** Purdue University Northwest**

**ITS52000**

**Web Analytics and Applications**

**Fall 2018**

**Final Research Paper**

**SENTIMENT ANALYSIS WITH TWITTER DATA**

***Submitted by***

**KALYAN PENDYALA**

**ALEKHYA JAGARLAMUDI**

**BHARATH KUMAR**

**Date: 09/19/2018**

**Abstract:**

Now a day’s twitter is considered as a source of huge data. People post their feelings in the form of tweets in twitter. These tweets talk about different things like politics, sports, technology etc. Among those tweets there are good tweets and bad tweets. We consider good tweets as tweets that talk about love, integrity and honesty etc., when it comes to bad tweets people talk about racism, threatening etc., so there comes the usage of sentiment analysis where we classify these tweets into two groups as hate speech and formal speech. This paper shows the tools and techniques that are used to classify the racist tweets from normal tweets.

**Background:**

**Natural language processing** (**NLP**) is a field of computer science, artificial intelligence and computational linguistics concerned with the interactions between computers and human (**natural**) languages, and, in particular, concerned with programming computers to fruitfully process large natural language corpora. NLP is basically inspired by the Neural Networks and how human brain works. From advances in Artificial Neural Networks and Machine Learning, NLP applications are also advanced.

Man ---Woman

King ---?

In the above shown example, we will know the answer as Queen as we got to know through our experiences. Likewise, we will enable the machines know the same by providing our experiences as the same provided above. These can be made understood by the machines through vector. Vector basically contains dimensions where it predicts the values using those dimensions.

**Machine learning** is basically the concept of making the machines learn the things and to make decisions from the data given as input and process the new data by its own going forward. like humans from the past, we learned from our ancestors how they survived, and we got to know the process of solving and making decisions from experience, machines are also made to learn and take decisions on their own like Siri i.e., understanding from the speech. It is designed as self-driven decision making than like data given explicitly by humans.

**Types of Machine Learning**

Machine Learning is sub divided into three types: Supervised, Unsupervised and Reinforcement. Let us see each of them in detail.

**Supervised Learning**

In this type, we provide the particular set of data to get desired outputs. Take the example of teacher in a class. The teacher provides the data to the students and the students just listen to what the teacher is explaining as they don’t have the idea at the first. It can be shown as y=f(x). Here, y is the output produced when the new data f(x) is processed. For each input, a output is evaluated. Once algorithm is trained, the output is predicted.

USECASES:

* + - Siri 🡪 Works with voice (Ex: Call Alex)
    - Weather forecast 🡪 Contains particular parameter (Cloudy, Sunny…)
    - Biometric Analysis 🡪 Thumbs, Retina etc.,
    - Banking 🡪 Credit worthiness of the card holder.
    - Health Care 🡪 Re-admission rates, Patients’ treatment.
    - Retail 🡪 Products that customer buy together.

**Unsupervised**

It contains only half input data. It works by modelling the distributed and underlying data in order to learn more about the data. For example, friends in a party (They keep on coming and it keeps on classifying based on color, age, class….). Here, the output may or may not be as expected data. It uses the concept of clustering where similar types of data are clustered and then processed. It only differentiates the similarities but doesn’t know the meaning of the data. This type of ML is so called because there will be no correct outputs as expected. It can be explained as the student-teacher relation i.e., When students start understanding by comparing things by their own that may be through books or through internet or any other resources.

USECASES:

* Banking 🡪 Behavioral customer classification by clustering
* Health care 🡪 Through normal and abnormal images
* Retail 🡪 Products to customers from past purchases

**Reinforcement:**

This allows the machines to automatically determine the ideal behavior within a specific content. It works on the trial and error basis i.e., if the correct output is obtained, the machine learns from it. Its wrong output s obtained, the machine gets penalty. Then it updates its policy and next time it makes its decision.

Ex: Training a dog with bell and food.

