STUDY ON EMOTION RECOGNITION FOR GAUGING PERFORMANCE OF SUPERVISED MACHINE LEARNING ALGORITHMS

[1] Bhargav Jayaram Moorthy, [2]Prof. Deepika K

[1]PG Student, [2]Assistant Professor
[1] Department of MCA,
[1]RV College of Engineering, Bengaluru, India.

Abstract: The face is an index of the mind, but it is a word that really illuminates the emotions. Many emotions are hidden in subjective posts. Once recognised, these emotions serve as the knowledge base for many organizations. Many business decisions are based on these feelings. To make a decision, there is a need to identify users' emotions. Organizations that make emotional decisions could be profitable when the emotions are correctly identified. This paper uses supervised machine learning algorithms to accurately predict the sentiment of Twitter messages. This paper also aims to compare machine learning algorithms to predict the emotion, to generate the score generated by the model, and to compare the performance of the algorithms.

***Index Terms* - classification, dataset, deep learning, emotion recognition, processing, text mining, Twitter.**

I. INTRODUCTION

Modern advances in the field of deep learning have led to progressive innovations in natural language processing. User-generated data on social media is growing by leaps and bounds in graphs every day, plausibly every ten minutes. User-generated data content emerges in bulk thanks to the progress and introduction of social media, blogging, digital marketing, and a myriad of cross-media platforms. The increase in user-generated content is rich in legitimacy and selective and subjective opinions or feelings rather than those published on the basis of objective and objective data. These contents generated by subjective opinion absorb rich emotions, expressions and legitimate feelings. As the wise proverb says, the face is the index of the mind but it is the words that really illuminate the emotions, there are many hidden emotions in the subjective messages. These emotions, once recognized, can serve as a knowledge base for many organizations. Based on these emotions, many business decisions are made that make it necessary to identify emotions accurately to make a decision. Organizations that make decisions based on emotions can make a profit when feelings are accurately identified or can suffer losses if they cannot identify the right emotions [1]. Despite modern advancements in the field, Deep Learning also advances natural language processing, previous and existing systems for understanding user-generated content published by users focus primarily on simple lexicons and word bag models[2][3]. Paul Ekman [4] conducted an intense study on human facial sentient expressions which resulted in a defined series of six universally recognizable basic emotions: anger, disgust, fear, joy, sadness, and surprise. Robert Plutchik [5] defined a wheel diagram with two contrasting pairs and eight basic emotions. Joy, sadness, confidence, disgust, fear, surprise, expectation. This paper considers each of these emotions as a separate category, ignoring the various levels of intensity that Plutik defines on the wheel of emotions.

The mood profile [6] is a psychological tool to assess an individual's mood. Each sentient adjective contributes to one of six categories. For example, euphoria contributes positively to the category of joy. The imperative of this paper lies to categorise the emotions and further use it for prediction.

II. RELATED WORK

Applications of natural language processing for sentiment mining are discussed. This in turn also cites the ideology of the field of text mining having a popular rise in research. The paper's basic aim lies in finding the polarity of texts and further classifying it into three categories namely, positive, negative and neutral; aiding in human decision making. The paper throws light on surveying of the imperative approaches used for analysis [7].

The process of extraction of opinion rich contents generated by users on multiple platforms like blog sites, social media platforms, forums, etc. are explained. The survey consists of covering varied techniques and approaches that are subject to opinion-oriented information. The imperative lies in seeking to address the challenges of existing systems like issues regarding privacy, manipulation, and economic impact [8].

The paper discusses that the research on this topic is an ongoing topic for research. The paper discusses the outcomes of various recent algorithms and the impacted results in the field of opinion mining. The main aim of the paper is to aid with abstract measures to brief about the algorithms' enhancements on various models for mining sentiments [9].

Discussions about unstructured natural language texts are briefed. It also discusses the fact that acquiring good feature sets is a challenging task, yet it is important. A novel concept extraction algorithm is presented based on a parser to extract features. All the concepts in the paper are used to classify a set of documents as positive or negative [10].

DAN2 and feature engineering is used to study brand-related Twitter analysis for emotion. The mild SA is revisited with the approach the challenges thrown by unique characteristics of Twitter. Dataset on Starbucks is used to train the proposed algorithms resulting in the comparison of the approaches proposed [11].

The paper proposes a solution called SentiCircle for semantic representation of words. It is a lexicon-based approach for Twitter sentiment analysis. The difference between the other lexicon models and the proposed solution is that it considers the co-occurring pattern of words to capture the semantics and update and assign the pre-assigned polarities [12].

