



Sentiment Analysis in Data of Twitter using Machine Learning Algorithms

Asst Prof. Dr. Sefer Kurnaz¹; Mustafa Ahmed Mahmood²

¹Computer Engineering & Altınbaş University, Turkey

²Information Technology & Altınbaş University, Turkey

¹sefer.kurnaz@altinbas.edu.tr; ²mustafa.mahmood1093@gmail.com

Abstract— microblogging websites like twitter and fb throughout this new generation is loaded with reviews and information. one in each of the hugest used micro-running a blog computing device twitter is anyplace individual's percentage their principles within the form of tweets after which it becomes one amongst the simplest sources for sentimental evaluation. opinions are wide taken care of into three training smarts for positive unhealthy for negative and neutral and then the strategy of reading versions of opinions and grouping them altogether these classes is assumed as sentiment evaluation. Information mining is basically accustomed discover applicable information from websites drastically from the social networking web sites. merging method with numerous fields like textual content mining human language era and device intelligence a tendency to rectangular degree capable of classify tweets almost pretty much nearly as top bad or neutral. the foremost stress of this evaluation is on the class of emotions of tweets facts accrued from twitter. within the beyond researchers were exploitation existing device mastering strategies for sentiment evaluation however the effects confirmed that current machine learning techniques were not supplying better outcomes of sentiment category. consequently, on enhance type finally ends up in the area of sentiment evaluation we tend to rectangular degree exploitation ensemble machine getting to know strategies for growing the potency and trait of projected approach. for the equal a bent to rectangular measure merging aid vector system with name tree and experimental results show that our projected technique is offering higher type ends up in phrases of f-measure and accuracy in distinction to character classifiers.

Keywords— Sentiment Analysis, big data, analysis tweets

I. INTRODUCTION

one benefit for the social media over the traditional media resources, that its miles maintained and controlled through customers. conventional media totally authorized the customers to have the statistics which changed into offered to them. information drift was one sided that means it flows from media supply to customers. at the same time as in social community the consumer has the functionality of responding to the occasions and information surrounding them also offering their thoughts and sharing them. that led to growing multiway mode of spreading the facts wherein the user publishes records with further information consisting of videos, pictures and links. this ends in producing facts version this is generated with the aid of the user. the consumer's social graph and the consumer connections in terms of the social network have a big role within the system of reading the records version to get essential data from huge quantity of "user generated content material" that's generated

on day by day basis. as, microblogging sites which include flicker, twitter and Facebook permit the user to share shortened multimedia and messages, these web sites become an immediate statistics supply where the users all over the global are capable of stay linked and feature numerous resources to achieve information [1].

in addition, twitter is verbal exchange platform globally from all around the global. twitter has been developed to hold tempo with the fast verbal exchange among the arena. its flexibility and speedy tracking have made changing messages and twitter among humans one of the maximum critical contacts. data is accumulated from twitter thru an application programming interface (api) [2].

One of the most significant twitter features which distinguish it from other social networks like Facebook is that the relation between the one being followed and the one who is following does not need to be two-ways. Following a user in twitter is the same as subscribing to blog; The user who is following get all twitter status updates of the user he follows. Twitter can be considered as a pretty good reaction to all what happens in the worlds since it is widely used in all daily life aspects. Among all the things that happening, most recent trends are of a great interest to the companies. Most recent trends could be put to analysing and when reacted to, identified. From marketing perspective, these most recent trends could be utilized for the purpose of responding with suitable activities, such as advertising products. Tweets analysing could be of a significance for the companies for generating an advantage to rivals. [3]

II. PROPOSED SYSTEM

A. About of Twitter:

Twitter have several other features. One feature permits the user to reply to other users tweets by choosing to click the reply button on the tweet made by another user. It is a way were a user is able to respond to the tweet made by another user. Additionally, a '@' symbol is added before the username a twitter user when another user wants to mention the first one in his tweet. Mentioning is a way where a user wants to refer to another one. Retweeting is another twitter common concept. Retweet means to share other user tweet to our followers, it plays a significant role in spreading information within twitter. Also, a Hashtag '#' sign is added by the user before relevant keywords in their tweets. This is utilized for the purpose of categorizing the tweets to be shown easily in the twitter search. Popular twitter hashtags turn to be trending subjects in twitter.

