Advances in Parallel Computing Technologies and Applications
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC210146

Finding State of Mind Through Emotion and Sentiment Analysis of the Twitter Text

Ashok kumar PM [a,1], Anitha A [a], Verma H [a] and Laxmannarayana M [a]
[a] *Department of Computer Science and Engineering, K L Deemed to be University, Vaddeswaram, Guntur District, Andhra Pradesh, India*

Abstract. In this paper, the main aim of the project is to identify and do analysis of sentiment and emotion of the person and through the analysis find the state of the mind of the person. After finding the state of the mind of the person we can help people through NGO. We know that now top people are using social media twitter, and that place people are posting their thoughts and feelings. In this paper, our job is to do a twitter text analysis and make recommendations based on human emotions and also find state of mind of the person. Here we collect a tweet from the tweeter and their posts and make an analysis of this post. Emotional analysis is the study area for analyzing people's reviews, emotions, attitudes, and feelings from a tweeter in a written language. Emotional analysis has applications such as data collection and analysis of that data. However, the large volume and unstructured nature of text or data poses a challenge to properly analyzing data. Similarly, skilled algorithms or computer techniques are needed to mine and reduce tweets and find emotional words. Many of the existing computer systems, models, algorithms in sensory diagnostics from such informal data rely on machine learning techniques on the voice bag process as its basis. Understanding public opinion from a tweeter can help improve future decision-making. Comment mines are a way to get knowledge about online services from tweeter blogs, micro blogs , and social media. Individual opinions vary from person to person, and Twitter tweets are the most important source of this type of data. However, the large volume and unstructured nature of text / ideas data poses a challenge to analyzing the efficient data system. we know that millions of people are posting their reviews and comments on Twitter. By performing a tweeter analysis we will use other data science techniques to make an example, processing, classification of Bayes naive, k means algorithm integration, etc.

Keywords. Emotion Analysis, Sentiment Detection, Term Frequency-Inverse Document Frequency (TF-IDF), Naïve Bayes, Random Forest Classifier, Xgb Classifier.

1. Introduction

Emotion is nothing but a feeling, thoughts of the people, according to that person having their relationship with others. Emotion is a very important characteristic of the human

[1] Ashok kumar PM, Department of CSE, K L Deemed to be University, Vaddeswaram.
E-mail: profpmashok@gmail.com

which defines the personality and behaviour of a human. Human express emotion [1, 2] in daily life. They use different processes for expressing emotion. Mainly people express emotion by giving a speech or by facial expression. But in upcoming days people are using technology and social networking for expressing their emotions in the form of text or video or in audio. By reading the various research papers we found that many researchers in different fields (computer science, data science) are already focusing on recognizing and analyzing emotion from the text. And different authors applied different methods to assess the true mood of the people which is in the form of text. We know that humans having so many types of emotion sp for detecting the correct emotion from the text, it's too difficult . For example- Happy, Angry, sad, etc . We can represent the emotion in the form of polarity that is positive, negative, and neutral. If any human will say "Today I am very happy", means emotion of the human behind that, he is very happy we can say that sentiment polarity is **positive**. If any human will say "Today I am very sad", means emotion of the human behind that, he is very sad, then we can say that sentiment polarity is **negative**. Now in upcoming days, Twitter data is very popular for sentiment analysis. And we know that in the world maximum people are using Twitter to express their opinion and ideas.

