



**Twitter Sentiment Analysis Using Machine Learning**

Table of Contents

[**Introduction** 3](#_Toc102766883)

[**Sentiment Analysis** 5](#_Toc102766884)

[**Approaches to Sentiment Analysis** 5](#_Toc102766885)

[**Lexicon-based Approach** 5](#_Toc102766886)

[**Machine Learning Approach** 6](#_Toc102766887)

[**Supervised Learning** 6](#_Toc102766888)

[**Unsupervised Learning** 7](#_Toc102766889)

[**Problem Definition** 7](#_Toc102766890)

[**Aims and Objectives** 8](#_Toc102766891)

[**Literature Review** 8](#_Toc102766892)

[**Summary of IEEE paper** 8](#_Toc102766893)

[**Summary of ACM paper** 9](#_Toc102766894)

[**Proposed Methodology** 10](#_Toc102766895)

[**Dataset Collection** 11](#_Toc102766896)

[**Exploratory Data Analysis** 12](#_Toc102766897)

[**Data Pre-Processing** 19](#_Toc102766898)

[**Plot for Sentiments** 19](#_Toc102766899)

[**Word Cloud for Negative Sentiments** 20](#_Toc102766900)

[**Word Cloud for Neutral Sentiments** 21](#_Toc102766901)

[**Word Cloud for Positive Sentiments** 21](#_Toc102766902)

[**Calculate Term Frequency Inverse Document Frequency** 23](#_Toc102766903)

[**Selected algorithms** 23](#_Toc102766904)

[**1.** **Bernoulli Naive Bayes** 23](#_Toc102766905)

[**2.** **K Nearest Neighbor** 24](#_Toc102766906)

[**3.** **Decision Tree** 24](#_Toc102766907)

[**4.** **Random Forest** 24](#_Toc102766908)

[**5.** **Logistic Regression** 25](#_Toc102766909)

[**6.** **Support Vector Machine** 25](#_Toc102766910)

[**Evaluation Parameters** 26](#_Toc102766911)

[**Precision** 26](#_Toc102766912)

[**Recall** 26](#_Toc102766913)

[**F1 Score** 26](#_Toc102766914)

[**Accuracy** 26](#_Toc102766915)

[**Confusion Matrix** 26](#_Toc102766916)

[**Results** 27](#_Toc102766917)

[**Sentiment Analysis Results** 27](#_Toc102766918)

[**Evaluation** 27](#_Toc102766919)

[**1.** **Bernoulli Naive Bayes** 27](#_Toc102766920)

[**2.** **K Nearest Neighbor** 28](#_Toc102766921)

[**3.** **Decision Tree** 28](#_Toc102766922)

[**4.** **Random Forest** 29](#_Toc102766923)

[**5.** **Logistic Regression** 29](#_Toc102766924)

[**6.** **Support Vector Machine** 30](#_Toc102766925)

[**Comparison of ML Algorithms for Sentiment Analysis** 30](#_Toc102766926)

[**Emotion Analysis Results** 31](#_Toc102766927)

[**Plot for emojis** 31](#_Toc102766928)

[**Word cloud for anger** 32](#_Toc102766929)

[**Word cloud for joy** 33](#_Toc102766930)

[**Word cloud for optimism** 33](#_Toc102766931)

[**Word cloud for sadness** 34](#_Toc102766932)

[**Evaluation** 34](#_Toc102766933)

[**1.** **Bernoulli Naive Bayes** 34](#_Toc102766934)

[**2.** **K Nearest Neighbor** 35](#_Toc102766935)

[**3.** **Decision Tree** 35](#_Toc102766936)

[**4.** **Random Forest** 36](#_Toc102766937)

[**5.** **Logistic Regression** 36](#_Toc102766938)

[**6.** **Support Vector Machine** 37](#_Toc102766939)

[**Comparison of ML Algorithms for Emotion Analysis** 37](#_Toc102766940)

[**Conclusion** 38](#_Toc102766941)

[**References** 39](#_Toc102766942)

# **Introduction**

The microblogging sites such as Facebook, Twitter, Tumblr, etc. that are present in today's world can not only be used for social relationships, but they can also be used as an important source of information. One of the most popular social blogs of this type is Twitter, a website where millions of users go every single day in order to interact with others and exchange tweets. Twitter is one of the most effective means of finding out what the public thinks about any specific topic. The process of analyzing as well as monitoring these tweets can provide valuable feedback to both the public and private sectors, as well as to individual users. Since these tweets are large, a sentiment analysis technique could be the most suitable for the analysis of these tweets since it is easy to determine the opinion of the users by going through the tweets without having to go through millions of them manually [1].

One of the available techniques used in Natural Language Processing (NLP) is Sentiment Analysis, which works by analysing the content of texts in order to determine whether they are positive, negative or neutral in nature. As a matter of common practice, people use the sentiment to discover how they feel about a specific topic. As well as helping in obtaining an idea of an individual's opinion or attitude, it is also referred to as opinion mining. In the context of sentiment analysis, there are three methods by which this can be done: using a machine learning approach, employing a lexicon of sentiments, and employing a hybrid approach. A person's ability to categorize a tweet as either a positive or a negative response is something that comes naturally to them, however, this method of categorizing tweets isn't sufficient to deal with these vast amounts of data over the Internet today. Machine learning approaches have been developed to overcome these issues. It is based on a machine learning classifier that is used in the classification of data in a machine learning approach (ML) [2].

 As for Twitter, its default length is 140 characters, where the average is 13 words as minimum, 15 words as maximum on average. There are several issues which arise from the tweets include misspellings, informal language, awkward slang words, opinions regarding entities, as well as symbolic words which make it all the more difficult to process and analyse the tweets. Since the web holds a vast amount of content, organizations must in order to effectively use the digital medium, they must undertake a thorough analysis of the content available. It is now easier for readers to express their feelings to new companies online, as most of them have gone online. There is no doubt that these details are vital to both the companies in question and to a variety of other parties including government agencies, politicians and even advertising agencies. In light of the basic concept of sentiment analysis, there is an urgent need to develop methods for automatically classifying opinions of users. There has been a lot of work done on sentiment analysis that has mainly focused on building machine learning models that are efficient [3].

# **Sentiment Analysis**

Sentiment analysis, which is also known as opinion mining is sub field of text mining. We can define this type of psychology as measuring the feelings and emotions that people have towards a particular entity or subject. An analysis of the data can be done from a number of angles, such as applying natural language processing methods (NLP), using lexicons with annotations that provide word polarity information, and even taking advantage of machine learning techniques. As a direct consequence of this proliferation of user-generated content on the Web, the field of sentiment analysis has been extensively investigated in the researchers' field since around the year 2000 [4].

Taking into consideration the availability of such voluminous information on the Internet, the task of manual analysis of such data is difficult and error prone for humans. As such, an automated analysis of the content is essential. Sentiment analysis can be used to achieve this. It is no longer necessary for consumers to ask their friends or family about the quality of a product, when the answers are readily available for them to find out. As well as product reviews, sentiment analysis is being used in the analysis of news articles and social media posts as well. The results of the sentiment analysis can be very useful in performing market research, as well as when deciding on political or government policies. In this project, we will determine the sentiments of Twitter data for which Ukraine Russia conflict tweets have been chosen as the data set on which our predictions will be based.

A lot of the decision-making processes of different organizations are driven by the opinions of users. As a result, the quality of service, as well as the enhancement of deliverables, are enhanced. In addition to reviews sites and blogs, microblogs (Twitter) as well as other datasets, there are a variety of sources of data to be found.

## **Approaches to Sentiment Analysis**

The main two methods for analysing sentiment in a platform are the lexicon-based approach and machine learning approach.