III. PROPOSED METHODOLOGY

This paper aims to compare algorithms to find out the best accuracy yielding model and use it for prediction. The methodology embraced in this project is agile methodology to iteratively develop in each incremental iteration. Compare the study of various algorithms implemented to check the accuracy, efficacy, and performance of each algorithm in terms of analysis, prediction and throughput. Supervised learning classification algorithms are used to perform the task. The main aim is to carry out preprocessing on the dataset to bring normalisation and to remove features that are mainstream. Counts obtained by studying from the unigrams and bi-grams are taken from the models. The methodology includes obtaining data, preparation on text- preprocessing, sentiment detection, classification and presentation. Data collection for the first stage of data processing. And given that the sources of information available pool warehouses. It is important to collect information (such as information retrieval and later), to form the sum of the available information, the sources are reliable and well structured. . After collecting the information, data and enter the time of preparation. Data preprocessing equipment is commonly referred to as the next step in processing the raw data swatch cleaning and organizing data. During the preparation, the raw data is carefully checked and is still safe. The steps to eliminate bad given that the (necessary, incomplete or inaccurate information) and start producing high quality information to improve business intelligence. Data entered into the computer data is also processed through a step prior to interpretation. Processing is done using machine learning algorithms, but the processed data sources (data, as social media, connected devices, etc.) and intended use (advertising research project, connected to medical device consumer health etc). The preprocessing step collects analyzed tweets and removes unwanted words, numbers, symbols and special characters. Preprocessing converts entire data to lowercase. If the data collected contains uppercase letters, boldface or words, they are converted to lowercase. The output of pre-processed tweets is more meaningful and easier to read than the stored tweets.

The naive Bayes classifier is a simple and powerful algorithm for classification activities. The Naive Bayes classification gives the best results when used for textual data analysis especially for natural language processing. Works well with conditional probabilities. Conditional probability is the probability of something happening because something else has already happened i.e., “given that something else has already occurred”. Conditional probabilities allow you to use previous knowledge to calculate the probability of an event.

The Bayesian naive classifier works as follows. Given that there is a set of training data D, where each tuple is represented by n-dimensional properties. Vector, X = x 1, x 2, ..., x n. It shows n measurements made on a tuple with n attributes or characteristics. Suppose you have the classes m, C 1, C 2, ..., C m. Given a tuple X, the classifier predicts that X belongs to C i iff it is: P (C i | X)> P (C j | X), where i, j ∈ [1, m] a n d i =/j. P (C i | X) is calculated as,

$$P(H \mid E) = P(E \mid H) * P(H) / P(E) \qquad (3.1)$$

Where,
P (H) is the probability that hypothesis H is true. This is known as pre-probability. P (E) is the probability of evidence factually P (E | H) is the probability of testing because the hypothesis h is true. P (H | E) Probability of hypothesis when there is evidence.

The Random Forest algorithm is a supervised classification algorithm. As the name suggests, this algorithm creates a forest with multiple trees. More the trees in the forest, the more robust the forest looks like. The random forest classifier was chosen for outperforming the decision tree algorithm for accuracy. It is essentially a bagging-based ensemble method. The classifier works as follows. Given D, the classifier first creates k bootstrap samples of D, where each sample points to Di. Di has as many tuples as D sampled with permutations from D. When sampled with permutation, some of the original tuples of D may not be included in Di and other tuples may occur multiple times. The classifier then builds a decision tree based on each D. To predict the use of the trained random forest algorithm, you need to pass a test function to the rules for each randomly created tree. Suppose you have formed 100 random decision trees from a random forest. Logistic regression is basically a supervised classification algorithm. For classification problems, the target (or output) variable y can only take discrete values for a particular set of functions (or inputs) X. This model creates a regression model that predicts the probability that a particular data item belongs to a category numbered "1". Logistic regression uses sigmoid functions to model data. This is the same as linear regression, assuming that the data follow a linear function and that the variable can have a binomial representation of only "0" or "1" s representing passage or failure. Logistic regression is predictive analysis. Use logistic regression to describe the data and the relationships between the dependent binary variables and one or more nominal, ordinal, or independent interval variables.

