Sentimental Analysis Using Machine Learning Techniques

Aadhya Kaul

Abstract— Recently, with the increasing development of the web content from social media, some studies such as sentimental analysis have received the attention of some technological sectors, industries and governments across the globe. In recent years, sentimental analysis has not only emerged as one of the most gripping topics for research but has also become popular in machine learning and artificial intelligence due to its high potential in opinion mining and user-friendly recommendation. In today's era users all over the world are using social web platforms as a mode to express their thoughts, opinion, etc. about various subjects and due to which it becomes next to impossible to analyze such a huge user generated data manually therefore special and effective techniques are required for the analysis of such data and provide combat to the textual data by using natural language processing. In this study various machine learning techniques are used to judge and analyze the extent and type of emotion expressed through a particular text . This paper has introduced various feature extraction and word embedding techniques such as bow and tf-idf which are applied on various machine learning techniques. Many machine learning techniques that have been applied in accordance to sentimental analysis have been shown in the paper.

Keywords- Bag-of-words, Feature extraction, Logistic Regression, Random Forest, Sentimental analysis, SVM, TF-IDF.

1 INTRODUCTION

The division or classification of sentiments based on : polarity i.e. positive, negative or neutral is Sentiment Analysis. It requires detecting sentiments within the text data by the use of text analysis techniques.

In today's world examining the person has become much important in various realms. It may be any form of business, the owner wants to have the knowledge of the sentiments of his customers. They need to have an automatic sentiment analysis system for detecting the sentiments of their customers. A class of Natural Language Processing on the sentiments is sentiment analysis[4]. A small contrast between sentiment analysis and opinion mining is that opinion mining has a structure which analyses the opinion on the products but analysis of Sentiment using sentiment analysis is a type of opinion mining which detects the sentiments through the web.

The other methods which are used in the sentimental analysis are referred to as Stemming and Lemmatization also known as text normalization or word normalization methods in the sphere of natural language processing and these particular techniques are used to form words, texts and documents for the further processing. These techniques are also used to remove unnecessary words from the given text. Stemming is a type of method which is used to extract the base form of words by removing the affixes from them and it is also used in the process of indexing of words in the text, therefore on the stem of the words are stored by the searching machine .by these processes stemming reduces the length of the text and increases the accuracy. Another algorithm which comes under stemming is the porter stemming algorithm removes suffixes from the words present in the text and replaces and this is done with the help of **PorterStemmer** class present in NLTK. Lemmatization is also a method like stemming ,in lemmatization we get the output as 'lemma' which is a root word and not a root stem , the output of stemming .Lemmatization provides us an appropriate word which means the same thing . This process is done with the help of **WordNetLemmatizer** class provider by NLTK . The difference between lemmatization and stemming is that stemming removes the unnecessary part from the words while lemmatizing finds an appropriate word and both techniques improve.

Fig. 1 Workflow

- *Aadhya Kaul is currently pursuing Bachelor's degree program in information technology engineering in Bharati Vidyapeeth's College of engineering. E-mail: aadhayakaul@gmail.com*

2 RELATED WORKS

The sentiment can be defined as an attitude towards a particular situation or an event. It generally involves various types of feelings such as happiness, sadness, nostalgia etc. A great amount of research work has already been done in the domain of sentimental analysis which a sub field of natural language processing that helps in the identification and

extraction of the textual data present on various social media platforms .sentiment analysis also known as opinion mining and emotion AI is used for categorizing the given data whether it is in the textual form or visual form into positive ,negative or neutral categories[2,16] .Most of the data present on the internet is in the textual form as it is the most readable and natural form for presenting the thoughts and opinions to the users but the already available methods which are based on the visual or textual data do not provide satisfying solutions or results as it becomes tedious to extract information from a single mode of data so therefore to solve this a fusion between the textual and visual data by analyzing textual data through convolutional neural networks and NLP and analyzing visual sentiment representation from training examples , these two modes are mixed together in order to predict the polarity of the sentiments from the given dataset.