USECASES:

* Banking 🡪 To create next best offer using predictive model
* Health care 🡪 Depending on the strategies to handle different cases
* Retail 🡪 To reduce excess stock by dynamic pricing

ML is a type of AI that provides computer with the ability to learn without being explicitly programmed. It focusses on development of computer program that can change when a new data is given. Most of the people are in a thought that ML (Machine Learning), AI(Artificial Intelligence) and DL(Deep Learning) are the same. But that’s not the fact. Let’s see the difference.

* AI is to make the machine carry out the task in better way as humans.
* ML is the subset of AI. It is about giving access of data to machines and to make them take decisions on its own i.e., it not only takes decisions but also adopt to changes that occur.
* DL is the subset of ML. It is to perform tasks and get better accuracy when both the former isn’t performing up to the mark.

**How to solve a problem using ML:**

ML contains a sort of algorithms to solve a situation. Algorithm is basically the logic of the writing a program. To tell a computer what to do, we need a program. Program is nothing but a logic in some language syntax where logic itself is the algorithm.

Let’s see how algorithm works in a flow chart.

Start 🡪 Init x=0 🡪 x++(increment of x) 🡪 print x 🡪 x<<0 🡪 END

If the value at x<<0 is Y, then the process is repeated by increasing the value of x by 1. If not, the process is ended, and the result is printed. This is just an example of printing he numbers from 1-20**.**

**Sentiment Analysis**

Sentiment Analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

There are mainly **three** approaches for performing the sentiment analysis. They are

1. Lexicon based - considers lexicon dictionary for identifying polarity of the text**(Purely NLP based)**
2. Machine learning based approach - Needs to develop classification model, which is trained using prelabeled dataset of positive, negative, neutral content.**(Combines concept of NLP and ML)**
3. Combined approach - Which uses lexicon dictionary along with pre-labelled data set for developing classification model. (**Combines approach 1 and approach 2)**

**Project Introduction:**

From the past few years, there has been a huge growth in the use of twitter and thereby analyzing the opinions of the people for information about their products and services. While there has been a considerable lot of research on how sentiments are communicated in online audits, news articles, highlights etc., Automatic part-of-speech tags and assets, sentiment lexicons have demonstrated helpful for sentimental analysis in different areas.

**Project Objective:**

The primary objective of this project is to classify the tweets that are extracted from twitter into hate speech and the formal speech. For the sake of simplicity, we say a tweet has hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets. In detail, the tweets can be analyzed by numbering them as ‘1’if the tweet is of racist/sexist and ‘0’ if the tweet is a non-racist/non-sexist one.

**Python Libraries used:**

**NumPy** is the fundamental package for scientific computing with python. It contains very useful things like N-dimensional array object, sophisticated functions and Linear algebra capabilities. It can also be used as an efficient multi-dimensional container of generic data; arbitrary data types can be defined.

**Scikit-learn** is a free machine learning library for python. It features various classification, regression, and clustering algorithms including SVM, Random Forest, Gradient Boosting, K-means etc., and designed to interoperate with NumPy and SciPy.

**Matplotlib** is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.

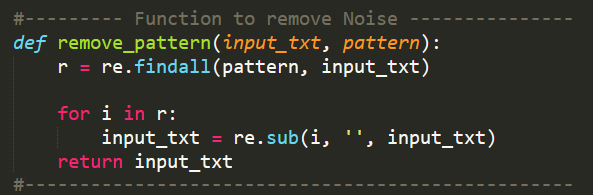
**Regular Expression** is a library that is used to clean the noise in the data. It is used at data pre-processing stage where we remove punctuations, short words and hashtags from the data.

**Tools used – Anaconda:**

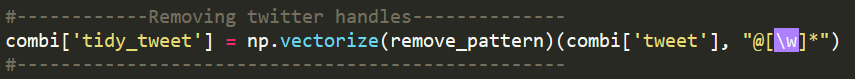
Anaconda is a free and opensource distribution for Python and R programming languages for data science and machine learning applications.