### **Lexicon-based Approach**

In this approach, lexicons and dictionaries are utilized. It is in this step where semantic orientated words or phrases that form part of the polarities of documents are used to determine their orientation. This algorithm makes use of lexicons or dictionaries in order to calculate a document's orientation, based on the structure of the words. In the domain of semantics, semantic orientation (SO) to capture the polarity and strength of words or phrases in the text is a measure of the degree of subjectivity and opinion. A document's overall sentiment orientation is determined by all of these words. Opinion lexicons can be created either manually or automatically based on a variety of factors. A manual approach to creating opinion lexicons can take a considerable amount of time, so it is necessary to include this into automatized methods of creation. There are two different types of manual lexicons, namely common lexicons and category specific lexicons. There are two types of sentiment words in the common lexicon: default words with the same sentiment value, split words, negation words and words with blind negation[8].

### **Machine Learning Approach**

The machine learning algorithms are divided into two categories one named as supervised and other as unsupervised. When using unsupervised learning methods, the training data is not used in order to classify, whereas supervised learning algorithms use training data before attempting to classify, i.e., a training phase and a testing phase that is used to validate the classification”[9].

### **Supervised Learning**

A well-labelled corpus is essential to the training of a classifier in supervised learning. The process of supervised learning is made possible by the use of various algorithms. In order to perform supervised learning with well-defined labelled data, it is essential to ensure that there is a proper definition of the data, without which training and testing of the model would be difficult. Based on the classification of supervised learning, there are two main types of it that are regression learning and classification learning. Regression involves labelling the data sets with labels in order to train the models, then using that data to build models, and then using those models to predict and improve the model by working iteratively on it. It is the purpose of classification to find the class label that can be used to provide the appropriate classification based on the positive, negative or neutral sentiment of the message. The principle of supervised learning is that the algorithm works on basis of labelled data to train an algorithm to classify and predict the sentiment of tweets based on the tagged data. As a result of supervised learning, decisions trees, random forests, and Nave Bayes classifiers are all algorithms that are found[10].

### **Unsupervised Learning**

Machine learning or lexicon can be used for unsupervised learning. In this method, only the data is put as input to the model, a label is not required on the model itself. An unsupervised learning process can also be known as pattern recognition. Clustering is one of the unsupervised learning processes. In general, a Sentiment Lexicon is used to implement an unsupervised approach to sentiment analysis. Several algorithmic approaches have been developed to classify texts based on semantic orientation, which involve extracting phrases from texts that contain adjectives and adverbs in order to estimate the semantic orientation of these phrases. An in-depth semantic analysis of the reviews is then used to classify them [11].

# **Problem Definition**

With the increasing number of social conflicts occurring across the globe, it is becoming ever more crucial to understand the opinions and feelings of the general population around the globe as well. Since the surveying of the whole population would take a lot of resources as well as take a lot of time to complete, it would be better for us to adopt a more modern approach. There is no doubt that the proliferation of social media, especially Micro-Blogs, has expanded the possibilities of raising awareness about issues. The aim of this research is to find emotion and sentiment from microblogging platform Twitter in the form of emotion and sentiment analysis. Currently having over 284 million active monthly users and having 500 million Tweets being generated each day, Twitter is making a significant impact on a significant portion of the population who aren't afraid to share their opinions. Our goal with this project is to provide a fast, accurate, and reliable way for people to extract sentiments and emotions from the data provided by Twitter.

Our problem is ***classification- based***. The ***Classification algorithm*** is a technique of supervised learning which uses training data to analyze new observations and identify the category on the basis of those observations. Classification is the process of learning to classify new observations from the presented dataset and then categorizing the new observations into a number of classes and groups.

Our hypothesis is that “*Given a set of tweets that contain multiple features and a wide range of different opinions, the goal is to extract opinions or expressions that describe a target feature, and categorize the opinions positively, neutral or negatively”.*

***The novelty of proposed work is that this topic is latest topic and not much work is found in literature, as it is ongoing topic and also our proposed method will achieve better results in terms of accuracy.***

# **Aims and Objectives**

As a primary objective of this study, we are interested in evaluating the performance of machine learning algorithms for sentimental analysis.

**Objectives:**

* To determine the algorithms and metrics that can be used to evaluate the performance of Machine Learning Classifiers.
* The purpose of this study is to identify which algorithm will outperform.
* To determine the sentiments of people towards the conflict of Russian Ukraine war.
* To determine the emotions of people towards the conflict of Russian Ukraine war.

The objectives are achieved after experiment is conducted, examining the performance metrics for the identified algorithms; this is followed by a comparison between the metrics for different algorithms. From the results obtained, the best algorithm which achieves best accuracy as compared to other algorithms for sentimental analysis of Twitter dataset will be identified.

# **Literature Review**

 An Analysis of Sentiment is a means to segregate positive, negative or neutral sentiments contained in texts based on the way they are expressed in the text. An analysis of sentiment can also be referred to as polarity or opinion mining. A growing number of people are engaging with social media platforms as a result of the advancement and growth of these platforms. 280-character Twitter-like social networking service where users are able to send their It becomes easy for sentiment analysis to analyse tweets because of the small number of characters that they have. for the sentiment analysis. There are 550 million tweets posted every day on Twitter. There is a fair representation of all age groups of people on Twitter as well as people from all genders. Hence, tweet analysis becomes somewhat representative of society's sentiments due to the sentiment analysis of twitter data.

## **Summary of IEEE paper**

This paper [12] aims to provide a comparison between Naive Bayes Classification method, Support Vector Machine Classification method, Maximum Entropy Classification method in order to draw conclusions about the effectiveness of each of these models.

This article discusses the various machine-learning techniques that are used to analyse Twitter data, such as Naive Bayes, SVM, and Maximum Entropy Methods. As part of the twitter analysis, certain aspects are analysed in order to mine sentiments in the text. The goal of this dissertation is to analyse the concept of opinion in relation to Twitter sentiment analysis.

In the study, it was found that the machine learning method, such as Naive Bayes, had the highest level of accuracy and could be considered as an appropriate base learning method. It was found in the study that Maximum Entropy methods can also be very effective in some cases as well.  The results are shown below:

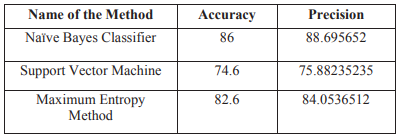


Figure 1: Comparison of accuracy of machine learning classifiers in paper.

A lot more effort needs to be put into improving the performance measures in future.

## **Summary of ACM paper**

The focus of this paper [13] is on using machine learning algorithms and Scikit-learn to analyse the sentiment of Twitter data in the sense of sentiment analysis. In order to achieve this, they analysed Twitter datasets which are publicly available via the NLTK Corpora, and take advantage of feature extraction techniques in order to create an efficient feature. They train and test various machine learning classifiers such as Bernoulli NB, SGD classifier, SVM, Multinomial Naïve Bayes, Logistic Regression, LinearSVC and NuSVC. As a result of experiments, it was demonstrated that Logistic Regression, SGD classifiers and BernoulliNB, are capable of achieving classification accuracy of at least 75%. The results are shown below:

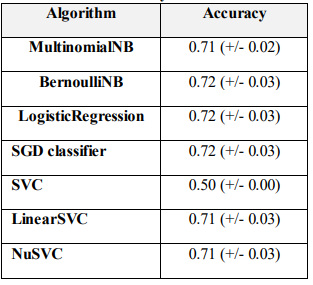


Figure 2:Comparison of accuracy of machine learning classifiers in paper.

In our work we are using various machine learning algorithms and perform classification. We have selected a dataset which has more features as compared to above papers. Our twitter dataset is about the tweets regarding conflicts between Russia and Ukraine conflict. We are performing sentimental analysis and also emotion analysis. We will see how people are taking this conflict and what are their emotions by visualizing the results. Our approaches will provide better results and increased accuracy as compared to above papers.

# **Proposed Methodology**

The proposed methodology is shown in below figure.

