**Twitter Sentiment Analysis Techniques**

**Abstract**

Sentiment analysis is a technique in NLP that can be used to detect the emotion of a sentence. It is also termed as emotion or feeling mining. Online social media sites like Twitter, Facebook, YouTube, and many more have these days became a lot more expressive. People share their thoughts or feelings on various subjects. Data used in this project has been taken from popular microblogging site Twitter which was available at a site called Kaggle.com. In this project a thorough study has been done on various possible machine learning algorithms under the domain of supervised learning to perform sentiment analysis. **Index Terms**: Sentiment Analysis, opinion, emotions, Machine Learning, Accuracy, NLP, support vector machine

1. **Introduction**

Sentiment can be referred to as the emotion or the intention that the person wants to convey through his or her actions or words or text. Sentiment and properly understanding it is quite crucial in today’s era, for the role it plays in making organizations able to comprehend the needs and desires of their customers and producing accordingly and deriving profit. There are numerous ways in which humans or animals convey their emotions or sentiments, be it by gesture, facial expression, words or just through eyes. Expressing one’s feelings or emotions is a very basic human instinct or can be referred to as reflex as many of the times humans tend to convey their sentiment without any prior knowledge or unconsciously even when they wish not to. For example if a person is annoyed and he/she might not want to show it but it can be conveyed through their tone or voice modulation or pitch of their voice. Even writers convey their emotions through their stories or poems and interpreting these emotions depends on the reader as to how he understands the text, his/her level of intellect and ability to analyze as well as the state of his mind not to forget the choice of the words of the writer to convey his sentiment. The aim of this project is to find the best methodology which classifies sentiments into positive and negative for identification of hate speech. The following paragraph gives us an outline of the project.

The next segment explains the introduction of the sentiment analysis. From that point onwards, different types and applications of sentiment analysis are displayed. Section III gives us a brief description of the algorithms employed by us. Part IV is literature review. Part V explained the framework of the project and experiments performed in detail. Finally, the sections VI & VII are about results & conclusion and our future work.

1. **Literature Survey**

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Year | Paper | Methodology/Algorithm |
| 1. | 2020 | A Review On Sentiment Analysis Methodologies, Practices And Applications | SVM, Naïve Bayes |
| 2. | 2015 | Sentiment analysis using product review data | Random Forest, SVM,  Naïve Bayes |
| 3. | 2012 | Large-scale machine learning at twitter | Logistic Regression |
| 4. | 2020 | Literature Review on Sentiment Analysis | SVM, Naïve Bayes |
| 5. | 2019 | A Study on Sentiment Analysis Techniques of Twitter Data | SVM, Naïve Bayes |
| 6. | 2016 | A Topic-based Approach for Sentiment Analysis on Twitter Data | SVM |
| 7. | 2013 | Opinion Mining on Social Media Data | Naïve Bayes |

1. **Definition & Motivation**

Sentiment analysis is simply recognizing positive or negative sentiments written in textual form. Sentiment analysis is a natural language processing technique used to detect if the data is positive, negative, or neutral, rather we can say that it’s a study of human behaviour in which we extract user’s view opinion from the text.

It’s also known as opinion mining or emotion AI.

Now-a-days Hate speech is a common occurrence on the Internet. Analyzing hate comments or hate tweets is quite difficult owing to the subjective nature of sentiments due to the lack of a general definition of hate speech. Hence achieving the objective of diminishing hate crimes seems difficult at times. Hence there is a dire need of an efficient, automated and reliable system that will enable organizations to analyze millions of texts and understand the sentiment without any confusion which can be employed in the required sentiment analysis. Early identification of users promoting hate speech could enable outreach programs that attempt to prevent an escalation from speech to action. This motivated us work on the area of sentiment analysis for prevention of hate crimes.

**Types of sentiment analysis-**

Depending upon the individual’s or business’s requirement sentiment analysis can be divided based on different factors:

Sentiment analysis

Aspect based

Emotion based

Polarity based

Fig(1)

**1.** **Polarity based sentiment analysis:**

Polarity is referred as orientation of emotions expressed in a sentence. It can be positive negative or neutral. When only the polarity is concerned, we can detect only the polarity of the data by using man NLP techniques and classifying the data as positive or negative or maybe neutral.

