# Determining the Most Popular Streaming Service using Machine Learning

**Sayan Ghosh, Dipshikha Sarkar, Lokenath Basu, S.R. Rajeswari**

***Abstract— Over the past years, twitter has become a popular medium for sharing views and ideas about personalities, brands, products or services. Analyzing sentiment of people to figure out the popularity of different streaming service by the twitter profiles is helpful for determining positive or negative views. This is a comparative analysis to predict or show which of the chosen streaming services is most familiar or liked by the public. To do this, different machine learning algorithms are used to computationally identify and categorize public opinions to draw a final result. The machine learning algorithms used here are Linear SVC, Naïve Bayes and Decision Tree. These help in receiving the data and predict the output within an acceptable range. The data in this case has been extracted from Twitter using Twitter API. Twitter API takes the parameters that can access many features of Twitter and also post and find tweets containing desired words. This includes data cleaning which refers to exclude the incorrect and unnecessary forms of data. This makes the way of data processing easier, faster and more compatible. On analyzing, the frequently used words are assessed. The classifying words are trained using the above mentioned algorithms. These algorithms are the supervised classifiers which are effective and efficient when the quantity of the data is huge. Using one or more algorithms helps to decide, compare and contrast the results. Once the classifiers are trained, testing is done. Testing gives the proper assessment of the data that is required for the desired results. The performance of the test set can be checked to draw a final result. Hence, comparing the results obtained for different streaming services helps to decide the most popular streaming service.***

***Keywords—Linear SVC, Naïve Bayes, Decision Tree, Twitter, Twitter API***

## I. INTRODUCTION

Social media has been a quality platform to share views, ideas and feedbacks in recent days. This can be used to improve personalities, services, products and brands. Comments and compliments are regarded as a valuable part of any utility. On analyzing public reviews, the popularity and goodness of any utility can be judged.

In recent days, online streaming apps and websites are taking over the television industry by storm. More and more people prefer these streaming services over television because it has quality content, exclusive shows, and movies and faster premiers. So, with so many streaming services in the country, it boils down to select the best ones.

The main objective of the project is to analyse the most popular streaming service using sentiment analysis of twitter data. This gives an idea of a popular streaming service that is preferred by the viewers. It helps to narrow down choosing the best streaming service out of many available in the market.

## II. RELATED WORK

Researches in [2] mainly focus on analyzing the tweets collected from Twitter written in English that are unstructured along with emoticons. Naïve Bayes method is chosen to build the classifiers that are used in their research. Works in [3] concentrate on excluding the lowercases, hashtags, usernames and URLs. Also, they have used decision tree to represent and segment out the choices in a tree by branching it. The authors in [4] preprocessed their datasets by removing noises, duplicate tweets, punctuations, stop words, numbers, etc. They further worked on changing the upper cases to lower cases. They represented each tweets by a vector so that it gets easily interpreted by the classifier. For this purpose, methods like decision tree and Naïve Bayes were used. In [5] data were fetched from Twitter for one month amounting 59,988. Various stop words and URLs were removed at the preprocessing stage. Naïve Bayes was used as one of the algorithm to run on the collected data. Since the tweets were in the multilingual form they converted them in English and excluded the stop words, URLs, tags, non-alphabetical characters for improving the data collected. Stop words mainly refer to the articles and the helping verbs used in the tweets.

## III. METHODOLOGY

The main focus here is to determine the most popular streaming service among Netflix, Hotstar and Amazon Prime Video. The tweets of the people regarding each streaming service in terms of how good or bad the services are taken as base on which the analysis is done. Here tweets are extracted from twitter using python. Python is a high level interpreted language which is used for machine learning algorithms. The tweets were extracted using the tweepy package. The necessary packages and libraries were installed.

**Table I Number Of Tweets Extracted**

| NETFLIX | HOTSTAR | AMAZON PRIME VIDEO |
|---|---|---|
| 7500 | 7500 | 7500 |

Revised Manuscript Received on December 15, 2019.

**Sayan Ghosh**, Student of computer science engineering in SRM Institute of Science and Technology Ramapuram, Chennai. Email:sayandnsm@gmail.com

**Dipshikha Sarkar**, Student of computer science engineering in SRM Institute of Science and Technology Ramapuram, Chennai. Email: sarkardipshikha@gmail.com

**Lokenath Basu**, Student of computer science engineering in SRM Institute of Science and Technology Ramapuram, Chennai. Email: lokenath.bose71@gmail.com

**Ms. S. R. Rajeswari**, Assistant professor, SRM Institute of Science and Technology Ramapuram, Chennai. She has completed her M.E. in CSE at Anna University. Email: sr.rajicse@gmail.com

To get more information about the tweets collected the word clouds were created using wordcloud package. This gives a better information about the most frequently used words.

### *A. Preprocessing*

The tweets extracted are highly dimensioned unstructured data and hence it has to be cleaned before analyzing. First the tweets collected which are stored in a list is converted into a csv file. Then the unnecessary parts are removed like URLs, usernames, accounts, punctuations and stopwords.

Once cleaning is done the tweets are saved in another csv file. The analysis is done on this new csv file containing cleaned tweets.