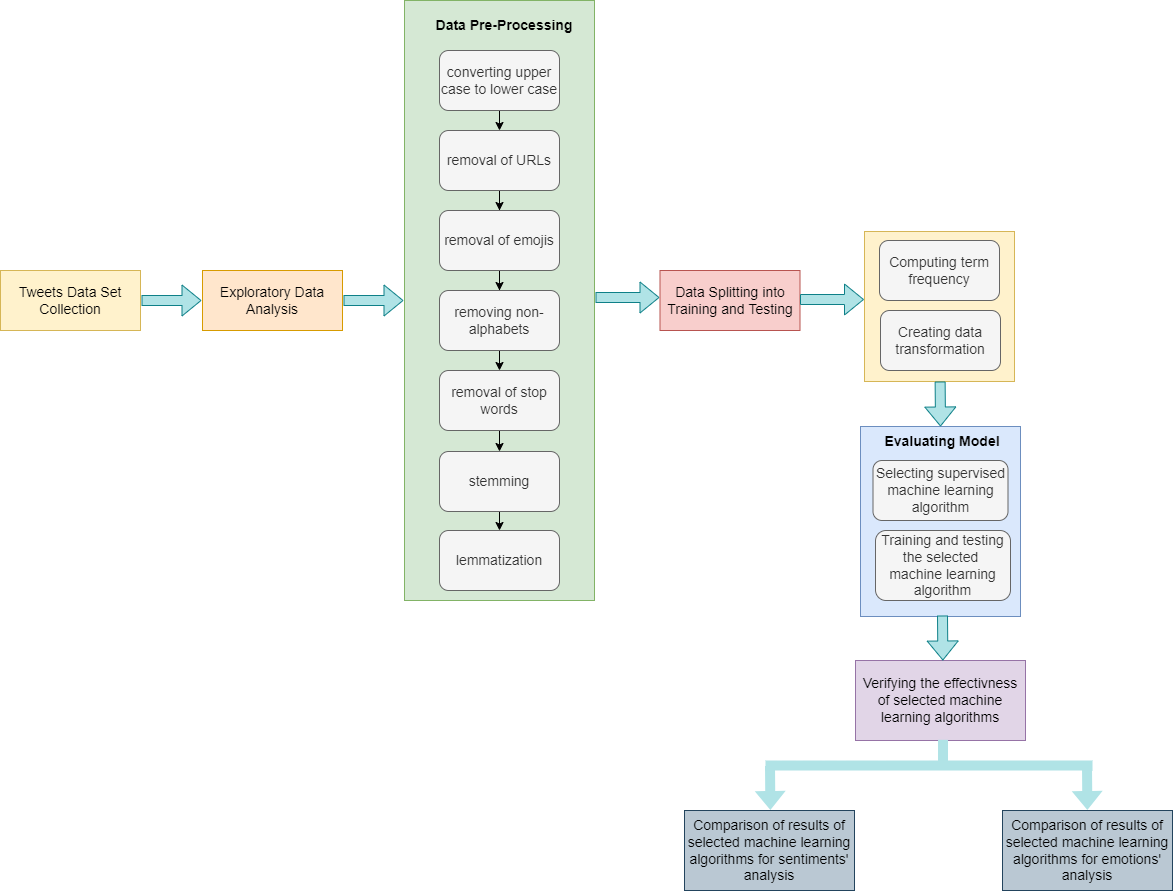


Figure 3: Workflow of proposed methodology.

## **Dataset Collection**

For the proposed project, we have selected dataset on topic of Ukraine-Russia Conflict topic as it is hot topic these days.

The invasion of the Ukraine by the Russians took place on the 24th February 2022. As with the ground war, so does the social media war start. This social media war has been a global effort and involves people from many countries [14]. In terms of popularity, Twitter is one of the most prominent microblogging services. As of now, there are multiple hashtags with tweets related to the Ukraine-Russia conflict attracting millions of mentions [15].

The dataset is downloaded from Kaggle[16]. A dataset of tweets relating to the ongoing conflict between Russia and Ukraine.

The dataset is a very small fraction of what is available in the real world, but may be able to help in quite a few different ways.

* What is the general feeling of the Twitter verse regarding the ongoing conflict.
* What are the sentiments and emotions amongst the personalities, countries, celebrities, and the like regarding this conflict.
* Based on the micro trends in the topics linked with the conflict in Ukraine, one can calculate the evolution of conflict on a daily basis.

## **Exploratory Data Analysis**

There are 20 M tweets in this dataset.

To extract the features required for proposed approach we perform dataset analysis and arrange tweets in following ways:

* **Dataset information**

First, we need to see the data in the dataset as follows:

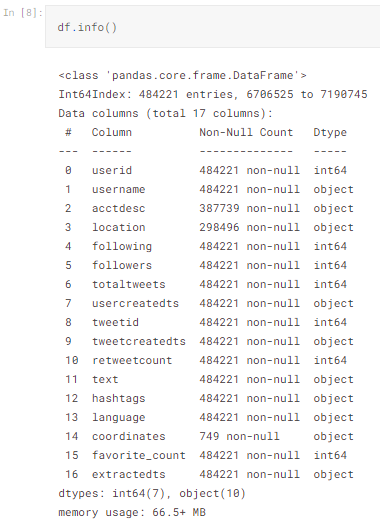


Figure 4: Information about dataset.

* **Arranging tweets according to language**

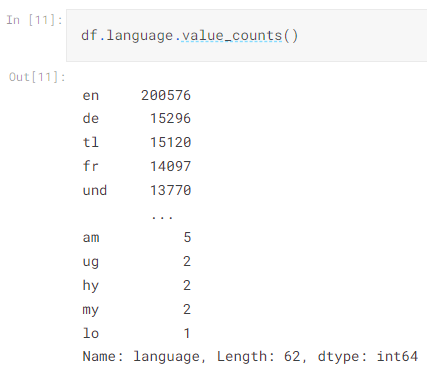


Figure 5: Arrangement of tweets according to language.

* **Plotting the tweets**

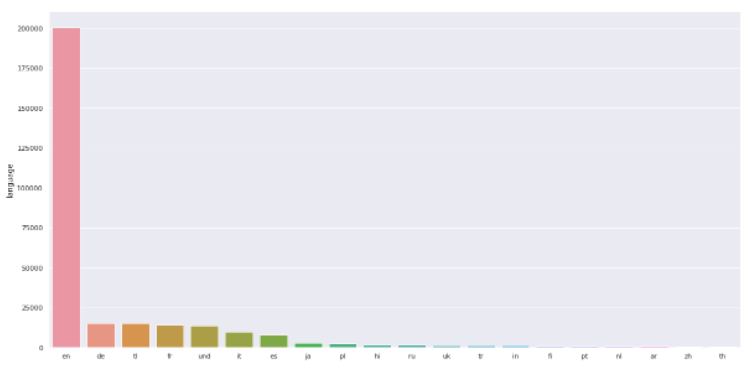


Figure 6: Plotting of tweets present in dataset.

* **Language-based unique tweets**

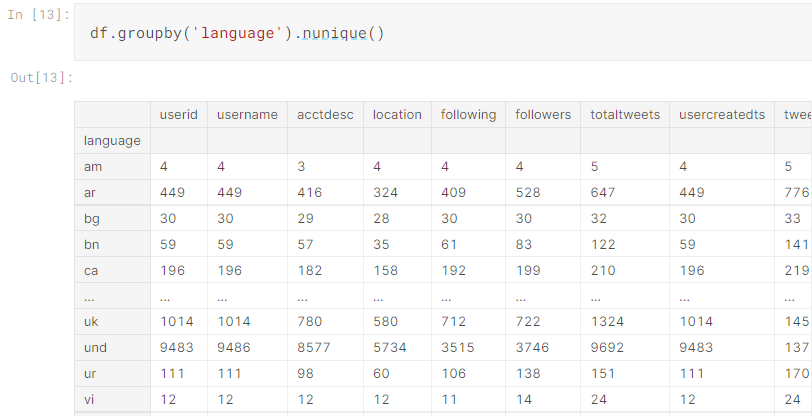


Figure 7: Unique tweets in dataset on language basis.