**Data Preprocessing:**

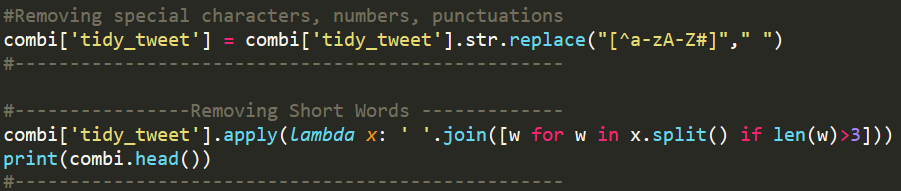
**Data preprocessing** is a**data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues. In Real world **data** are generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate **data.** Noisy: containing errors or outliers. Inconsistent: containing discrepancies in codes or names.



In the below code we are removing twitter handles. Twitter handles are usernames. Since we don’t need the usernames for the sentiment analysis, we are removing these handles.

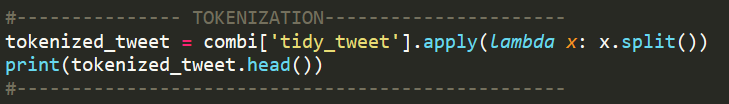


In the below tweets we are removing special characters, numbers and punctations. If we leave them as it is and submit the data to the algorithms. Algorithms consider each of these punctuations, numbers and special characters as vectors and produce false results. Our main aim of the prediction will be spoiled with these characters. We are removing them with the help of a library called Regular Expressions.



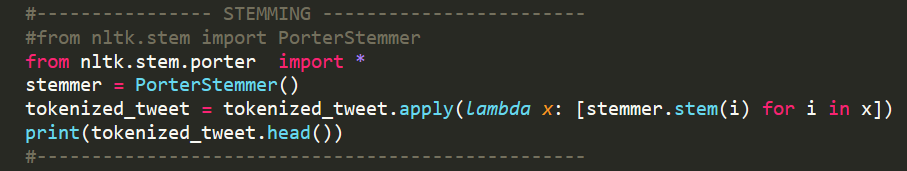
**Tokenization**

**Tokenization** is the process of dividing text into a set of meaningful pieces. These pieces are called **tokens**. For example, we can divide a chunk of text into words, or we can divide it into sentences. Depending on the task at hand, we can define our own conditions to divide the input text into meaningful tokens.



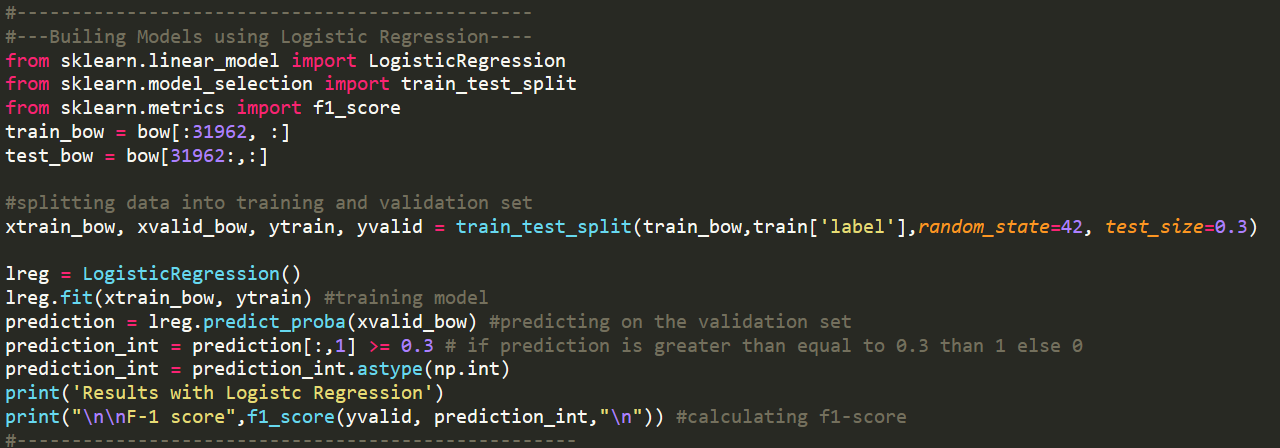
**Stemming**

[Stemming](http://en.wikipedia.org/wiki/Stemming) is the process of reducing a word into its stem, i.e. its root form. The root form is not necessarily a word by itself, but it can be used to generate words by concatenating the right suffix.