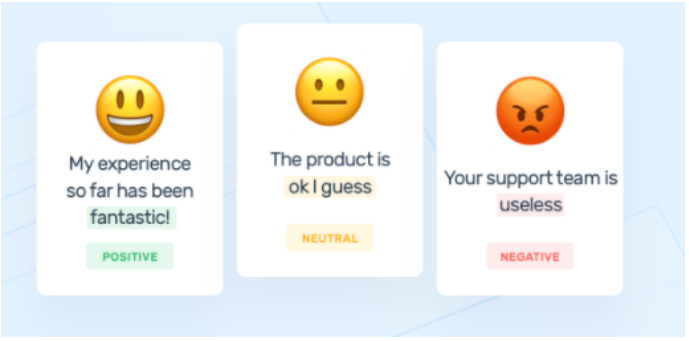

Figure 1. Different type of emoji of sentiment

2. Related Work

Emotions have already been addressed by religious, psychological, and philosophical researchers [1–5] since the beginning of the study fields. In 1872, Darwin attempted to explain the link between emotions and evolution. Following that, cognitive scientists [6] discovered that emotions are a product of other brain system processes. With development technology, emotional analysis was field of information technology research, also [20] performs the soil corpus transfer learning with better results. Because human emotions play such a significant role in human interaction, it was necessary to develop a method for robots to accurately classify emotions. Most methods and techniques for recognizing emotional responses through speech, facial expressions [7], body language [8], writing [9], and other methods of communication are related with current technology. The emotional interpretation of the text will be the emphasis of this study. Emotional discoveries in texts such as chapters, books, music lyrics, theatrical texts, logs, weblogs,

Facebook postings, twitter posts, and customer reviews were scrutinized for a prolonged period by the computers experts engaged. Investigators cited different sizes of people's feelings in various ways. 'Segmentorical' and 'Dimensional' models are the two types of emotions that exist [10]. Elman's, Shaver's, and Oakley's emotional models divide feelings and emotions into a few major categories. On the other side, state-of-the-art models [11] like Plutchik's model, Circumplex Model, OCC (Ortony, Clore, and Collins) Model, and Loveim's model segregate sentiments in great detail utilising a variety of sizes (e.g. data structures). Lists, trees, wheels, cubes, and other objects are used to create sub-species. Because of the enormous quantity of participants and posts in recent emotional analysis research, support platform [12] postings have been employed for textual emotional analysis. About 2.46 billion individuals use social media sites such as Facebook [13], Twitter [14], Instagram [15], and YouTube, and are members of one or more of them. Discovering and controlling the flow of information on any online community might require locating powerful persons and organizations.

Real-time data analysis [16] is not a one-time thing. Whenever data is not heard we need to do an analysis of why we should not use previous analysis results. It's growing method allows the existing effect to be updated using only new data conditions, without re-processing past events. This can be helpful in situations where all databases are not available if the data changes [17] over time.the paper. Do not number text heads-the template will do that for you.

3. Methodology

In this paper we are using machine learning algorithm for implementation. Following steps are –

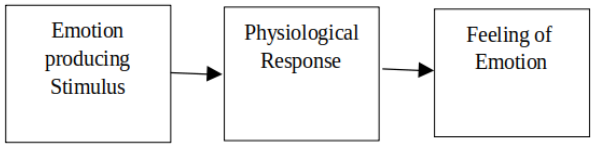

Figure 2. Emotion Calculation Flow

3.1. Data collection

With 330 million members worldwide, Twitter is among the most famous social networking networks today. People express their views on a range of political, governmental, and international issues through their daily lives. People communicate their thoughts in 280 character bursts and occasionally audio/video clips. The comments are known as twitter posts and are open to the general public. Some people have the ability to like, remark on, or modify postings. On Twitter, people may follow just one or even become buddies with each other. Twitter, unlike many other social networking sites, permits single-direction connections [18], which means that one user may follow another without having to re-establish the connection. This is a partnership that resulted in the creation of a communication network.

The database we utilized for our experiment comprises a collection of tweets, remarks, and replies, as well as their respective users' information. The 'Emotion in Text data set,' 'ISEAR,' 'Amazon Reviews,' 'US airlines data set,' 'EmmoBank,' 'TREC,' and other text databases are used for emotional and emotional analysis. Existing Twitter data sets, on the other hand, could not be used for our research since most of them contain messages or user-to-user communication. We wanted an emotional network [19] based on user content for our research, not on who follows who. We too needed feedback on those tweets and marketers / commenters details. In our emotional network, We required user communication based on their feelings about a certain issue. In our experiments, we've picked a few significant news and difficulties and will gather tweets with various emotions. *Syria keywords, *DonaldTrump, *SchoolShooting, *Christmas2017, *NewYear, *ValentinesDay2018, *Terrorism, *olympicgames2018, *WomensDay, *Oscars2018.

3.2. Data Pre-Processing

The impact of new data science approaches may be seen in the image below. Data analysis is a statistical technique that entails turning original data into an accessible format. Data from the real world is frequently imprecise, inconsistent, and/or deficient in specific behaviors or styles, as well as including numerous mistakes. Fixing data is a tried and true method of resolving such problems.