* **Most retweeted tweets**

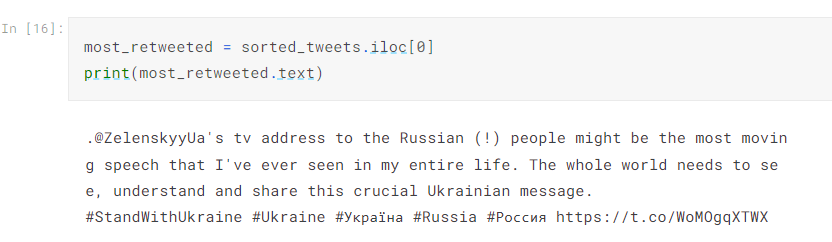


Figure 8: Most retweeted tweets in dataset.

* **Word Cloud**

The Word Cloud is a type of data visualization that enables, for instance, how to visualize text data, encompassing these very words in a visual representation that indicates their frequency and importance statistically [17].

The word cloud of tweets is shown below:



Figure 9: Word cloud of tweets present in dataset.

* **Plotting the tweets according to their location**

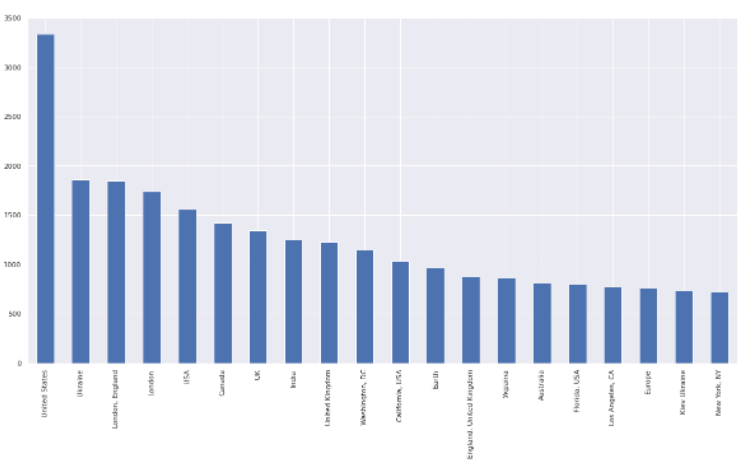


Figure 10: Plotting of tweets according to location.

* **Word Clouds using based on unique tweets**

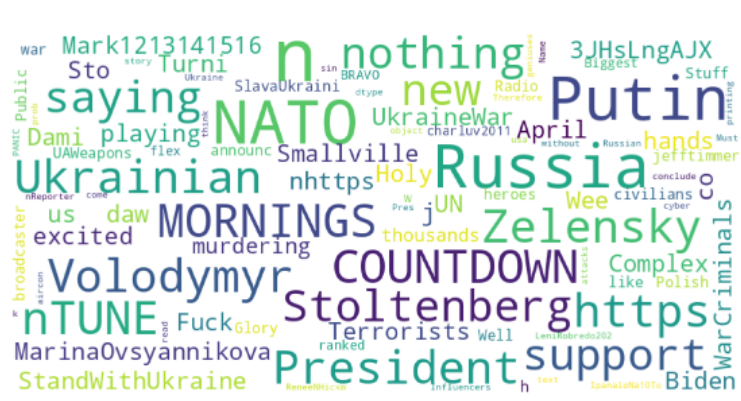
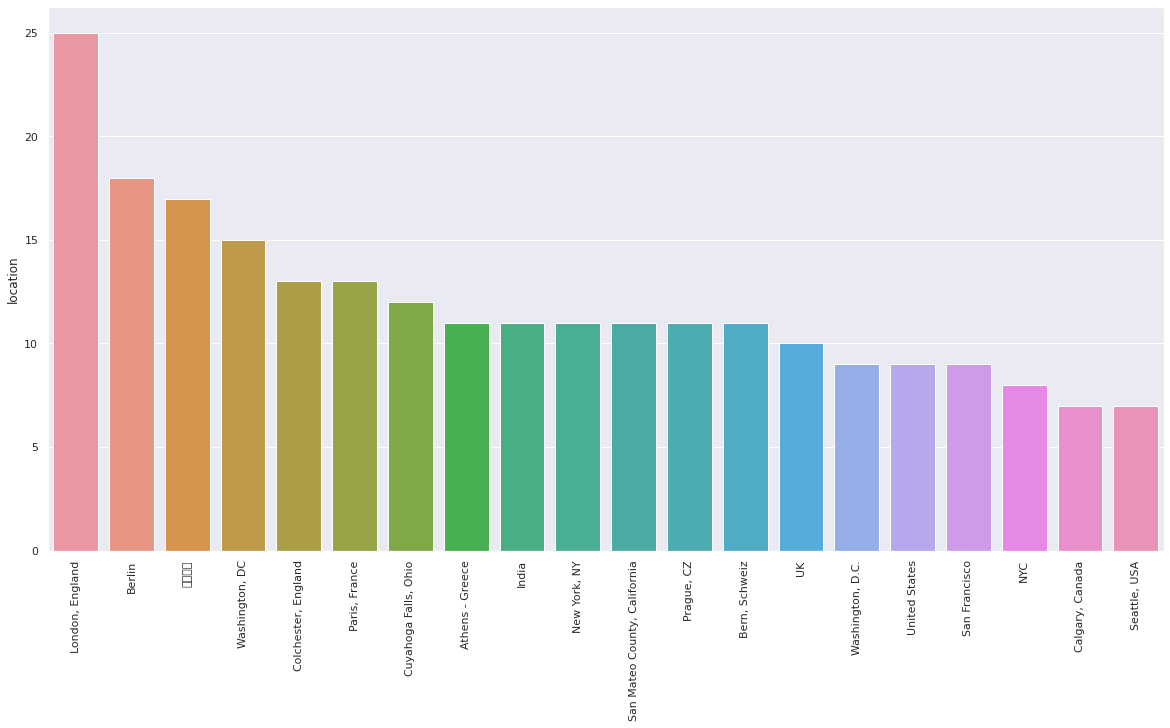
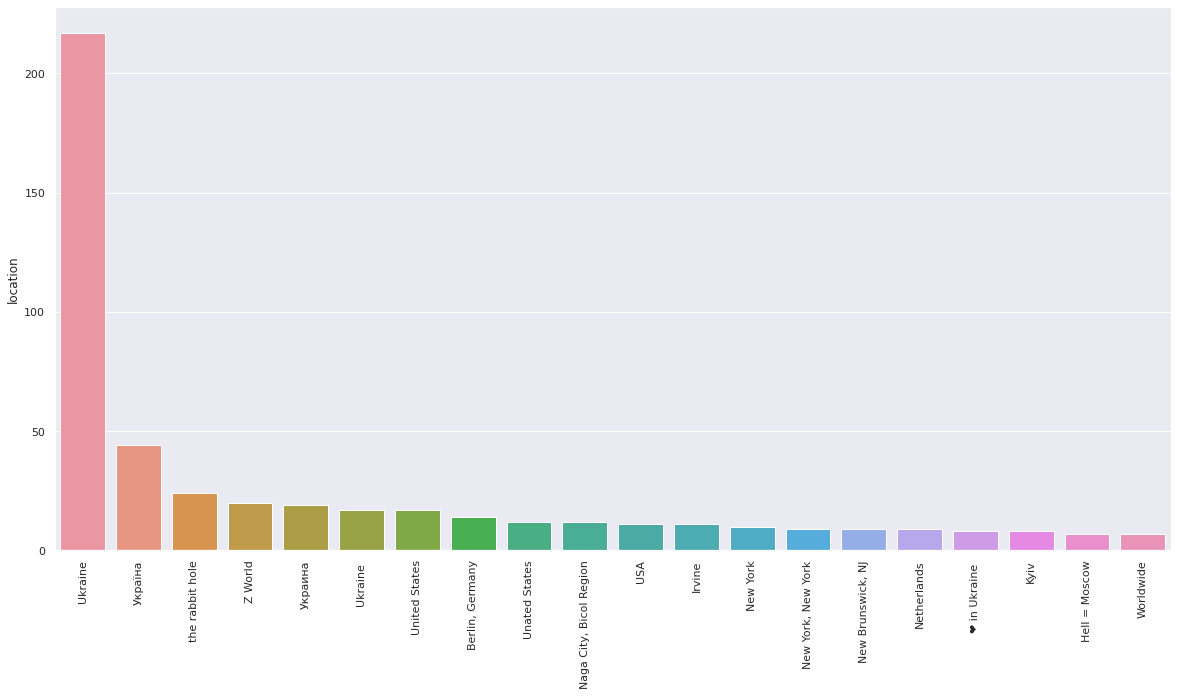


Figure 11: Word Cloud of unique tweets in dataset.

