MACHINE LEARNING BASED SENTIMENT ANALYSIS ON TWITTER DATA

***Ms. Khushboo Saglani*[1]*,Prof.Dr. Nitin Janwe*[2]**

[1] Department of Computer Science and Engineering,
[1]Rajiv Gandhi College Of Engineering Research and Technology, Chandrapur, India.

Abstract: With the increase of social networking epoch and its growth, the net has become one in all the powerful platforms for online learning, exchanging ideas and sharing opinions. Social media contain an enormous amount of the info for sentiment analysis within the type of tweets, blogs, and updates on the status, posts, etc.This paper addresses the matter of sentiment analysis in twitter; that's classifying tweets in keeping with the sentiment into positive, negative or neutral. Analysing the emotions of the general public became important to seek out the reviews of the shoppers for any product within the market, predicting political elections and predicting socio economic phenomena like stock market. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream by using convolutional neural networks.

***IndexTerms* – Twitter, Sentiment, Natural Language Processing, CNN.**

I. INTRODUCTION

Twitter has emerged as a significant micro-blogging website, having over 100 million users generating over 500 million tweets each day. On Twitter, users are allowed to share their opinions within the type of tweets, using only 140 characters. This ends up in people compacting their statements by using slang, abbreviations, emoticons, short forms etc. together with this, people convey their opinions by using sarcasm and polysemy. so as to extract sentiment from tweets, sentiment analysis is employed. The results from this could be utilized in many areas like analyzing and monitoring changes of sentiment with an occasion, sentiments regarding a specific brand or release of a specific product, analyzing public view of presidency policies etc. lots of research has been done on Twitter data so as to classify the tweets and analyze the results. During this paper we aim to review some research during this domain and study the way to perform sentiment analysis on Twitter data using Python.

There are various approaches for sentiment analysis on Twitter, one of them is machine learning. Deep learning models have achieved great results in computer vision6 and speech recognition in recent years. To solve NLP (Natural Language Processing) problems, machine learning is also useful by using a general learning algorithm combined with a large sample of data to learn the classification rules. Several methods do it with traditional algorithms such as SVM or Naïve Bayes, most of such methods consider text word by word, classify a sentence to positive or negative by analyzing the word in the text, sometimes information lost by extracting key word without other word.

CNN (Convolutional Neural Networks) is one in every of the machine learning models which has achieved impressive ends up in image recognition several years ago and has achieved remarkable ends up in tongue processing recently. there's a convolutional layer to form a bit of words that will be considered together. During this paper, we propose an approach to parsing Twitter data to grasp situations within the universe that supports a CNN model to try and do sentiment analysis. We adopt convolutional neural networks as our sentiment analysis model because within the image analysis and classification field, CNN can extract a district of features from global information, with the convolution operation, a bit of information information will be extracted together because the features, and it's ready to consider the link among these features. For computer vision, like image analysis, it's ready to extract a component of pixel data information, not only extract the pixels one by one, the features information will be extracted piece by piece, the piece contains multi pixels data information; after we transfer the text into matrix, it also can be considered as same as a picture pixels' matrix, so we will do the identical operation to the text data to form the input features to the model will be trained in another effective way.

The procedure for sentiment analysis involves the following major steps:

A. *Data Collection and Preprocessing*
To obtain the twitter data, we first need to access the Twitter API and keys such as secret key and control key.. After the necessary data has been obtained, we further proceed to the next step that is data cleaning in order to obtain the data in the useful format. This is efficient because the tweets contain a lot of noise factors. Since the tweets have a limitation of maximum 140 characters, people tend to use slang words , emoticons etc which lead to inappropriate dataset. The preprocessing techniques include: Tokenization and Removal of non-English Tweets, URL, stop words and hashtags . Once the data has been cleaned up, we perform its conversion into a data frame.

B. *Common learning algorithms*
After we have the appropriate dataset, we need to follow a sentiment analysis technique, stating that there are two important sentiment analysis methods, one is machine learning and the other is Lexicon-based Approach. The first approach deals with a supervised classification algorithm. The second approach depends on the understanding that the polarity of a text sample can be acquired on the grounds of the polarity of the words which comprise it. There are many different ways to obtain the polarity

of sentence: Natural Language Processing (NLP), Support Vector Machine (SVM), Case-Based Reasoning (CBR), Artificial Neural Network (ANN).