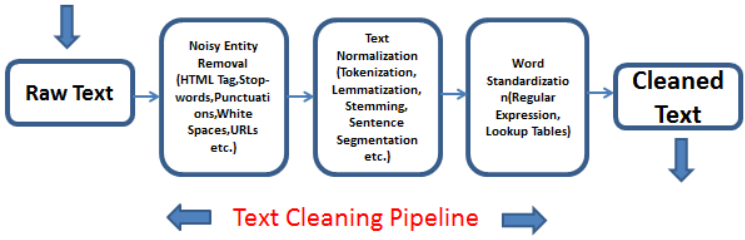

Figure 3. Text Cleaning Pipeline

We can obtain sentence tokenization (splitting text into sentences) as well if we wish to. However, we would have to include a pre-processing pipeline in our "nlp" module for it to be able to distinguish between words and sentences.

- **Decoding data**: In this stage, complex symbols in the twitter are converted in to simple text characters, inspite of encoding techniques like UTF-8.
- **Removal of Stop-words**: Most of the words in the English sentence are prepositions, articles, conjunction (like and, I, the, am, an, we) etc., which does not convey information for emotion detection. The job of the Count Vectorizer is that it removes all stop words present in the twitter text, which is already defined in twitter text.
- **Removal of Punctuations**: Another main difficulty in text processing is the presence of symbols like ".", ",","?". Some are important and needs to be processed and some needs to ignored. In this stage, punctuations are removed.
- **Removal of URLs**: Similarly, the presence of URLs and hyperlinks in tweet data in the form of comments, likes, dislikes, emoticons, reviews plays a major role in automatic understanding of detecting sentiments and emotions.

3.3. Feature Extraction

In this research, Term Frequency–Inverse Document Frequency (TF-IDF) are employed as feature representation for text analytics. Let's take a quick look at this. Count Vectorizer simply measures the number of times each word appears in the tweet text. First, we tokenize the whole sentence in to words and integerID is assigned. A count is maintained for each the occurrence of each token. Followed by, lower case transformation, punctuation removal, which occurs by setting all the parameter values to default values.

Tf-Idf Vectorizer is a single model that incorporates all of the features of CountVectorizer and Tf-Idf Transformer. Frequency of each word is maintained for each tweets. Further, Normalization is also done to convert in to Normalized feature vector.

3.4. Classification

We provide an outline of the methodologies used to identify sentiment analysis and emotion recognition in tweets in this section. We outline the general process involved in this paper:

Algorithm 1 Pseudo Code of Overall process

Input: Dataset D and Training data T

for each post TW in collection D **do**
    Perform Tweet Tw Extraction process in dataset D
    Cut C is initialized to the top root node Tweet Tw
    Extract the Feature Vector F
    Sub-set Features Fe are extracted from Feature Vector F
for each Sub-Set features E in Dataset D **do**
    Extracted features E are compared to training data T using Naïve Bayes/ Random Forest/ Xgb Classifier model and store result in P
if Polarity P is positive **then**
    Output the Positive Result
else if Polarity p is negative **then**
    Output the Negative Result
else
    Output the Neutral Result

3.4.1. Naïve Bayes Classifier

Naive Bayes may be a basic methodology for performing identification: Mathematical equations that predicts class labels to issue situations, expressed as vectors of feature integers, in which the category name-tags are derived from a fixed number. In many real situations, maximum likelihood is used to estimate parameters for naive Bayes models; in other terms, the naive Bayes model may be used without adopting Bayesian probability or utilizing Bayesian methods. Nave Bayes is a contingent ability to handle that assigns probabilities to the current instance given a drag instance to be categorized, represented by a vector $x = x_1, x_2, ...x_n$ reflecting some n properties (input variables), it assigns to the present instance probabilities. This model is used with a decision rule in the nave Bayes

classifier. The most prevalent approach is to choose the most likely hypothesis; this is known as the utmost a posteriori or MAP choice rule. A Bayes classifier is the function that assigns the category label for some k as follows:

$$P\left(\frac{C_k}{x}\right) = \frac{P(C_k)P\left(\frac{x}{C_k}\right)}{P_x}$$

$$y = argmaxP_{k=1,2,\ldots k}(C_k)\prod_{i=1}^{n} P\left(\frac{X_i}{C_k}\right)$$