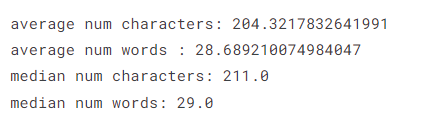
* **Bar Plot on User Account Age and classified based on the location of the account**
  + ***Top 1000 sorted values***



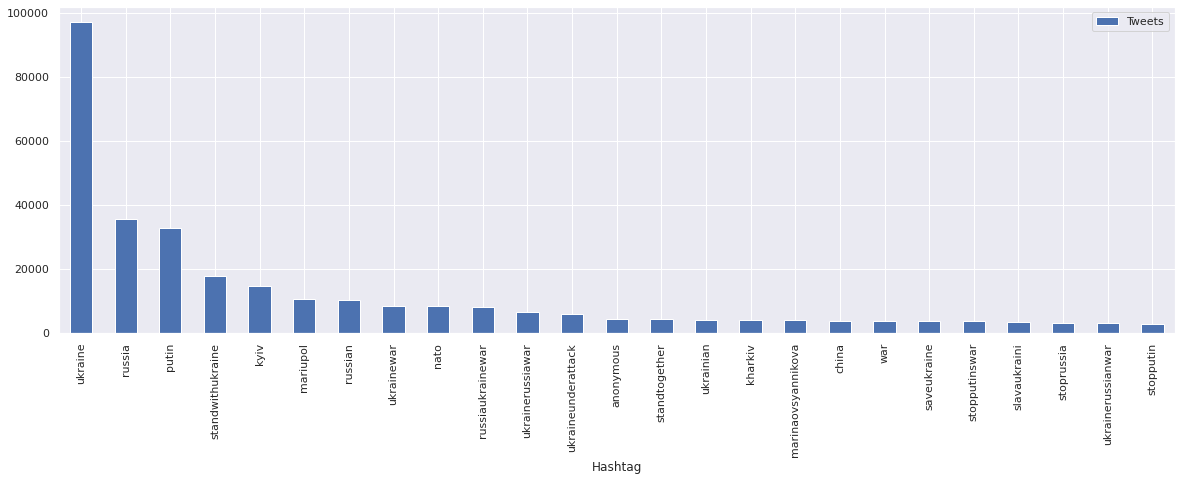
* + ***Bottom 1000 sorted values***



* **Hashtag Analysis**







## **Data Pre-Processing**

Pre-processing of the text data is an essential step because it prepares the raw text for finding information from it, in other words, it facilitates the extraction of information from the text so that machine learning algorithms can be applied to it. If we skip this first step, then it is more likely that the data you may be working with is noisy and inconsistent. During this process, the objective is to get rid of noise that doesn’t contribute very much to determining the sentiment of the tweet, including punctuation marks, emojis, special characters, numbers, and terms that don’t carry much weight when considered in context with the tweet itself.

Before performing data pre-processing, we have added a ‘sentiment’ column in the file. And we have set sentiments for each tweet as ‘positive’, ‘negative’ and ‘neutral’.

### **Plot for Sentiments**

We will first plot for sentiments to get visualization about sentiments’ distribution.

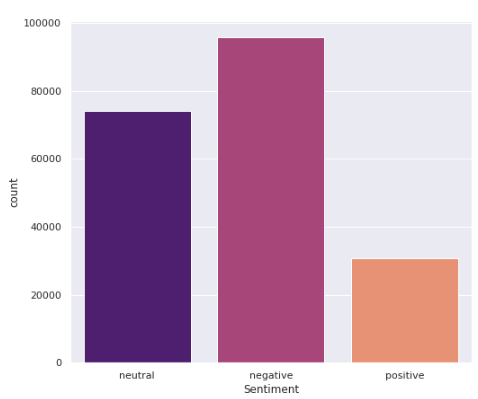


Figure 12: Graph showing distribution of sentiments in dataset.

The graph shows that negative sentiments are more in the tweets’ dataset on the selected topic.

### **Word Cloud for Negative Sentiments**



### **Word Cloud for Neutral Sentiments**



### **Word Cloud for Positive Sentiments**



We are using Python and we are using NTLK (Natural Language processing Toolkit) package and perform following pre-processing steps:

* **Converting tweets into lower case**

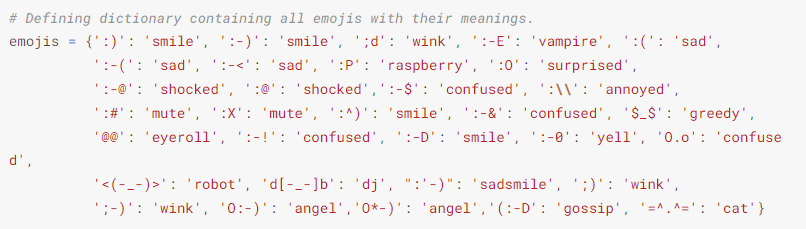
We have first converted tweets into lower case for make it easy to process.

* **Removing URLs**

All the URLs are replaced with URL so that it will not cause any effect on classification.

* **Removing emojis**

We have first defined dictionary in which we have explained meaning of each emoji. Then all emojis are replaced by their meaning.

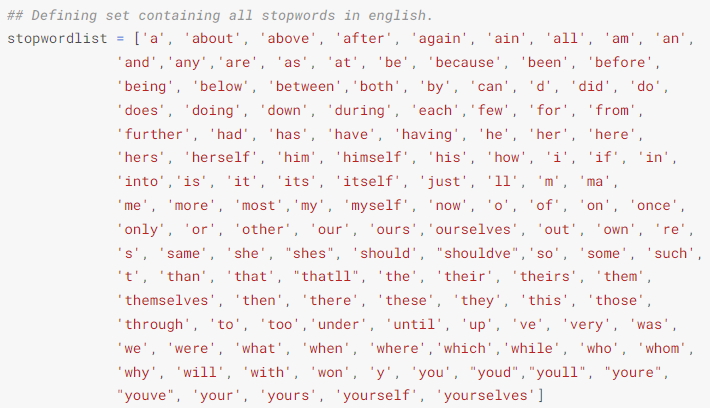


* **Removing non-alphabets**

We have replaced all non-alphabets.

* **Removing stop words**

A stop word is used to filter out unimportant words, so that the application can deal with the important words instead. We have defined a dictionary of stop words and then applied stop word removal technique.



* **Stemming**

Stemming is the process by which a word is reduced to its root word stem, the part of a word that attaches to prefixes and suffixes and to the roots of words, also referred to as lemma.

* **Lemmatization**

Lemmatization involves the use and analysis of a vocabulary, as well as morphological analyses of words, normally aiming to remove only those that are formed from inflectional elements, returning either the base form or the dictionary form of the word, which is called lemma.