B. Deep Sparse Autoencoder:

From the structural connect of recognize, the autoencoder is an axisymmetric base hit hidden-layer neural join [4]. The autoencoder encodes the input sensor broadcast by per the disoriented layer, approximates the minimum lapse, and obtains the best-feature hidden-layer conceit [5]. The work of genius of the autoencoder comes from the unsupervised computational pose of cro magnon man perceptual training [6], which itself has some down-to-earth flaws. For concrete illustration, the autoencoder does not revoke any practical achievement through copying and inputting hallucination into suggested layers, during it bounce reconstruct input message with fancy precision. The rare autoencoder inherits the sense of the autoencoder and introduces the sparse comeuppance term, adding constraints to feat learning for a concise conceit of the input statement [7, 8].

$$p_j = \frac{1}{n} \sum_{i=1}^n [a_j(x(i))] \quad (1)$$

The hidden layer is unbroken at a lower worth to confirm that the typical activation worth of the distributed parameter is outlined as, and therefore the penalty term is employed to stop from deviating from parameter . The Kullback–Leibler (KL) divergence [9] is employed during this study because the basis of social control. The mathematical expression of KL divergence is as follows:

$$kL(p||p_j) = p \ln \frac{p}{p_j} + (1 - p) \ln \frac{1-p}{1-p_j} \quad (2)$$

When doesn't deviate from parameter, the KL divergence worth is 0; otherwise, the KL divergence worth can bit by bit increase with the deviation? the price operate of the neural network is about as C (W, b).

III. PROPOSED METHOD

A. Data collected:

First, you must create a Twitter account to be able to use the Twitter API. It is not only difficult to fill out information on the Twitter site and the company will provide you with a password and a username by logging on to Twitter. You can read the tweets and send and receive comments on any topic you want. As for how to get a developer account, through your own account you can log on to Twitter developers by creating an account to write tweets through some important details. After the creation of a developer account can be used client key, the client secret key, access to the code key and access to the secret key in Access to Tweets.

B. Pre-processing:

1. Tokenization:

Tokenization is a process of splitting text strings into tokens, which are represented by words in sentences. As part of Twitter data analyses, this method will help with Named Entity Recognition. From the tweet dataset we can request sample data and tokenize them to see the results.

2. Removing numbers and punctuation:

Cleaning punctuation is a significant phase of the pre-processing. Punctuations does not alter the meaning of word and they cause a noise on input. Punctuations could also alter the word count as a word with punctuation and without punctuation are counted differently. Thus, apostrophes, semicolons, triple dots, exclamation marks, dots, commas, dots and other punctuations must be removed from the data set.

3. Changing uppercase letters to lowercase:

Words with upper-case or lower-case are counted in a different way in a case sensitive situation. Unfortunately, they share the same meaning in a sentence. Therefore, capital letters are converted to their lower case.

4. Stemming:

This preprocess step is performed on word roots. If base form of words has the same meaning, then they can be analyzed as one term. Without stemming, they could be assumed as different and unique words. Stemming process is needed to reduce word variants and increase the accuracy.

C. Feature Extraction:

After the pre-processing step, presently we are going to separate highlights from it and that we will try distinctive blends of highlights like unit-grams, POS labelling, twitter express highlights so on. Highlights are tokenized words and will be modified over to numerical vectors that are then spoken to as numerical data.

D. Algorithms Classification:

Before the test text is analyzed as vectors, the text is first converted to the carrier range to find word similarity. Then, preprocessing organizes this data. When text is processed, it is converted into a set of words and all characters without a meaning are isolated and removed. Besides, we have applied Deep Sparse Autoencoder classifier, it is popular and powerful tools for classification and prediction.