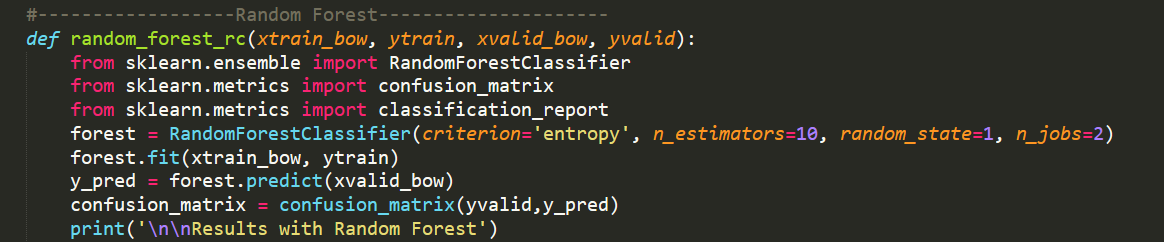


**Algorithms used:** (Logistic Regression, Random Forest, Decision Tree)

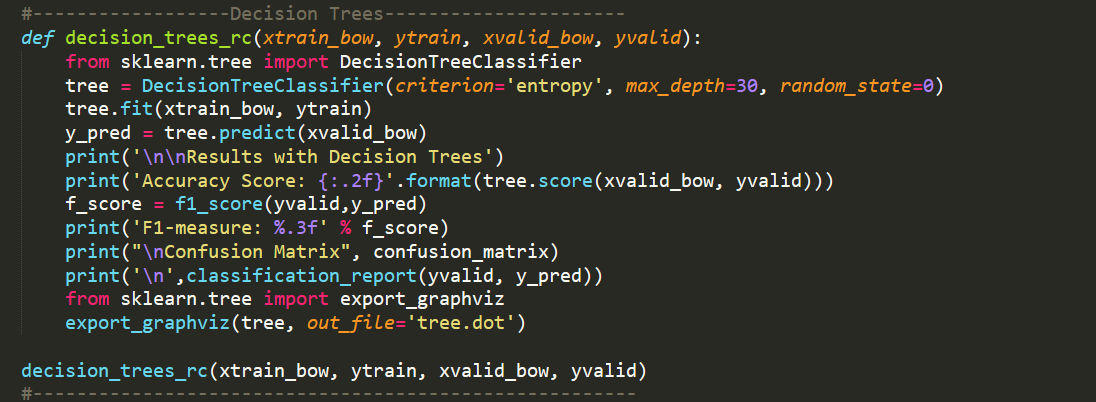
**Logistic Regression** is a statistical method to analyze a dataset in which there are one or more independent variables that determine an output. The output is measured with a binary variable in which there are only two possible outcomes. Like all the regression analyses, the logistic regression is a predictive analysis.  It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.



**Random Forest** algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result.

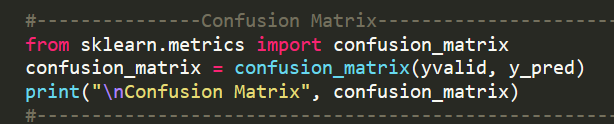


**Decision tree** has many analogies in real life and it turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though it is a commonly used tool in data mining for deriving a strategy to reach a particular goal, it’s also widely used in machine learning.



**Performance Metrics:**

**Confusion Matrix** is one of the easiest metrics used to find the correctness and accuracy of the model. It is used in the Classification problem where the output is two or more classes. All the other performance metrics are based on the confusion matrix.