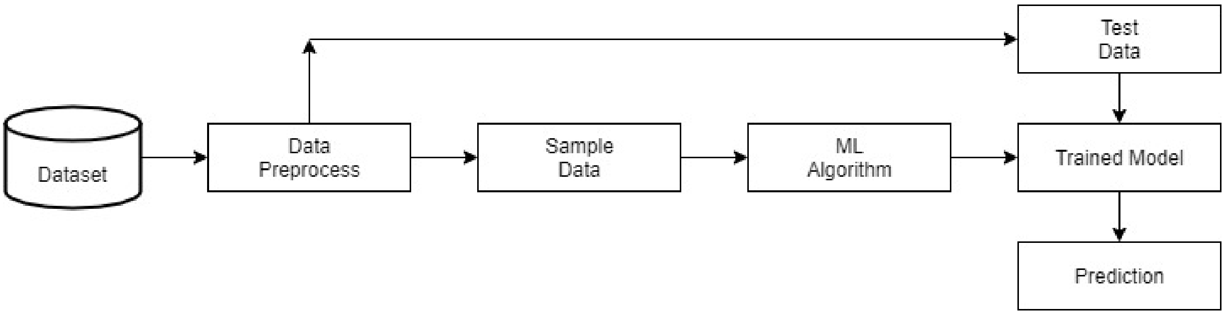

Fig. 3.1: Block diagram of the proposed project

The data is preprocessed and fed to the algorithm to train. The test data is fed to the model to validate. The user then chooses the input dataset and selects the model to be trained. The model returns with an Accuracy score percentile generation. The user then inputs the tweet for prediction followed by the model predicting the sentiment.

IV. Challenges

This section elucidates the challenging factors in the model that can occur while running the model. Data acquisition is an important factor. The quality of the data is directly proportional to the accuracy generated by the algorithm. Acquiring the right data is a challenging factor but given the right data in decent quantifiable measures, emotion prediction may not be challenging after all. The graphical user interface could be made more user friendly.

V. Results

This section elucidates the results derived after the experiment was conducted. The experiment was conducted on supervised classification algorithms. The following algorithms were used for the experiment; Naïve Bayes, Logistic Regression, and Random Forest.

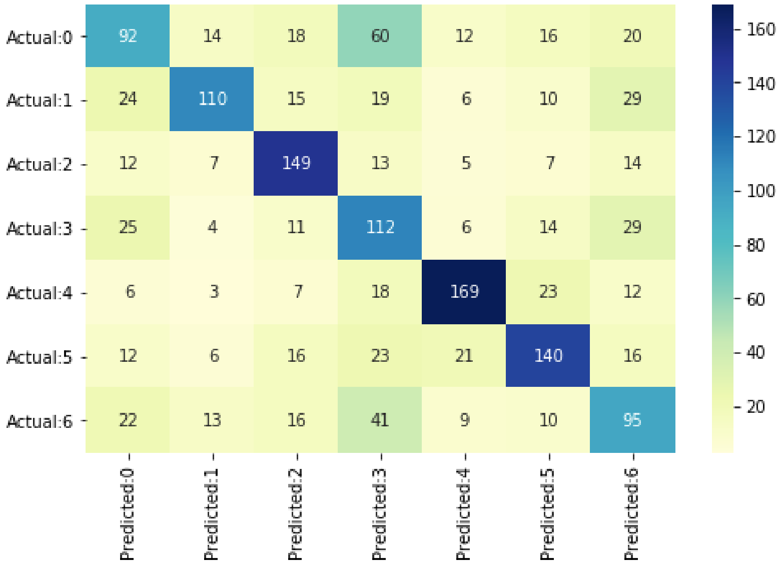

Fig. 5.1: Heat map of Naïve Bayes

Figure 5.1 elucidates the heat map of Naïve Bayes algorithm. Upon training the data and preprocessing it for training the model, the model gives a close to accurate in terms of accuracy. The chart shows the percentage of predicted words on actual words.

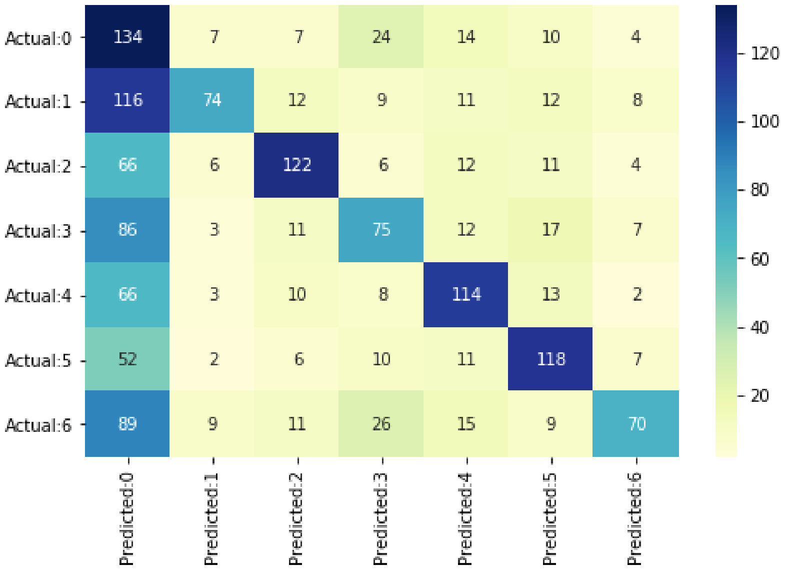

Fig. 5.2: Heat map of Random Forest

Figure 5.2 gives a broad picture of the output generated by Random Forest Algorithm. Upon preprocessing the data and feeding it to the model for training, the model gives a higher accuracy. The chart depicts the percentage of actual words and predicted words.