**2. Emotions based sentiment analysis:**

Here we need to detect emotions (like sadness, happiness, anger) of the user or customer towards a service or product. These types of analysis use lexicons or complex algorithms to classify. Emotion detection may have useful applications, such as Gauging how happy our citizens are. Different indexes have different definitions; most evolve around economic, environmental, health, and social factors, Pervasive computing, to serve the individual better. This may include suggesting help when anxiety is detected through speech, or to check the tone of an email before sending it out, Understanding the consumer. Improving perception of a customer with the ultimate goal to increase brand reputation and sales.

**3. Aspect based sentiment analysis:**

In these types of systems, we can know the polarity of data towards a particular aspect or subject, or feature. For example, “I like the phone but the camera wasn’t too good”, here in this data, there is a positive polarity towards the phone and also a negative polarity towards the camera. Aspect-based sentiment analysis can be used to analyze customer feedback by associating specific sentiments with different aspects of a product or service. Aspect-based sentiment analysis allows businesses to automatically analyze massive amounts of data in detail, which saves money, time, and means teams can focus on more important tasks.

**Application of sentiment analysis in real life**

**1.Business and organization:**

   Sentiment analysis helps business organizations to analyze their product via people’s opinions on it and it also helps in perception for new product and in product or service benchmarking. Another use case of sentiment analysis is market and competitor research. Analyzing competitor’s content helps to find out what works with the public that company may not have considered.

**2. Individuals:**

  Sentiment analysis helps individuals to check people’s reviews of a particular product or service before purchasing it.

**3. Social Media:**

Sentiment analysis helps in gathering general opinion on a particular topic from social media.

In today's world social media is getting a complex place in which there can be too much negativity and hatred which needs to be stopped or else it can lead to hatred in society. So, hate speech needs to be monitored and filtered out this can be achieved by applying machine learning and sentiment analysis together.

**4. Ad’s placement:** Using sentiment analysis tools can help publishers and advertisers understand how consumers react to certain topics. Media companies can use those insights to deliver more poignant marketing messengers to consumers. And so on and so forth there a vast application of sentiment analysis in real life like customer support.

1. **Methods and Approaches**

**Machine learning approaches:**

Many machine learning algorithms are widely used to analyze sentiment.

Many complex algorithms are used to train a machine to perform sentiment analysis. here machines are trained using training data and then tested in test data.

**Supervised learning:** In supervised machine learning, we always have an independent variable and a dependent variable, using training data we determine the relation between them and then predict the dependent variable for distinct values.

a**. Logistic regression:**

It is a classification algorithm that determines the binary outcome of independent variables, Logistic regression is used to calculate the probability of a binary event occurring, and to deal with issues of classification. For example, predicting if an incoming email is spam or not spam, or predicting if a credit card transaction is fraudulent or not fraudulent. In a medical context, logistic regression may be used to predict whether a tumor is benign or malignant. In marketing, it may be used to predict if a given user (or group of users) will buy a certain product or not. An online education company might use logistic regression to predict whether a student will complete their course on time or not. The idea here is to divide the training set into positive and negative tweets. Count all the words and make a python dictionary of their frequencies in positive and negative tweets.

1. **Support vector machine:**

It is another supervised learning algorithm like linear regression but more advanced

Here the main purpose is to create a hyperplane in N-dimensional space where N is the number of features that are used to classify the positive or negative data with the help of Support vectors. To separate the two classes of data points, there are many possible hyper planes that could be chosen. Our objective is to find a plane that has the maximum margin. The maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

1. **Naive Bayes:**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

There are three types of classifiers:

Multinomial Naïve Bayes: Feature vectors represent the frequencies with which certain events have been generated by a multinomial distribution. This is the model typically used for document classification.

Bernoulli Naïve Bayes: In the multivariate Bernoulli event model, features are independent Booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks.

Gaussian Naïve Bayes: In Gaussian Naïve Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution.

Here a probability is assigned to a particular word that is considered positive or negative and classify the data.

1. **Random forest:**

As its name implies, a random forest consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes become our model’s prediction or the random forest model combines the predictions of the estimators to produce a more accurate prediction. Random forests are very good for classification problems but are slightly less good at regression problems. In contrast to linear regression, a random forest regressor is unable to make predictions outside the range of its training data.