The two broad categories in which sentimental analysis can be divided are sentiment classification and feature based opinion mining and further it can also be classified into two types namely corpus-based approach and the other being the dictionary-based approach. The corpus-based approach involves placing the sentimental words into the corpus for learning purposes and by doing this sentiment score for the words can be obtained ,on the other hand in dictionary-based approach the sentimental scores of the given words are obtained by extracting the words from the text as well as using the lexical databases for the same, on the contrary feature-based opinion mining transforms the given text into feature vector in order to perform sentiment analysis by using feature engineering and learning algorithms for the transformation.

As we know there has been a lot of research done on the textual statics for analyzing the sentiments while the research on the visual data is still in its initial stages because analyzing the sentiments through the visual data is a laborious and tedious process due to many reasons[5]. As predicting sentiments through visual data is tough due to which deep learning models require a large amount of supervised data of images which is hard to gather, therefore the solution to this is the introduction of deep learning architecture which has shown immense success in the sphere of computer vision. Deep learning architecture involves convolutional neural networks for the categorization of various visual activities and predicting their emotion[15]. CNN models operate in a particular way for the identification of the sentiments through the visual data, the network has a multi-layered architecture which grasps feature representation from the raw pixels through layer-wise transformation and CNN eventually provides superior results as it has robust and accurate feature learning ability.

The other method which does not require expertise in natural language processing or deep learning for the prediction of the sentiments is known as VADER and it is also referred to as a lexicon, the algorithms used in this are optimized to sentiments demonstrated in many social media platforms[3].

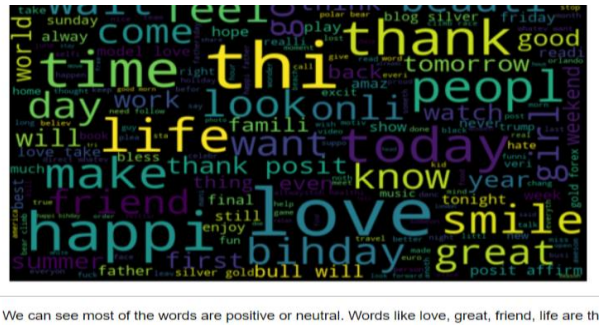
3 EXPERIMENTAL SETUPS AND METHODS

3.1 Dataset

The dataset used is a publicly available twitter dataset of tweets of different people. Data used is divided into train and test data.Train set has 31,962 tweets and test set has 17,197 tweets. The data is labelled as 0 and 1 for negative tweets and positive tweets. In the train dataset, we have 2,242 (7%) tweets labeled as racist or sexist, and 29,720 (93%) tweets labeled as non racist/sexist[8]. So, it is an imbalanced classification challenge. The tweets have unwanted text patterns like usernames, hashtags, punctuations, numbers, and special characters which have to be removed for classification. The words used in tweets can be classified as common words, racist words and non-racist words.

Fig. 2 Common Words Used

Fig. 3 Racist Words Used



We can see most of the words are positive or neutral. Words like love, great, friend, life are th