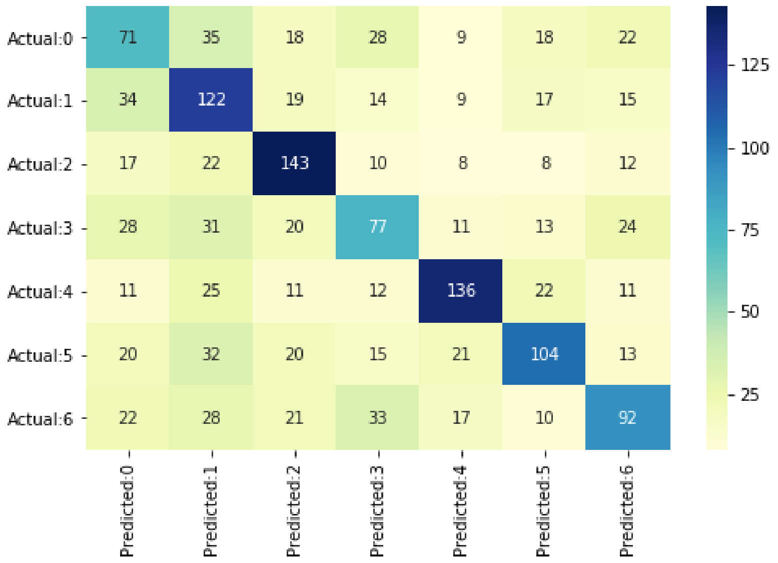

Fig. 5.3: Heat map of Logistic Regression

JETIR

Figure 5.3 elucidates the performance of Logistic Regression algorithm. The trained data is given as the input to the model, upon which the algorithm generates the accuracy almost similar to Naïve Bayes, both equally high.

VI. Conclusion

The proposed project has the aim centred to explore the possibilities to mine opinions from the sentient rich user generated content. “Data is the modern oil”, with that being said, it is critical to extract data generated by users on varied multiple platforms. Business organisations use opinion mining to enhance their products, analyse user sentiments over it, weigh brand value among the market, and also remain competent. The system is not only beneficial for business purposes, but also for research and development organisations. The project also boasts the comparison functionality between algorithms. The user enters the dataset containing the actual contents derived from twitter by users. Upon feeding the dataset, the user trains the model after selecting the algorithm to be trained. As specified in Figure 5.3, the process takes place. The model is trained and it generates the accuracy score percentile for the specific algorithm trained. The user then gives the input string to get the prediction.

References

[1] J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market,” J. of Computational Science, vol. 2, no. 1, pp 1–8, 2011.

[2] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, “The Predictive Power of Online Chatter,” Proc. of the 11th ACM SIGKDD Int. Conf. on Knowledge discovery in data mining, pp. 78–87, 2005.

[3] G. Mishne and N. Glance, “Predicting Movie Sales from Blogger Sentiment,” AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 155–158, 2005.

[4] P. Ekman, “An Argument for Basic Emotions,” Cognition & Emo- tion, vol. 6, no. 3, pp. 169–200, 1992.

[5] R. Plutchik, “A General Psychoevolutionary Theory of Emotion,” in Theories of Emotion. Academic Press, 1980, vol. 1, pp. 3 – 33.

[6] J. C. Norcross, E. Guadagnoli, and J. O. Prochaska, “Factor struc- ture of the Profile Of Mood States (POMS): Two Partial Replica- tions,” J. of Clinical Psychology, vol. 40, no. 5, pp. 1270–1277, 1984.

[7] Harpreet Kaur, Veenu Mangat, Nidhi, “ A survey of sentiment analysis techniques”, 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) 10.1109/I-SMAC.2017.8058315.

[8] B Pang and L. Lee, "Opinion mining and sentiment analysis", FoundTrends Inform Retriev, pp. 1-135, 2008.

[9] Medhat Walaa, Ahmed Hassan and Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal 5.4, pp. 1093-1113, 2014.

[10] Agarwal Basant et al., "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach", Cognitive Computation, vol. 7, no. 4, pp. 487-499, 2015.

[11] Zimbra David, M. Ghiassi and Sean Lee, "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks", 49th Hawaii International Conference on System Sciences (HICSS), 2016.

[12] Saif Hassan et al., "Contextual semantics for sentiment analysis of Twitter" in Information Processing & Management, vol. 52, pp. 5-19, 2016.