After pre-processing steps, the data is split into training and testing data. then TF-IDF is performed and data is transformed before giving input to machine learning algorithms.

### **Calculate Term Frequency Inverse Document Frequency**

TF-IDF algorithm is very commonly used to transform textual input into meaningful representations of numbers, which are then used to fit a machine learning algorithm.

After performing TF-IDF. the machine learning algorithms are applied and results are observed.

## **Selected algorithms**

### **Bernoulli Naive Bayes**

The Bernoulli Naive Bayes is regarded as a version of Naive Bayes algorithm, which offers a lot of potential in the context of a binary distribution when the output label may be present or not[18].

The Bernoulli NB framework implements a naive Bayes training and classification approach based on a multivariate Bernoulli distribution; in other words, multiple features could be present but each of them would be assumed to represent a binary (Bernoulli) variable. Consequently, in order to use this class, sample data must be bound to binary-valued feature vectors; for any other data type, the Bernoulli NB instance will binarize it (in accordance with the binarize parameter).

The decision rule is based on Bernoulli's naive Bayes as follows:



The decision rule formula states that x must be binary in order for it to be applied

### **K Nearest Neighbor**

Generally, the K Nearest Neighbor algorithm can be categorised as Supervised Learning and is commonly used in classifications (most commonly) and regression analysis. The algorithm can also be used to determine whether missing values exist within a dataset and to resample a dataset if necessary. As the name suggests, K Nearest Neighbor is considering K the nearest neighbors (i.e., data points), in order to predict whether a new data point should belong to the class or continuous value [19].

### **Decision Tree**

Decision Trees are supervised learning techniques that can be used both for solving Classification problems and for solving Regression problems, although they are more commonly used for solving Classification problems. The classifier is a tree-structured one, where the nodes represent features, and the branches represent decision rules, and the leaves represent the outcomes in the classifier. Each node in the tree starts from the root node of the tree and dives up the instance space into two or more sub-spaces depending on whether it meets an attribute test condition or not. Upon moving down, the hierarchy tree branch that corresponds to the value of the attribute, we are able to create a new node. Similarly, the above process is repeated in regard to the subtrees rooted at the new node, in order to achieve classification of all of the records in the training set. When planning the construction of a decision tree, the process usually operates from the top-down, where each step of the decision tree is chosen based upon the attribute that will do the best job of dividing the data[20].

### **Random Forest**

The Random Forest algorithm is considered to be one of the most popular machine learning algorithms in the field of supervised learning. As a result it can be used for classification as well as regression based problems in Machine Learning. An ensemble learning algorithm is a technique that is based on the concept of multiple classifiers being used in conjunction with one another in order to solve a complex problem and to optimize performance.

The Random Forest, as its name implies, is a classifier that consists of a number of decision trees operating on different subsets of the data. Then the average is taken for the purpose of improving predicted accuracy of the data." The Random Forest does not rely on one decision tree but instead takes predictions from each tree and predicts the final results based on majority votes[21].

### **Logistic Regression**

The Logistic Regression algorithm is one of the most popular methods used by Machine Learning, a type of training method that is supervised from the outset. In this case, it is used to predict categorical independent variables using a given pair of independent variables.

Logistic regression is used to predict the output of a dependent variable that is categorical in nature. In other words, that variable must have a discrete or categorical outcome. In other words, it can either be Yes or No, 0 or 1 or True or False. In its place of giving the exact value which lies between 0 and 1, however, the classifier provides the probabilistic values, whose values lie between 0 and 1.

Logistic Regression is a form of regression that is somewhat similar to Linear Regression, except that their way of doing this is different. Regression problems are usually resolved by using linear regression, while classification problems are usually resolved by using logistic regression[22].

### **Support Vector Machine**

It is a supervised machine learning technique called Support Vector Machine (SVM) which is capable of being used for both classification as well as regression tasks. Nonetheless, it is most commonly used in classification tasks. When we use the SVM algorithm, every data item is plotted as a point in a n-dimensional space (n is the number of data items you have) with the value for every data item being based on the coordinates of that data item. We perform classification by identifying a hyperplane that gives a very clear differentiation between two classes.

Support vectors are nothing more than the coordinates of a single observation. An SVM classifier is a frontier that best separates the two classes from each other (hyperplane/line).

Several different hyperplanes are possible for separating the two types of data points and there are many possibilities. As a goal, we are interested in finding a plane with the largest margin, or in other words, the maximum distance between the points of the two different classes of data. This method provides some reinforcement by increasing the margin distance, by which future data points will be classified with a greater degree of confidence[23].

## **Evaluation Parameters**

The performance of algorithms is evaluated using following metrics:

## **Precision**

An observation's precision is determined by the ratio of correctly predicted positives to all predicted positives. Precision is calculated as:

## **Recall**

Recall is defined as the proportion of observed positive events to all those that occurred in the actual class - yes.  Recall is calculated as:

## **F1 Score**

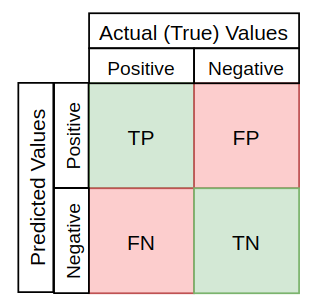
The F1 Score is computed by averaging Precision and Recall. It is calculated as:

## **Accuracy**

The accuracy of a prediction is the most intuitive measure of a system's performance since it is simply a percentage of correctly predicted observations divided by the total number of observations. Accuracy is calculated as:

## **Confusion Matrix**

A confusion matrix is used to display the performance of a classifier based on the four values which are TP, FN, FP and TN as shown:



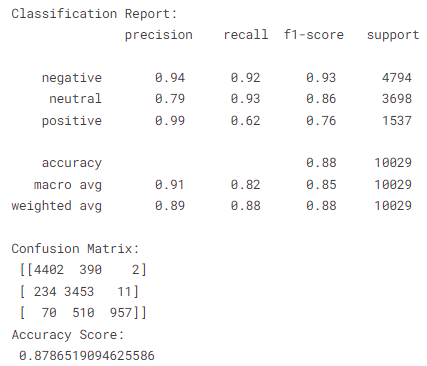
# **Results**

## **Sentiment Analysis Results**

### **Evaluation**

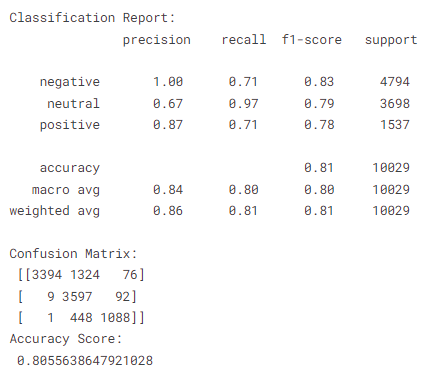
### **Bernoulli Naive Bayes**

The first algorithm applied is Bernoulli Naïve Bayes and following results are achieved.