	label	sentiment	text	len
0	3204	sad	agree the poor in india are treated badly thei...	270
1	1431	joy	if only i could have spent the with this cutie...	93
2	654	joy	will nature conservation remain a priority in ...	85
3	2530	sad	coronavirus disappearing in italy show this to...	94
4	2296	sad	uk records lowest daily virus death toll since...	69
5	4624	fear	joe biden's coronavirus web address lands on a...	111
6	2596	sad	respected sir in our telangana all private tea...	263
7	4131	fear	so is also 20 times more lethal than influenza...	63
8	3184	sad	thull is passing the most dangerous and ultra ...	249
9	3175	sad	thull is passing the most dangerous and ultra ...	248

Figure 4. Sentiment dataset

3.4.2. Random Forest Classifier

A random forest is a stochastic predictor that employs averaging to improve predicted accuracy and control overfitting by matching a range of decision tree classifiers on various sub-samples of the dataset. The random forest may be thought of as a classification method made up of numerous decision trees. It employs bagging and randomization in the construction of each individual tree in order to create an uncorrelated forest of trees whose committee forecast is more accurate than that of a single tree. The most important parameters during this are : n_estimators : Number of trees in random forest. Default is 10. criterion: “gini” or “entropy” similar to decision tree classifier. min_samples_split: the least proportion of working sets necessary to divide at each node. The default value is 2.

3.4.3. Xgb Classifier

We’d simply use a normal machine learning model, such as a decision tree, to train a single model on our dataset and utilise it for prediction. We might fiddle with the parameters or add more details, but we’re still working with a single model in the end. Even if we build an ensemble, each classifier is constructed and presented to our data independently. Rather of teaching all of the models separately, boosting teaches them in sequence, with each new model being learnt to rectify the faults made by the preceding ones. Models are added in a logical order until there are no more improvements to make. Gradient Boosting is a method of training new models to forecast the residuals (i.e. mistakes) of previous models.

4. Results and Analysis

In this experiment we taken emotion and sentiment data set which is available in Kaggle for sentiment analysis. Here we calculate accuracy by using different algorithm. The

Table 1. Comparison Table

Algorithm	Training Accuracy(%)	Validation Accuracy(%)
Decision Tree Classifier	82	78
Naive Bayes Classifier	88	83
Random Forest	95	90
Xgb Classifier	99	96

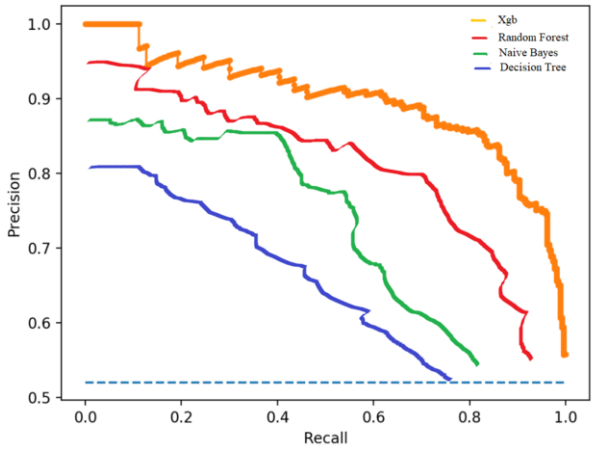

Figure 5. Comparison of Precision-recall curves of Various Classifier.

number of genuine positives separated by the entirety of genuine positives and wrong positives breaks even with accuracy. It shows how well a demonstrate predicts the positive lesson. The positive prescient value is alluded to as precision. Recall is calculated as the proportion of the number of genuine positives separated by the entirety of the genuine positives and the wrong negatives. Recall is the same as sensitivity. A precision-recall bend may be a plot of the exactness (y-axis) and the recall (x-axis) for distinctive limits, much just like the ROC bend. Clearly in Figure 5, Xgb classifier performs well both in cases of training and validation data sets.