IV. RESULTS

Deep Sparse Autoencoder used to classify the features that extracted from dataset after pre-processing the dataset. Several parameters are calculated to evaluate the proposed method.

As shown in Table 1 the proposed method presented high results when 10 statistical parameters are calculated. Then, the obtained results compared with well-known studies presented in this field. According to the comparison in the Table 2 our method presented remarkable results compared to previous studies. As shown in the Table 2 the proposed method presented best results than methods proposed in [10,11,12] which these studies represented the commonly known researches in this field.

Table 1: Results of Proposed Method

Parameters	Results
Sensitivity	0.9600
Specificity	1.0000
Precision	1.0000
Negative Predictive Value	0.9600
False Positive Rate	0.0000
False Discovery Rate	0.0000
False Negative Rate	0.0400
Accuracy	0.9800
F1 Score	0.9760
Matthews Correlation Coefficient	0.9680

Maximum Accuracy after applying the classification technique is 0.98.

Table 2: Results Comparison

Methods	Results
Deep Learning with Bi-LSTM [10]	94
SVM [11]	91
Maximum Entropy algorithm [12]	70.04
Proposed Method	98.00

V. CONCLUSIONS

Opinion mining is a field where a large data volume is being generated via person-to-person communication. Analysing sentiments is a field used in various applications such as advertising, social media and forensic. By the assistance provided by opinion mining, the corporations could estimate their market and learn which changes considered to be necessarily needed to make for next product up gradation. Furthermore, providing mechanisms to construct a plan on their item. Customers can also use opinion mining for buying a product which they never used before, as the customer always prefer sentiments and reviews associated to the product before they purchase it.

REFERENCES

- [1] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimera, (2004) "Community analysis in social networks," *Eur. Phys. J. B*, 38(2), pp. 373–380.
- [2] G. Simos, (2015). "How Much Data is Generated Every Minute on Social Media," *We Are Soc. Media*, 19.

- [3] S. Wakade, C. Shekar, K. J. Liszka, and C.-C. Chan, "Text mining for sentiment analysis of Twitter data," in Proceedings of the International Conference on Information and Knowledge Engineering (IKE), 2012, p. 1.
- [4] M. Kang, K. Ji, X. Leng, X. Xing, and H. Zou, "Synthetic aperture radar target recognition with feature fusion based on a stacked autoencoder," *Sensors*, vol. 17, no. 1, 2017. View at Google Scholar · View at Scopus
- [5] J. Leng and P. Jiang, "A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm," *Knowledge-Based Systems*, vol. 100, no. C, pp. 188–199, 2016. View at Publisher · View at Google Scholar · View at Scopus
- [6] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996. View at Publisher · View at Google Scholar · View at Scopus
- [7] D. Yang, J. Lai, and L. Mei, "Deep representations based on sparse auto-encoder networks for face spoofing detection," Springer International Publishing, 2016. View at Google Scholar
- [8] J. Xu, L. Xiang, Q. Liu et al., "Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 119, 2016. View at Google Scholar
- [9] A. Youssef, C. Delpha, and D. Diallo, "An optimal fault detection threshold for early detection using Kullback-Leibler Divergence for unknown distribution data," *Signal Processing*, vol. 120, pp. 266–279, 2016. View at Publisher · View at Google Scholar · View at Scopus
- [10] Min-Yuh Day, Yue-Da Lin, *Deep Learning for Sentiment Analysis on Google Play Consumer Review*, 2017 IEEE International Conference on Information Reuse and Integration (IRI)
- [11] K. Amaravathi*, N. Lokeswari, 2018 IJSRSET | Volume 4 | Issue 8 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099 Themed Section : Engineering and Technology
- [12] Ankit Kumar Soni, *Multi-Lingual Sentiment Analysis of twitter data by using classification algorithms*, 978-1-5090-3239-6/17/\$31.00©2017IEEE