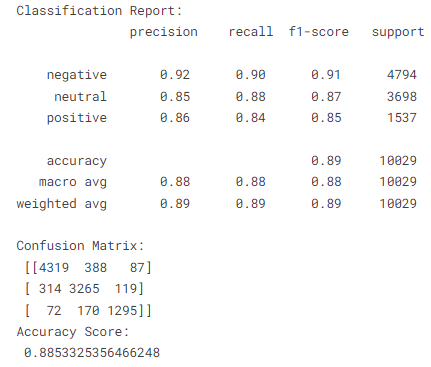
### **K Nearest Neighbor**

The results of KNN are shown below:



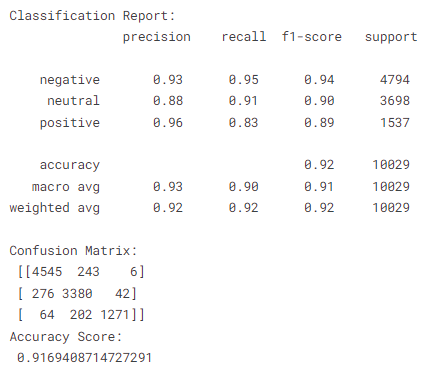
### **Decision Tree**

The following figure shows the results achieved after applying Decision Tree:



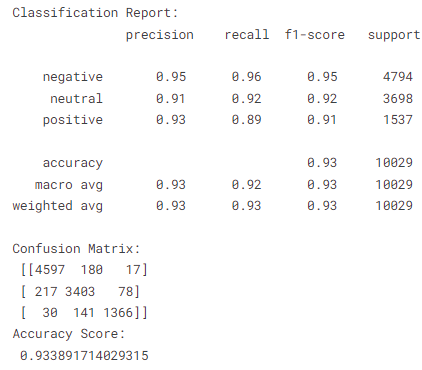
### **Random Forest**

The following figure shows the results achieved after applying Random Forest:



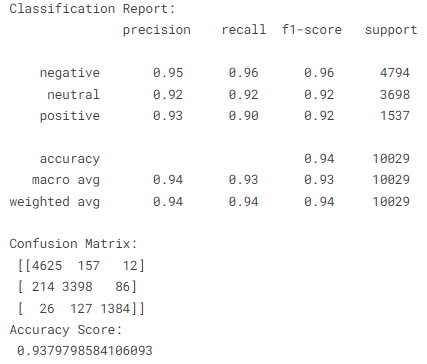
### **Logistic Regression**

The results of Logistic Regression are shown below:



### **Support Vector Machine**

The results of SVM are shown below:



### **Comparison of ML Algorithms for Sentiment Analysis**

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| ***Bernoulli Naïve Bayes*** | 87% |
| ***KNN*** | 80% |
| ***Decision Tree*** | 88% |
| ***Random Forest*** | 91% |
| ***Logistic Regression*** | 93% |
| ***SVM*** | 93.70% |

Table 1: Comparison of accuracy achieved by machine learning algorithms for sentimental analysis.

Figure 13: Graph showing accuracy achieved by machine learning algorithms for sentimental analysis.

From the results it can been seen that SVM performs best as compared to other algorithms. The accuracy of SVM is 93.70%. the logistic regression also performs good as its accuracy is 93%. On other hand, the accuracy of random is 91% that comes at third position. Afterwards decision tree performs well and gave accuracy of 88%. Then Bernoulli Naïve Byes gives us accuracy of 87%. KNN does not perform well and achieve lowest accuracy which is 80%.

## **Emotion Analysis Results**

We have added a ‘emotion’ column in the file. And we have set sentiments for each tweet as ‘anger’, ‘joy’, ‘optimism’ and ‘sadness’.

### **Plot for emojis**

We will first plot for emotions to get visualization about emotions’ distribution.

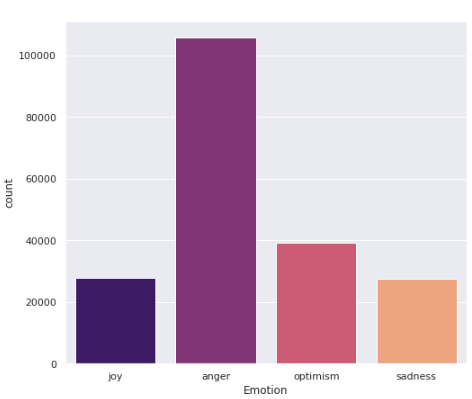


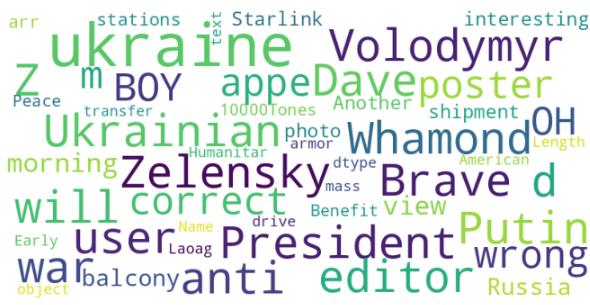
Figure 14: Graph showing distribution of emotions in dataset.

The graph shows that emotions of people in tweets on the selected topic are more towards anger.

### **Word cloud for anger**



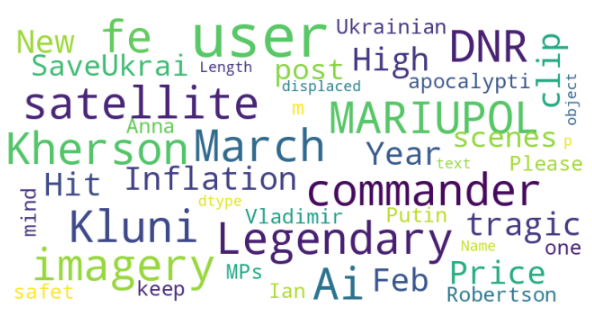
### **Word cloud for joy**



### **Word cloud for optimism**



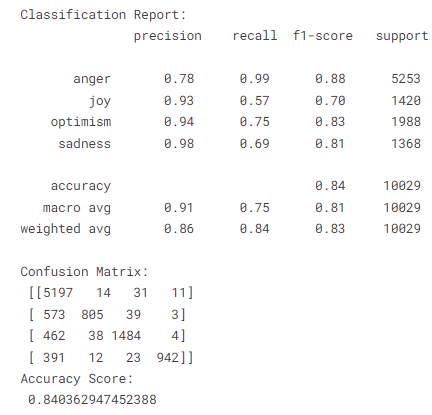
### **Word cloud for sadness**



### **Evaluation**

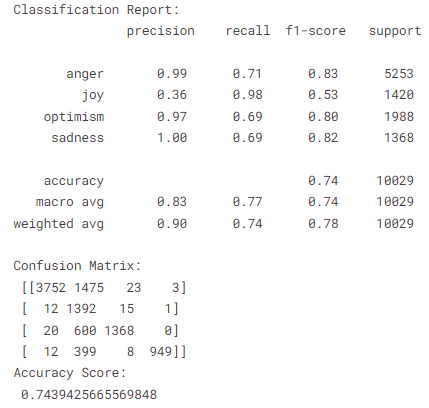
### **Bernoulli Naive Bayes**

The first algorithm applied is Bernoulli Naïve Bayes and following results are achieved.



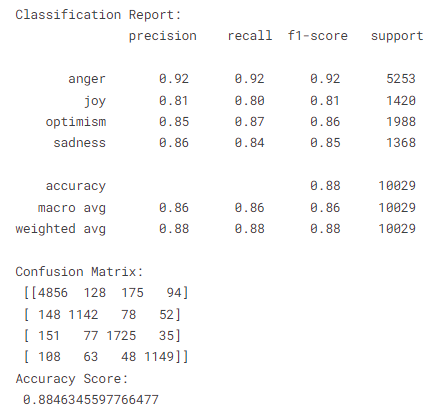
### **K Nearest Neighbor**

The results of KNN are shown below.