This paper covers the comparison and analysis of all the research and methodologies that have been used to implement sentiment analysis on Twitter data in the past decade.

REVIEW OF SENTIMENT ANALYSIS ON TWITTER DATA

Sentiment analysis could be a language Processing task which is being handled at many levels of granularity. It contains three levels of granularity as follows-a document level classification , a sentence level classification and a phrase level classification. There are various studies on classifying sentiments using machine learning models, like Support Vector Machine (SVM), Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and ensemble. The foremost recently-used features for such machine learning models are n-grams.

Read .J[1], used unigram features for sentiment binary classification and obtained 88.94% accuracy using the SVM. Within the study of Gautam and Yadav[3] , they utilized the SVM together with the semantic analysis model for the sentiment binary classification of Twitter texts and achieved 89.9% accuracy using unigram features.

Barnaghi et al[4] , use unigrams and bigrams and apply Term Frequency Inverse Document Frequency (TF-IDF) to search out the load of a selected feature in a very text and hence filter the features having the most weight. The TF-IDF could be a very efficient approach and is widely utilized in text classification and data processing.

Anton and Andrey[6] developed a model to extract sentiment polarity from Twitter data. The features taken out were words containing emoticons and n-grams . The experiment demonstrated that the SVM performed better than the Naïve Bayes. The most effective overall performing method was the SVM together with unigram feature extraction, gaining a precision accuracy of about 81% and a recall accuracy of 74%.

Malhar and Ram[5] proposed the supervised method to categorize Twitter data. The results of this experiment demonstrated that the SVM performed better than other classifiers and, employing a hybrid feature selection, achieved an accuracy of 88%. The experiment attempted to mix principal component analysis (PCA) alongside the SVM classifier to cut back feature dimensionality. Further, unigram, bigram, hybrid (unigram and bigram) feature extraction methods were used. Malhar and Ram showed that integrating PCA with the SVM with a hybrid feature selection could help in reducing feature dimensions and therefore the results obtained a classification accuracy of 92%.

Pablo et. al. [15] had described the variations of Naive Bayes classifiers for detecting polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline and Binary. The features undertaken by classifiers were Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters.

Go et al. (2009)[12] use distant learning to get the sentiment data. They make use of tweets containing positive emoticons and negative emoticons . They build models using Naive Bayes, MaxEnt and Support Vector Machines (SVM), and reports that SVM performs better than other classifiers. For feature extraction, they try a Unigram, Bigram model in combination with parts-of-speech (POS) features. They note that the unigram model performs better than all other models.

Agarwal et al[13]. approached the task of classifying the sentiment from twitter, as a 3-way task of classifying sentiment into positive, negative and neutral classes. They worked with three styles of models: unigram model, a feature based model and a tree kernel based model. For the tree kernel based model they constructed a brand new tree representation for tweets. The feature based model utilizes 100 features and also the unigram model utilizes over 10,000 features. They wind up with the conclusion that features which contain prior polarity of words in combination with their parts-of-speech tags are most important for the classification task. The tree kernel based model performed better than the opposite two.

Kouloumpis et al[7]. explored the usefulness of various linguistic features for classifying the sentiments of tweets. The emoticon (EMOT) and hash-tagged (HASH) datasets were used to train the model and the iSieve dataset was used for the testing. In this study, various feature sets were produced with the help of unigrams, bigrams, lexicons, part-of-speech and micro-blogging elements. The AdaBoost classifier was trained using these selected features in various combinations. The results showed that microblogging features were the most useful. The best results were achieved when n-gram features were used along with lexicon and micro-blogging features. An F-score of 0.65 was obtained with HASH and EMOT datasets combined.

Masud et al[8]. employed a vocabulary-based system for sentiment classification, which classifies the tweets as positive, negative, or unbiased. This system differentiate the scored slang utilized in tweets. The experimental outcomes demonstrated that the scheduled framework performed better than existing frameworks, gaining 92% precision in double characterization and 87% in multi-class grouping.