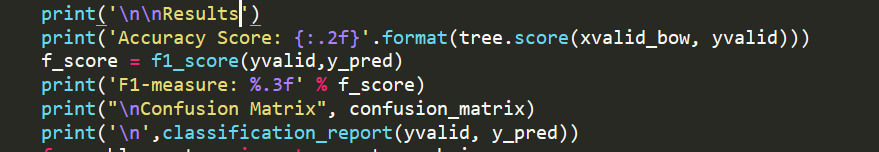
**Accuracy** is number of correct predictions made by the model over all kinds of predictions made.

**Accuracy** = True Positive + True Negative / True positive + False positive + False Negative + True negative

**Precision** = True Positive / True Positives + False Positives

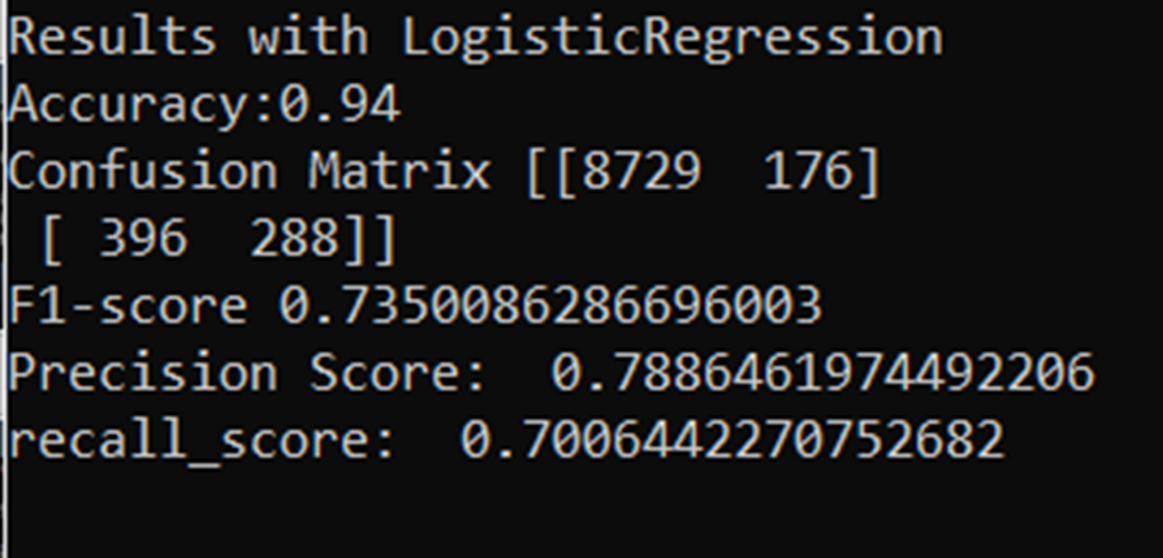
**Recall** = Number of True positives / Actual Number of Positives