Fig. 4 Non-Racist Words Used

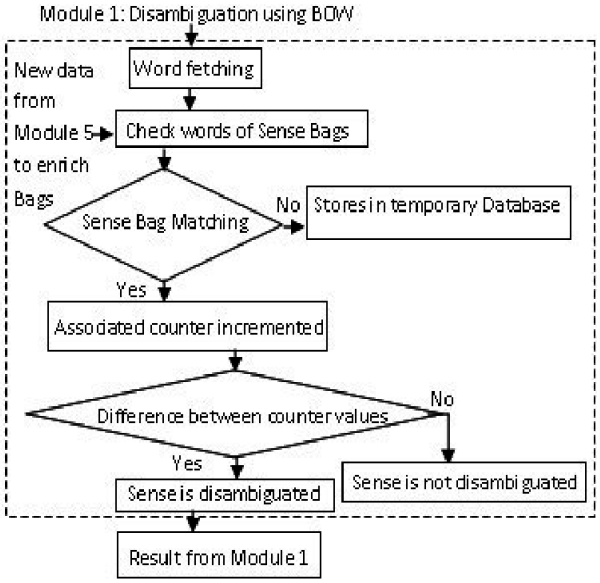


Fig. 5 Bag-Of-Words Flowchart

3.2 Feature Extraction

Sorting and merging the Factors into features, successfully decreasing the amount of data that has to be processed, Whereas still accurately and entirely describing the original data set is known as feature extraction.Word Embedding is a method where the text is represent using vectors.The popular forms of phrase embeddings are:

3.2.1 Bag-Of-Words

The BOW version is the easiest type of text representation in numeric form. As the term suggests, symbolizing a sentence as a bag of words vectors.It is a characterization of text that describes the frequency of phrases.
It consists of 2 components:

1. Thesaurus of known phrases.
2. Count of the existence of known phrases.

Bag of Words Model

1. Collect data
2. Designing the Vocabulary:
   Make a table of all the phrases in our thesaurus.
3. Create document Vectors

Assign each word a score. Scoring Method can be like "0" for absent words and "1" for present.

Further the vector for corresponding sentences can be formed according to the model vocabulary.

3.2.2 TF-IDF

Term Frequency-Inverse Document Frequency also referred to as TF-IDF which is a method that is commonly applied for text mining and additionally data processing.The primary intention of TF-IDF is to determine the significance of a specific phrase present in certain data or a text. The significance of that phrase existing in that the text is directly proportional to the frequency of that phrase occur in the text but is counterbalanced by the frequency of words present in the collection.Term Frequency-Inverse Document Frequency is split into two parts known as term frequency and inverse document frequency[14]. The part term frequency is defined as the proportion of the number of times the term appears in a specific text into the entire number of items present in that specific text.On the other hand the other part that's inverse document frequency is defined as the log of the entire number of documents upon the number of documents with term in it. The formula to calculate the weight of TF-IDF is given below.