5. Conclusion and Future Work

In this research, we performed experiments to detect to identify the state of the mind through Naïve Bayes, Random Forest, Decision Tree and Xgb classifier. At first, we performed data harvesting from different twitter accounts to perform data cleansing. In data cleansing, we removed special punctuations, stop, frequent, rare words. We converted emojis in to words. The specialty is we considered tweet replies , likes and dislikes also in this work. We used TF-IDF for extracting features in the tweets. Implememtation part

is done in two phases: a) Training and Testing Phase. In learning phase, Different models are implementation with the help of Naive Bayes Classifier, Random Forest, Decision Tree Classifier, Xgb Classifier. During validation phase, the trained model are tested with new data set. After experimental results, we found out that Xgb classifier performs well in detecting state of mind through tweets.

References

[1] Parveen H, Pandey S. Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. In2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT) 2016 Jul 21 (pp. 416-419). IEEE.

[2] Anand J, Sivachandar K. An edge vector and edge map based boundary detection in medical images. International Journal of Innovative Research in Computer and Communication Engineering. 2013 Jun;1(4):191-3.

[3] Garg Y, Chatterjee N. Sentiment analysis of twitter feeds. InInternational Conference on Big Data Analytics 2014 Dec 20 (pp. 33-52). Springer, Cham.

[4] Anand J, Flora TA, Philip AS. Finger-vein based biometric security system. International Journal of Research in Engineering and Technology eISSN. 2013 Dec;2(12):197-200.

[5] Sailunaz K, Alhajj R. Emotion and sentiment analysis from Twitter text. Journal of Computational Science. 2019 Sep 1;36:1-18.

[6] Zucco C, Calabrese B, Agapito G, Guzzi PH, Cannataro M. Sentiment analysis for mining texts and social networks data: Methods and tools. Wiley Online Library, Advanced Review. 2019:1-32.

[7] Patil HP, Atique M. Sentiment analysis for social media: a survey. In2015 2nd International Conference on Information Science and Security (ICISS) 2015 Dec 14 (pp. 1-4). IEEE.

[8] Wongkar M, Angdresey A. Sentiment analysis using Naive Bayes Algorithm of the data crawler: Twitter. In2019 Fourth International Conference on Informatics and Computing (ICIC) 2019 Oct 16 (pp. 1-5). IEEE.

[9] Sibia EV, Mareena G, and Anand J. Content Based Image Retrieval Technique on Texture and Shape Analysis using Wavelet Feature and Clustering Model. International Journal of Enhanced Research in Science Technology & Engineering. 2018;3(8):224-9.

[10] Munezero M, Montero CS, Sutinen E, Pajunen J. Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. IEEE transactions on affective computing. 2014 Apr 14;5(2):101-11.

[11] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media?. InProceedings of the 19th international conference on World wide web 2010 Apr 26 (pp. 591-600).

[12] Riquelme F, González-Cantergiani P. Measuring user influence on Twitter: A survey. Information processing & management. 2016 Sep 1;52(5):949-75.

[13] Kafeza E, Kanavos A, Makris C, Vikatos P. T-PICE: Twitter personality based influential communities extraction system. In2014 IEEE International Congress on Big Data 2014 Jun 27 (pp. 212-219). IEEE.

[14] Francalanci C, Hussain A. Influence-based Twitter browsing with NavigTweet. Information Systems. 2017 Mar 1;64:119-31.

[15] Lahuerta-Otero E, Cordero-Gutiérrez R. Looking for the perfect tweet. The use of data mining techniques to find influencers on twitter. Computers in Human Behavior. 2016 Nov 1;64:575-83.

[16] Joshi S, Deshpande D. Twitter sentiment analysis system. arXiv preprint arXiv:1807.07752. 2018 Jul 20;180:35-9.

[17] Sadhana SA, SaiRamesh L, Sabena S, Ganapathy S, Kannan A. Mining target opinions from online reviews using semi-supervised word alignment model. In2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM) 2017 Feb 3 (pp. 196-200). IEEE.

[18] Selvakumar K, Ramesh LS, Kannan A. Enhanced K-means clustering algorithm for evolving user groups. Indian Journal of Science and Technology. 2015 Sep 1;8(24):1-8.

[19] Sulthana AR, Jaithunbi AK, Ramesh LS. Sentiment analysis in twitter data using data analytic techniques for predictive modelling. InJournal of Physics: Conference Series 2018 Apr 1 (Vol. 1000, No. 1, pp. 1-8). IOP Publishing.

[20] Xu X. Machine learning-based prediction of urban soil environment and corpus translation teaching. Arabian Journal of Geosciences. 2021 Jun;14(11):1-5.