**F1 Score = 2** x (Precision \* recall) / Precision + Recall

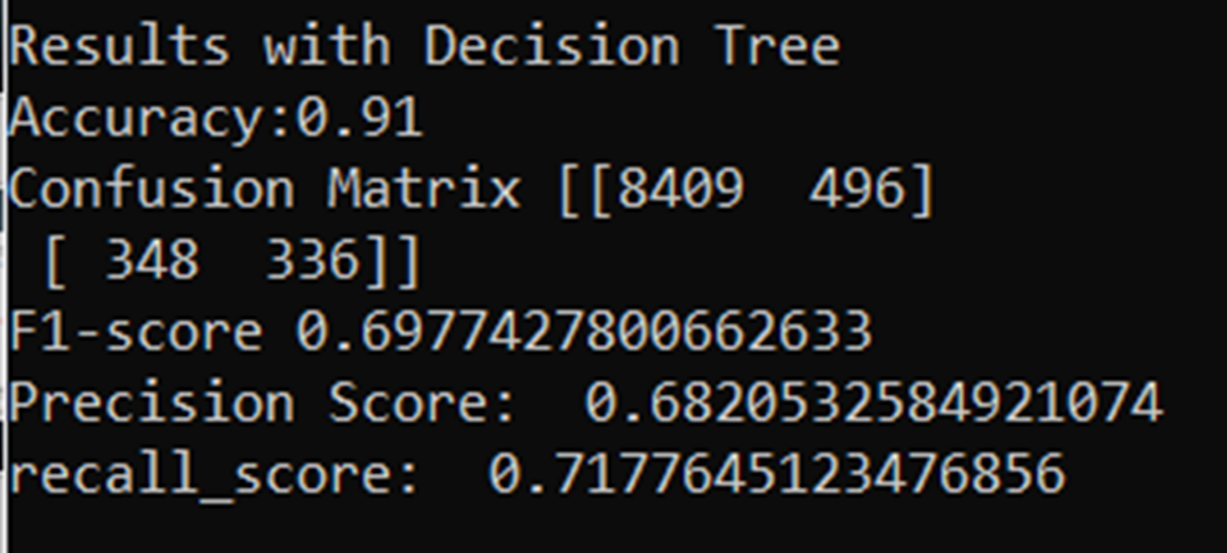


**Results:**

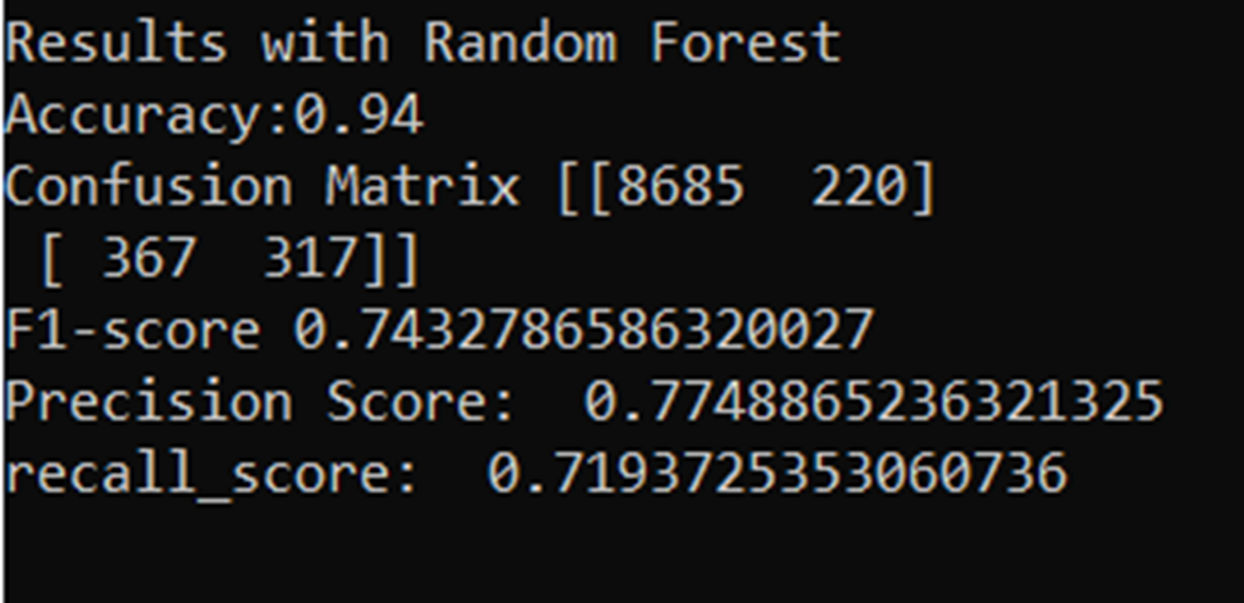
The following are the results with logistic regression algorithm. After testing this algorithm, we achieved the accuracy of 95% and F1 score of 55%. In recall we got 95%.