Another important effort for sentiment classification on Twitter data is done by Barbosa and Feng (2010)[9]. They utilized the polarity predictions from three websites as noisy labels to train a model and use 1000 manually labeled tweets for training and another 1000 manually labeled tweets for evaluation. They recommended the use of syntax features of tweets like, hashtags, link, retweet, punctuation and exclamation marks in combination with features like prior polarity of words and POS of words. The results show that the features that improve the performance of classifiers are the combination of prior polarity of words with their parts of speech.

Po-Wei Liang et.al.(2014)[10] utilized the Twitter API to collect the tweets. Their training data falls into three varieties of categories ie, camera, movie and mobile. The data is classified as positive, negative and non-opinions. Tweets that contain opinions were filtered. Naive Bayes model was implemented with Unigram feature and the Naive Bayes simplifying independence assumption was employed. They also evicted the useless features by using the Mutual Information and Chi square feature extraction method. Finally , the sentiment of a tweet is predicted. i.e. positive or negative.

Bifet and Frank(2010) [11] used Twitter streaming data afforded by Firehouse API , which gave all messages from each user which are publicly available in real-time. They implemented multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. Their results show that the SGD-based model, when used with an appropriate learning rate, was better than the rest used.

Lin and Kolcz[14] implemented integrating multiple classifiers into large-scale twitter data. They experimented to train logistic regression (LR) classifiers from the hashed 4-grams as features. The training dataset contains about one to 100 million examples with ensembles of 3 to 41 classifiers. The experiment showed that the accuracy of sentiment analysis of Twitter data using multiple classifiers was better than with a single classifier. The best performance was obtained when the classifiers are total 21 and the instances was 100 million, achieving a classification accuracy of around 0.81.

Jiang et al[2].implemented binary classification using the SVM and achieved 67.8% accuracy for Twitter texts, where they used two forms of target-independent features ie, twitter content features and sentiment lexicon features.

CONCLUSION

Twitter sentiment analysis comes under the group of opinion and text mining. It focuses on analyzing the feelings of the tweets and feeding the info to a machine learning model so as to coach it then check its accuracy, in order that we are able to use this model for future use in line with the results. It comprises steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of virtually 85%-90%.In this paper, we offer a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, Research results show that machine learning methods, like SVM and naive Bayes have the best accuracy and might be considered the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few effort in human-labeled document .We also studied the consequences of varied features on classifier.Hence we are able to conclude that more the cleaner data, more accurate results are often obtained.

REFERENCES

[1] Read, J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, Ann Arbor, Michigan, 27–27 June 2005; pp. 43–48.

[2] Jiang, L.; Yu, M.; Zhou, M.; Liu, X.; Zhao, T. Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, 19–24 June 2011; Volume 1, pp. 151–160.

[3] Gautam, G.; Yadav, D. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Proceedings of the 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, India, 7–9 August 2014; pp. 437–442.

[4] Peiman Barnaghi, John G. Breslin and Parsa Ghaffari, “Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment”, 2016 IEEE Second International Conference on Big Data Computing Service and Applications.

[5] M. Anjaria and R. M. R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning," in 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), 2014, pp. 1-8.

[6] A. Barhan and A. Stikhomirov, "Methods for Sentiment Analysis of Twitter messages," in the 12th Conference of FRUCT Association, 2012.

[7] E. Kouloumpis, T. Wilson, and J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," Icwsm, vol. 11, no. 538-541, p. 164, 2011.

[8] F. M. Kundi, A. Khan, S. Ahmad, and M. Z. Asghar, "Lexicon-based sentiment analysis in the social web," Journal of Basic and Applied Scientific Research, vol. 4, no. 6, pp. 238-48, 2014.

[9] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.

[10] Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[11] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[12] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

[13] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, “Sentiment Analysis of Twitter Data”, In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 30–38

[14] J. Lin and A. Kolcz, "Large-scale machine learning at twitter," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 2012: ACM, pp. 793-804.

[15] Pablo Gamallo, Marcos Garcia, “Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.