$$TFIDF\ (term) = TF\ (term) * IDF\ (term)$$

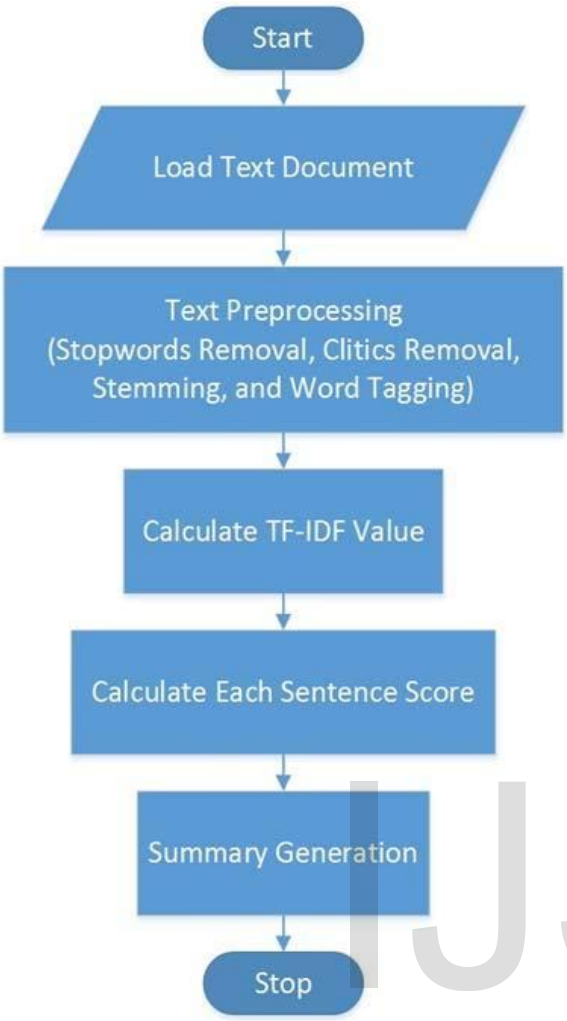


Fig. 6 TF-IDF Flowchart

3.3 ALGORITHMS USED

3.3.1 Logistic Regression

It is basically a model based on statistics, in basic form it uses a logistic function. It is based on the notion of chance , also known as the algorithm of evaluation, and is employed in solving classification problems in machine learning. It is used to assign observations into a set of the classes. By seeing from the mathematical point of view, a binary logistic model has two possible values- pass/fail which further has indicator variables. The values of these variables are "0" and "1". The value labelled "1" in the logistic model is a representation in linear form and combination of one or more predictors("independent variables"). The variables which are independent can be any, a continuous variable or a binary variable. The value which is labelled "1", its probability has the variation of value between "0"(value certainly "0") and "1"(value certainly "1"). The value which has been labelled "1" is also known as log-odds, which is the logarithm of odds. Logit is called the unit of measurement of log-odds[13].

The logit model can be represented as:

$$l = log_b(p/p-1) = {}_0 + {}_1 + {}_2 \qquad (1)$$

where l are the log odds, b is the base case of the logarithm and $_i$ are the values of parameters.
In this we have two predictors *x1, x2* and one response variable which is *Y* for which we denote p = P(Y=1).

Once the values of $_i$ get fixed, one can easily find the value of log odds, which results in either getting the value of log-odds as 0 or 1. The use-case seen in any logistic model is, given an observation(x1,x2) and then we have to compute the value of $\boldsymbol{p}$ such that the value of $\boldsymbol{Y}$ remains 1.

3.3.2 Naïve bayes

Naïve Bayes, is a supervised machine learning algorithm, one of easiest machine learning techniques to implement and at the same time is very reliable [6]. It is a very efficient computing method compared to all other machine learning techniques. It is used to determine the probability of the given data and classify it into positive or negative as done on the sentimental analysis dataset. Apart from being the most effective methods for machine learning, it is also a very efficient learning technique for data mining. Naïve bayes algorithm is derived from the bayes theorem which is used in finding the relation between likelihood of two occasions A and B. Naïve bayes has turned out to be a very useful technique in case of large datasets and is based on bayes theorem to determine the relation of probabilities between two occurrences. The conditional probability of two occasion A and B which is portrayed as P(A) and P(B) in which probability of occurrence A is conditioned by the probability of the event B and vice versa is shown as P(A | B) and P(B | A). therefore, the Bayes theorem is represented as [2].

$$P(A \mid B) = P(A)P(B \mid A)/P(B)$$

This formula entitles us to compute the relationship between two various occurrences and also helps us to find the probabilities of these two different events. By using this formula, we can evaluate the chances of an occurrence in view of the case of its event. Therefore, we can determine whether the event is certain or uncertain and can also classify the dataset into positive or negative. Some challenges which are still faced by the researchers in this method are looked upon and soon will be overcome.

3.3.3 Support Vector Machine

The motive of using SVM classifiers is to find an X -dimensional plane which must be a hyperplane(dimensions equal to the number of features) that classify the data points distinctly. Our approach starts with classifying two sets of data points distinctly, then we try to find the plane in which there is the greatest distance between the points of both the distinct classes(keeping in mind that the plane has maximum margin)[7]. The hyperplanes which we find are actually the decision making boundaries which help to classify the

various data points. The data points that lie on each side of this plane are classified and represent various classes in a combined data set.