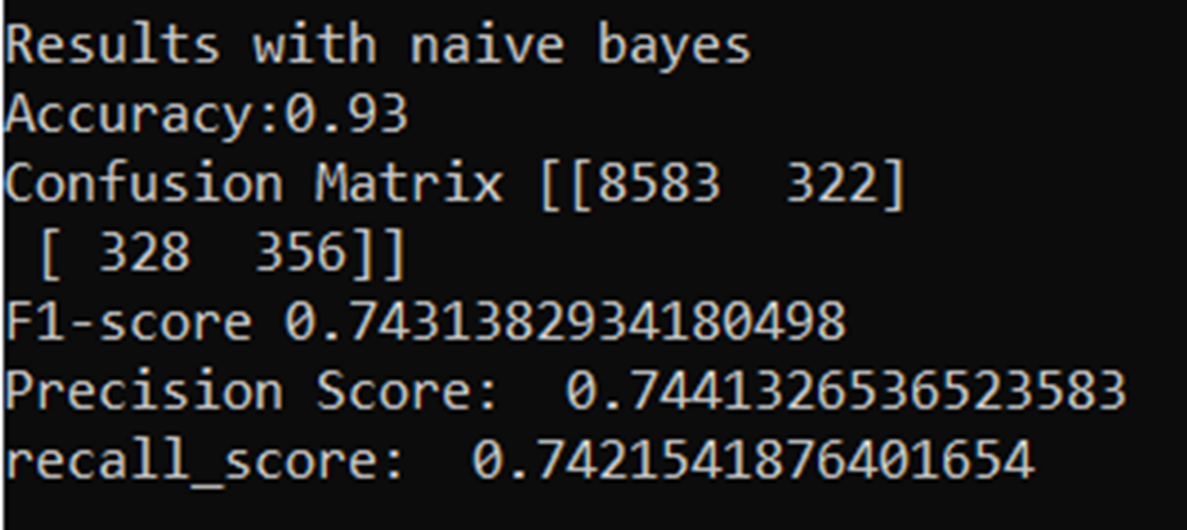
The next results belong to Decision Trees algorithm one of the famous algorithms in machine learning. With decision trees we achieved less accuracy when compared to the Logistic Regression. Here we got 94% accuracy and 45% of F1 measure.



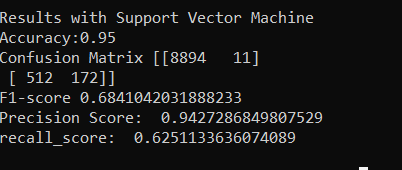
Our final results are with the Random Forest Algorithm. In this algorithm we got 94% accuracy which is same as Decision Trees and F1 measure is little bit high when compared to the Decision Trees here we got 51%.



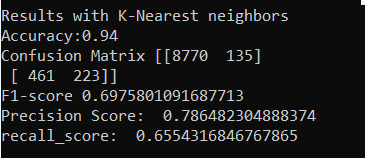
Naïve Bayes



Support Vector Machines



K-Nearest Neighbors



**Conclusion:**

This research and project are mainly focused on detecting the hate speech from the twitter tweets. We built models with some of the famous machine learning algorithms like logistic regression and Decision Trees. We also achieved good performance metrics. This model can be applied to the real-world data. It works efficiently to detect the racist tweets in the twitter.

**References:**

1. T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Ruckstieß, and J. Schmidhuber. ¨ PyBrain. The Journal of Machine Learning Research, 11:743–746, 2010.
2. T. Zito, N. Wilbert, L. Wiskott, and P. Berkes. Modular toolkit for data processing (MDP): A Python data processing framework. Frontiers in Neuroinformatics, 2, 2008.
3. B. Pang, L. Lee, S. Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques, 2002.
4. B. Pang and L. Lee. "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts" in Proceedings of ACL, 2004.