains "100", and the third is a blue button labeled "Submit".

Fig:2 Possibility of Positive Tweet

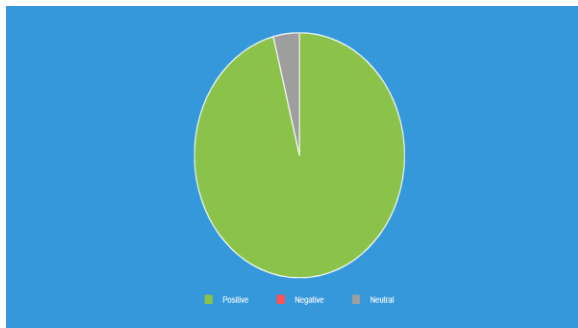


Fig:3 Possibility of Negative Tweet

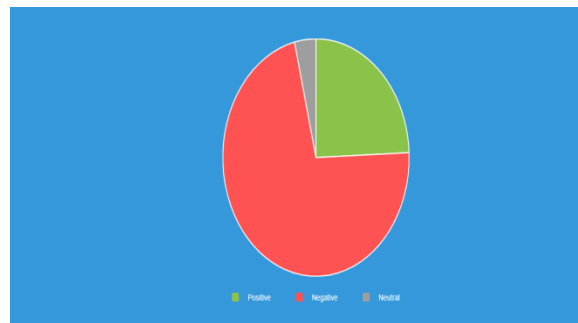
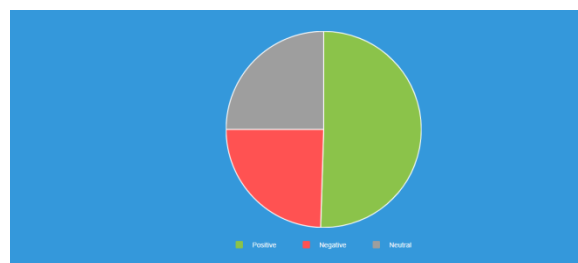


Fig:4 Possibility of Neutral Tweet



REFERENCES:

-
- [1] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," *Icwsn*, vol. 11, no. 122–129, pp. 1–2, 2010.
- [2] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, 2014, pp. 94–97.
- [3] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the 20th international conference on Computational Linguistics*, 2004, p. 1367.
- [4] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 625–631.
- [5] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis," *Emot. Sentiment. Soc. Expressive Media*, p. 9, 2013.
- [6] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 628–632.
- [7] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol. 150, no. 12, pp. 1–6, 2009.
- [8] M. Bouazizi and T. Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter," *IEEE Access*, vol. 3536, no. c, pp. 1–21, 2017.
- [9] R. Sara, R. Alan, N. Preslav, and S. Veselin, "SemEval-2014 Task 9: Sentiment Analysis in Twitter," in *Proc. of the 8th International Workshop on Semantic Evaluation*, 2014, pp. 73–80.
- [10] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in twitter," in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 451–463.
- [11] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter," *Proc. 10th Int. Work. Semant. Eval.*, pp. 1–18, 2016.
- [12] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, 2000.