The dimensions of a hyperplane depend specifically on the number of input features, which we want to classify. For instance,

1. When the number of input features to be classified is 2 then hyperplane will be a line.

2. When the number of input features to be classified is 3 then the hyperplane happens to be a 2-dimensional plane.

It becomes a difficult task to classify using SVM classifier when the number of input features to be classified exceeds the number 3. The cost function of SVM classifier can be represented as:

$$min\ 1/m\ _{i=1\ to\ m}[y^{(i)}cos\ t1(\theta^T\ x^{(i)}) + (1 - y^{(i)} +$$

$$\lambda/2m\ \Sigma_{i=1\ to\ n}\theta^2{}_j \qquad (2)$$

For the SVM we take our two logistic regression y=1 and y=0 terms described previously and replace with

1. $cost_1(\theta^T x)$

2. $cost_0(\theta^T x)$

3.3.4 Random Forest

This is an ensemble method which comes into function after creating a multitude of decision trees. This method is used for classification, regression and other tasks. The algorithms which combine two or more different algorithms to classify the same or various kinds of classifying objects. For instance if we take the predictions of SVM, Naive Bayes etc. then it will take a vote for the final consideration of the method for the test object[9].It's features include the following:

1. Being one of the most accurate classifiers for many data sets, it is the most accurate algorithm of all the available.

2. It's performance is very efficient when it runs on large databases.

3. Another advantage comes when it can handle thousands of variables without even deleting a single variable.

4. It has the ability of estimating the missing data and also it results in great accuracy when a large amount of data points are missing

Ensemble algorithms:

Ensemble algorithms are used for the combination of multiple algorithms for similar or different kinds of classifying objects.

Working of Random Forest Algorithm **:** Random forest classifier algorithm works as follows.

Firstly, the python libraries are imported, and dataset is loaded into the data frame

Secondly, the dataset is divided into train and test dataset.

Thirdly, random forest regression model is created and fitted into the training data

Lastly, the test set outcome is predicted, and a confusion matrix is made[12].

3.3.5 XGBoost

These days xgboost is one of the most widely used techniques in machine learning applications and has become very popular among the data scientists in the industry. The term xgboost actually refers to as extreme gradient boosting and its main purpose is to enforce certain algorithms to amplify the performance of the machine learning model and also increase the computational speed to perform those algorithms efficiently. Xgboost has the capability of solving problems on large scale and it can be used in complex projects with much ease .it is considered to be one of the most well-built classifiers among all the other classifiers present, it can solve any kind of regression, classification problems and give very fine results .xgboost mainly comprises of three types of boosting which are stochastic , gradient and regularized boosting.in this project of sentimental analysis xgboost has been trained to provide the intensity of the emotion present in the text and classify it into positive or negative.

4 RESULT

This paper signifies the usage of different feature extraction and word embedding techniques like BOW and TF-IDF on several machine learning techniques such as Logistic regression, Random forest classifier, Support vector machines, Naive bayes as well as XGboost.

The effectiveness of every individual model is based on the F1-Score and accuracy. F1 Score is the Harmonic Mean of the precision and recall. The values for the F1 Score lie between [0, 1]. It tells us about how much the precision of our classifier is. F1 Score aims to find a balance between the precision and recall[10].

Precision is known as the positive predictive value and Recall is called sensitivity.

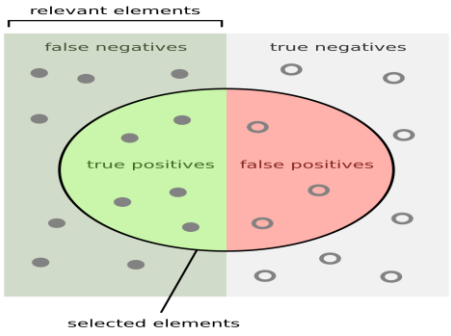


Fig. 7 Actual and Prediction Graph

$precision = TP/(TP+FP)$
$recall = TP/(TP+FN)$
$F1 = 2 * ((pre * recall)/(pre + recall))$

Accuracy is a good evaluator for classification models. This is the most commonly used metric to evaluate how well the model is doing.