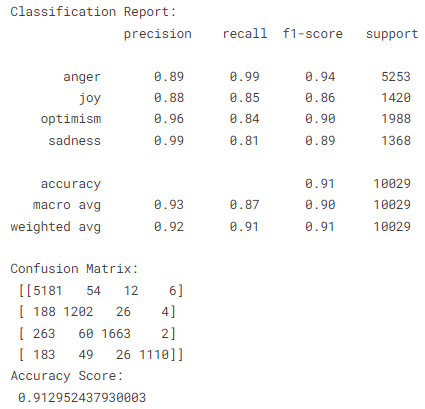
### **Decision Tree**

The following figure shows the results achieved after applying Decision Tree:



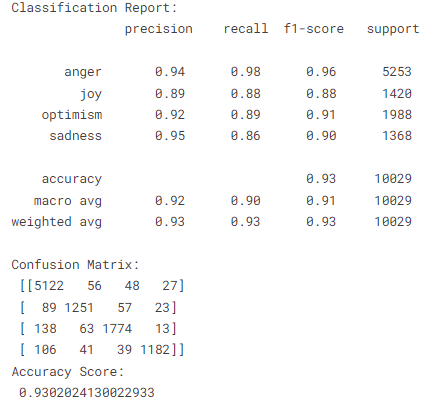
### **Random Forest**

The following figure shows the results achieved after applying Random Forest:



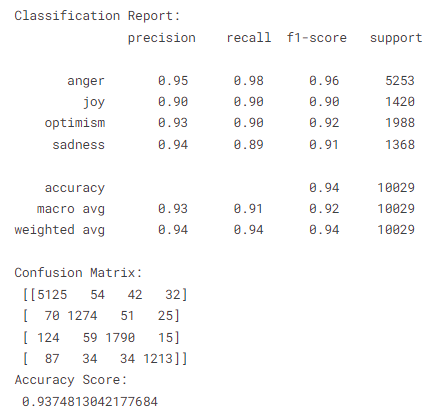
### **Logistic Regression**

The following figure shows the results achieved after applying Logistic Regression.



### **Support Vector Machine**

The following figure shows the results achieved after applying SVM:



### **Comparison of ML Algorithms for Emotion Analysis**

|  |  |
| --- | --- |
| **Algorithms** | **Accuracy** |
| ***Bernoulli Naïve Bayes*** | 84% |
| ***KNN*** | 74% |
| ***Decision Tree*** | 88% |
| ***Random Forest*** | 91% |
| ***Logistic Regression*** | 93% |
| ***SVM*** | 93.70% |

Table 2: Comparison of accuracy achieved by machine learning algorithms for emotions’ analysis.

Figure 15: Graph showing the accuracy achieved by machine learning algorithms for emotions’ analysis.

From the results it can been seen that SVM performs best as compared to other algorithms. The accuracy of SVM is 93.70%. the logistic regression also performs good as its accuracy is 93%. On other hand, the accuracy of random is 91% that comes at third position. Afterwards decision tree performs well and gave accuracy of 88%. Then Bernoulli Naïve Byes gives us accuracy of 84%. KNN does not perform well and achieve lowest accuracy which is 74%.

# **Conclusion**

The usage of social media sites like Twitter, Facebook, and WordPress is a major source of information exchange in our society today. As a social network, Twitter primarily relies on the opinions of the public regarding a product, event, or topic and therefore contains large amounts of raw data. There is a lot of importance and difficulty in synthesising and analyzing this data due to the sheer size of the datasets. An analysis of sentiments is the method that is chosen for this data analysis in that it does not delve into all the tweeted, but rather refers to the sentiments connected with each tweet in terms of positive, negative, and neutral opinions.

An innovative method of using machine learning algorithms to perform sentiment analysis on the twitter data set has been found to be useful in this project. As a result of the implementation of the developed algorithms, noise can be removed or data can be filtered in the form of Natural Language Processing (NLP) techniques, as well as pre-processing to improve output. Moreover, in order to perform pre-processing or filtering of the textual Twitter data in order to remove noise from it, the process of pre-processing must be performed in a series of steps. This process involves filtering and processing of tweets, in order to reduce the size of dataset. In the first step, dataset was explored to view the distribution of data. For visualization purpose, word cloud was used which helps to show the frequency of prominent words in dataset. Then tweets were converted into lower case, then URLs, emojis are replaced. afterwards stop words were removed and stemming along with lemmatization was performed.

After pre-processing steps, term frequency was calculated and data was transformed. Then dataset was split into training and testing data. Then data was put as input to classifiers. The selected classifiers were Bernoulli NB, KNN, Logistic Regression, Decision Tree, Random Forest and SVM. Each classifier was trained and tested. then results were evaluated using various evaluation metrics. First, we have performed sentiment analysis using selected classifiers. Then we have performed emotion analysis using selected classifiers. The results showed that SVM classifier outperformed as compared to other algorithms.

# **References**

[1] J. Carvalho and A. Plastino, “On the evaluation and combination of state-of-the-art features in twitter sentiment analysis,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1887–1936, 2021.

[2] W. Y. Chong, B. Selvaretnam, and L.-K. Soon, “Natural language processing for sentiment analysis: an exploratory analysis on tweets,” in *2014 4th international conference on artificial intelligence with applications in engineering and technology*, 2014, pp. 212–217.

[3] J. F. Raisa, M. Ulfat, A. Al Mueed, and S. M. S. Reza, “A review on Twitter sentiment analysis approaches,” in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 375–379.

[4] R. Feldman, “Techniques and applications for sentiment analysis,” *Commun. ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[5] A. Qazi, R. G. Raj, G. Hardaker, and C. Standing, “A systematic literature review on opinion types and sentiment analysis techniques: Tasks and challenges,” *Internet Res.*, 2017.

[6] M. Birjali, M. Kasri, and A. Beni-Hssane, “A comprehensive survey on sentiment analysis: Approaches, challenges and trends,” *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021.

[7] H. Wang, L. Liu, W. Song, and J. Lu, “Feature-based Sentiment Analysis Approach for Product Reviews.,” *J. Softw.*, vol. 9, no. 2, pp. 274–279, 2014.

[8] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-based methods for sentiment analysis,” *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, 2011.

[9] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, “Machine learning-based sentiment analysis for twitter accounts,” *Math. Comput. Appl.*, vol. 23, no. 1, p. 11, 2018.

[10] J. Jain, P. Panchal, N. Suryawanshi, and A. Shinde, “Sentiment analysis using supervised machine learning,” *Imp. J. Interdiscip. Res.*, vol. 2, no. 6, pp. 1362–2454, 2016.

[11] M. S. Usha and M. I. Devi, “Analysis of sentiments using unsupervised learning techniques,” in *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, 2013, pp. 241–245.

[12] L. Mandloi and R. Patel, “Twitter sentiments analysis using machine learninig methods,” in *2020 International Conference for Emerging Technology (INCET)*, 2020, pp. 1–5.

[13] S. Elbagir and J. Yang, “Sentiment analysis of twitter data using machine learning techniques and scikit-learn,” in *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 2018, pp. 1–5.

[14] D. R. Marples, “Ukraine in Conflict,” *An Anal. Chronicle, London E-International Relations*, 2017.

[15] I. Sabatovych, “Do social media create revolutions? Using Twitter sentiment analysis for predicting the Maidan Revolution in Ukraine,” *Glob. Media Commun.*, vol. 15, no. 3, pp. 275–283, 2019.

[16] “Ukraine Conflict Twitter Dataset (27.98M tweets) | Kaggle.” https://www.kaggle.com/datasets/bwandowando/ukraine-russian-crisis-twitter-dataset-1-2-m-rows (accessed May 06, 2022).

[17] “Best Free Word Visualization Tools.” https://monkeylearn.com/blog/word-visualization/ (accessed May 06, 2022).

[18] “sklearn.naive\_bayes.BernoulliNB — scikit-learn 1.0.2 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.naive\_bayes.BernoulliNB.html (accessed May 06, 2022).

[19] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, “KNN model-based approach in classification,” in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2003, pp. 986–996.

[20] S. R. Safavian and D. Landgrebe, “A survey of decision tree classifier methodology,” *IEEE Trans. Syst. Man. Cybern.*, vol. 21, no. 3, pp. 660–674, 1991.

[21] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?,” in *International workshop on machine learning and data mining in pattern recognition*, 2012, pp. 154–168.

[22] R. E. Wright, “Logistic regression.,” 1995.

[23] S. V. M. Vishwanathan and M. N. Murty, “SSVM: a simple SVM algorithm,” in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (Cat. No. 02CH37290)*, 2002, vol. 3, pp. 2393–2398.