Various Approaches Implemented

Sentiment Analysis

Supervised learning

Unsupervised learning

linear classifier

Decision tree classifier

Probabilistic classifier

Logistic regression

Support vector machine

Random rainforest

Naive bayes

Fig(2)

1. **Project Design and Methodology**

Data Collection

Data Processing

Data Analysis

Classification Algorithms

Sentiment Analysis

Feature Extraction

Fig(3) Framework of Twitter Analysis

1. Data Collection:

The data used in this project is taken from Kaggle.com. For training the models, we are given a labelled dataset of 31,962 tweets. Each tweet in the dataset consists of a tweet id, its label and the tweet. Following is a brief terminology associated with tweets. **Handle**: Users of Twitter use the “@” symbol to refer to other users on the microblog. Referring to other users in this manner automatically alerts them. **Hashtags**: Users usually use hashtags to mark topics. This is primarily done to increase the visibility of their tweets. In the sample of tweets are given labels, where label ‘1’ denotes the tweet is racist/sexist and label ‘0’ denotes the tweet is not racist/sexist.

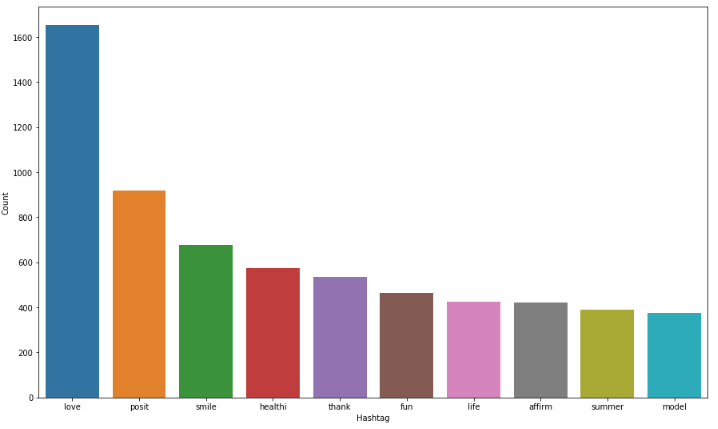
This figure shows a wordcloud in which the more frequent words are shown in bigger sizes.

 Fig(4)

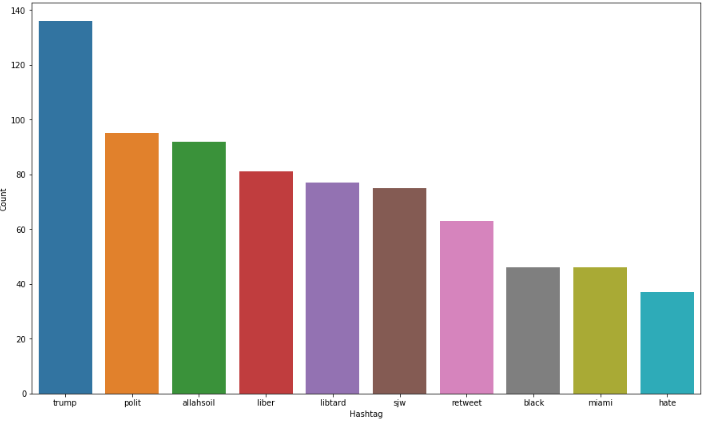
1. **Data Analysis:**

In this step various methods are used to visualize the content of data available.

These figures tries to depict frequencies of every hashtag.



Fig(5)



Fig(6)

Fig(5) shows all the most frequent positive hashtags and Fig(6) shows most frequent negative hashtags.

1. **Data Processing:**

Pre-processing refers to the extraction of several features from a set of random data collected and it can be achieved through the following methods:

* **Tokenisation**: Tokenisation is the process where every statement is analyzed and searched for unnecessary symbols like special characters, spaces, or as in twitter the @ symbols and then these symbols are removed from the dataset so as to get an outcome which only comprises of words. This makes the process of analysing the data set much more efficient and time saving as the data set might contain millions of texts.
* **Stop words/short words removal**: Stop Word Removal refers to the process of removing the large number of words which have the size of less than three letters which are also known as stop words and include the prepositions, verbs, pronouns etc. Following this the outcome will still have numerous unnecessary words which aren't as relevant for the feedback so these words are identified and removed the remaining text is completely filtered and contains the important words as well as reduces the number of words to be considered by a very significant amount.
* **Stemming**: The data still will be very lengthy and needs to be a lot more precise hence there has to be the process of Stemming which will enable to reduce each and every word to its basic or simply the root form and recombine the entire dataset to make sense of the information. Here the outcome that has been derived will be referred to as "cleaned tweets" which are much more efficient to study. We have recombined all the words to get our ‘cleaned tweets’.
* **Case Normalisation**: Further the data set comprises of letters of both upper and lower cases and hence it will be time consuming and inefficient to analyze such data, so the text is analyzed again and all of the letters are either converted into lower case or into upper case which is known as Case Normalisation.

1. **Feature Extraction:**

Sentiment tokens are extracted from the original dataset. They are also known as features, which will be used for sentiment categorization. In order to train the classifiers, each entry of training data needs to be transformed to a vector that contains those features, called a feature vector. To do this we have use a tool known count vectorizer. It is used to transform a given text into a vector on the basis of the frequency of each word that occurs in the entire text. We have extracted all the words and into a matrix called Bag of Words. The features are ranked according to their respective frequencies.

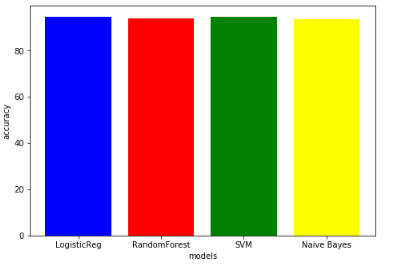
1. Classification Algorithms & Sentiment Analysis

We have split the dataset for training and testing in the ratio of 3:1. Accuracy is considered by us as our evaluation metric. We have implemented 4 models based on 4 different algorithms namely Random Forest, Support Vector Machine, Naïve Bayes & Logistic Regression. Even after achieving high accuracy our model suffers from a fundamental drawback. It doesn’t detect sarcasm. Consider the following sentence: “My flight’s been delayed. Brilliant!”. Most humans would be able to quickly interpret that the person was being sarcastic. Without contextual understanding, a machine looking at the sentence above might see the word “brilliant” and categorise it as positive.

1. **Results**

COMPARISON OF VARIOUS ALGORITHMS

|  |  |
| --- | --- |
| MODELS | ACCURACY |
| Random Forest | 94.01% |
| Support Vector Machine | 94.73% |
| Naïve Bayes | 93.56% |
| Logistic Regression | 94.69% |

Fig(7)

After comparing all the models, we found that SVM algorithm is most efficient in classifying data into positive/negative comments. It is able to do so with a accuracy of 94.73%.

1. **Conclusion & Future work**

Twitter is a source of vast unstructured and noisy data sets that can be processed to locate interesting patterns and trends. In this project, we mainly focus on the basics of sentiment analysis. There are various approaches and methods to identify sentiment from content. In this project, our examination represents various machine learning procedures in the domain of supervised learning. The study shows that machine learning method SVM, have the highest accuracy and can be considered as the baseline learning methods. In future work, discovering the result of various other text datasets and more models based on neural networks can be done. More work in the future is needed to improve performance measures and on prediction accuracy is needed.

**References**

1. Pooja Mehta, Dr.Sharnil Pandya, “A Review On Sentiment Analysis Methodologies, Practices And Applications”, International journal of scientific &technology research (2020)
2. Xing Fang and Justin Zhan, “Sentiment analysis using product review data”,Journal of Big Data (2015)
3. J. Lin and A. Kolcz, "Large-scale machine learning at twitter,“ ACM SIGMOD International Conference on Management of Data (2012)
4. Hemamalini, Dr.S. Perumal,”Literature Review on Sentiment Analysis” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH(2020)
5. Abdullah Alsaeedi, Mohammad Zubair Khan, “A Study on Sentiment Analysis Techniques of Twitter Data”(2019)
6. Pierre FICAMOS, Yan LIU, “A Topic based Approach for Sentiment Analysis on Twitter Data”, International Journal of Advanced Computer Science and Applications (2016)
7. Po-Wei Liang, Bi-Ru Dai, “Opinion Mining on Social Media Data‖”, IEEE 14th International Conference on Mobile Data Management (2013)