$Accuracy = (TP + TN)/(TP + TN + FP + FN))$

From the analysis, we are able to comprehend that the XGBoost model works pretty well with our data improving the f1 score and accuracy relative to the other machine learning algorithms applied to our model.

TABLE 1

VALIDATION F1-SCORE

Algorithms used	Bag-of-Words	TF-IDF	F1-Score	Accuracy
Logistic Regression	0.53	0.54	0.62	0.94
SVM	0.50	0.51	0.61	0.94
Naïve Bayes	0.51	0.52	0.44	0.86
Random Forest	0.55	0.56	0.50	0.95
XGBoost	0.52	0.54	0.66	0.96

5 CONCLUSION AND FUTURE SCOPE

The main aim of writing the paper on sentiment analysis was to determine the sentiments and the opinions that a particular text exhibits[8]. In near future the demand of sentimental analysis will increase exponentially, as these days due to the swift increase in the consumption of social media. People are becoming more and more dependent on social media for entertainment purposes. Apart from entertainment, social media plays with the minds of people so it's important to analyze what people read on social media into positive or negative and protect people from getting negatively affected by such data available on these platforms. In this research paper various algorithms of machine learning have been tested on the twitter dataset to classify the sentiments of the given text into positive or negative.For this, various feature extraction methods are applied on some techniques such as Bag of Words in addition to TF-IDF applied on methods such as logistic regression, random forest classifiers, support vector machine and XGBoost and the effectiveness of this model is calculated by taking into consideration the F1-score and accuracy.

6 REFERENCES

[1]Mary, Sherin. (2019). *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review*. 10.1007/978-3-030-22868-2_90.

[2]Bhatt, A., Patel, A., Chheda, H. and Gawande, K., 2015. Amazon review classification and sentiment analysis. *International Journal of Computer Science and Information Technologies*, *6*(6), pp.5107-5110.

[3]Haque, T.U., Saber, N.N. and Shah, F.M., 2018, May. Sentiment analysis on large scale Amazon product reviews. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)* (pp. 1-6). IEEE.

[4]Coyne, Emilie & Smit, Jim & Güner, Levent. (2019). *Sentiment analysis for Amazon.com reviews*. 10.13140/RG.2.2.13939.37920. [5]Shrestha, N. and Nasoz, F., 2019. Deep Learning Sentiment Analysis of Amazon. Com Reviews and Ratings. *arXiv preprint arXiv:1904.04096*.

[6]Tan, W., Wang, X. and Xu, X., Sentiment Analysis for Amazon Reviews. [7]Yue, L., Chen, W., Li, X., Zuo, W. and Yin, M., 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, pp.1-47.

[8]Kim, H. and Jeong, Y.S., 2019. Sentiment classification using convolutional neural networks. *Applied Sciences*, *9*(11), p.2347. [9]Khondokar, M. and Islam, M., 2019. Sentiment Analysis from Social Media Image using CNN.

[10]Yang, P. and Chen, Y., 2017, December. A survey on sentiment analysis by using machine learning methods. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 117-121). IEEE.

[11]Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S., 2017. Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, *8*(3), p.27.

[12]Chen, X., Wang, Y. and Liu, Q., 2017, September. Visual and textual sentiment analysis using deep fusion convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)* (pp. 1557-1561). IEEE.

[13]Chen, L.C., Lee, C.M. and Chen, M.Y., 2019. Exploration of social media for sentiment analysis using deep learning. *Soft Computing*, pp.1-11.

[14]Ain, Q.T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B. and Rehman, A., 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl*, *8*(6), p.424.

[15]Rana, S. and Singh, A., 2016, October. Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)* (pp. 106-111). IEEE.

[16]Chandra, Y. and Jana, A., 2020, March. Sentiment Analysis using Machine Learning and Deep Learning. In *2020 7th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1-4). IEEE.