**Table II Tweets Before And After Cleaning**

| ***Tweets before Cleaning*** |
|---|
| "@anonymityshine @ShazieZea I started watching The Bros (Movie) this morning on Netflix. So far it's fun." |
| ***Tweets after Cleaning*** |
| "I started watching The Bros (Movie) this morning on Netflix. So far it's fun." |

### *B. Model Building*

After cleaning the tweets, two csv files are downloaded which contains pre-classified positive and negative words. These files are used to compare each words with the tweets to classify whether a tweet contains more positive or negative words.

Next three supervised algorithms are applied for training the system: Linear SVC, Naïve Bayes, and Decision Tree.

- Linear SVC: The objective of a Linear SVC (Support Vector Classifier) is to fit to the data provided, returning a "best fit" hyperplane that divides, or categorizes the data.
- Naïve Bayes: is defined as classifier used to determine the most probable class label for each object.
- Decision tree: are flexible algorithms used to assign□label based on the highest score. Random forest: is a supervised algorithm for constructing multiple decision tree.

## IV. RESULTS AND DISCUSSION

The data extracted directly from Twitter API were used to train and test the models. The positive and negative csv files are used to determine the sentiment of every tweet. This model combined several algorithms to get the fit model for the used data.

The following tables shows us the accuracy of the algorithms used for the streaming services.

**Table III Accuracy (Netflix)**

| NETFLIX | |
|---|---|
| **ALGORITHM** | **ACCURACY** |
| LINEAR SVC | 0.84 |
| NAÏVE BAYES | 0.82 |
| DECISION TREE | 0.74 |

**Table IV Accuracy (Hotstar)**

| HOTSTAR | |
|---|---|
| **ALGORITHM** | **ACCURACY** |
| LINEAR SVC | 0.83 |
| NAÏVE BAYES | 0.80 |
| DECISION TREE | 0.72 |

**Table V Accuracy (Amazon Prime Video)**

| AMAZON PRIME | |
|---|---|
| **ALGORITHM** | **ACCURACY** |
| LINEAR SVC | 0.79 |
| NAÏVE BAYES | 0.77 |
| DECISION TREE | 0.71 |

## V. CONCLUSION

Opinion analysis or opinion mining is a field of study that analyzes people's sentiments, attitudes or emotions towards certain entities. The reviews given by the public for or against some streaming services has been handled here. It is an important issue in the recent days to talk about as most of the people are switching to online streaming services to catch up their favorite shows rather than waiting for them to be casted in the television. Analyzing public reviews is a better way to understand which streaming service is mostly liked by the audience and by what ratio. Reviews from Twitter has been collected and selected as the data to be used for the analysis purpose. This is a computational process to analyze and compare among different streaming services that has been taken in to account. Computational analysis gives much accurate results and it is done using several useful and efficient supervised algorithms.

As a result of the training and testing of tweets and algorithms used Netflix is the most popular streaming service.

The main idea is to create a system that will help to find out the preferred streaming service by the help of algorithmic comparisons based on the reviews given by the audience on the twitter platform. This kind of analysis has strong commercial interest because it is important of the companies often wants to know how much their product or service are being perceived. It can also be a sign to change or develop any existing feature or advantage to get an increased rate of demand for that product or service. Public can get an idea to decide what is best in the market along with a range of various other similar products or services to choose from. In the future, this comparison can be more precisely done by the usage of some other efficient algorithm better than the existing ones used here.

## ACKNOWLEDGMENT

We would like to show our gratitude to all participants involved in this work at SRMIST, Ramapuram..

## REFERENCES

1. SAHAR A EL_RAHMAN, Feddah Alhumaidi AlOtaibi and Wejdan Abdullah AlShehri's 'Sentiment Analysis of Twitter Data' in 2019 International Conference on Computer and Information Sciences (ICCIS).
2. M.TRUPTHI, SURESH PABBOJU and G.NARASIMHA's 'Sentiment Analysis on Twitter Using Streaming API' in 2017 IEEE 7th International Advance Computing Conference (IACC).
3. Megha Rathi , Aditya Malik, Daksh Varshney, Rachita Sharma and Sarthak Mendiratta's 'Sentiment Analysis of Tweets using Machine Learning approach' in Proceedings of 2018 Eleventh

International Conference on Contemporary Computing (IC3), 2-4 August, 2018, Noida, India

4. MOHAMMED H. ABD EL-JAWD, RANIA HODHOD and YASSER M. K. OMAR's 'Sentiment Analysis of Social Media Networks Using Machine Learning' in 2018 14th International Computer Engineering Conference (ICENCO).
5. PULKIT GARG, HIMANSHU GARG and VIRENDER RANGA's 'Sentiment Analysis of the Uri Terror Attack Using Twitter' in International Conference on Computing, Communication and Automation (ICCCA2017).
6. VICTORIA IKORO, MARIA SHARMINA, KHALEEL MALIK and RIZA BATISTA-NAVARRO's 'Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers' in 2018 Fifth International Conference on Social Networks Analysis, Management and Security(SNAMS).
7. PRAKRUTHI V, SINDHU D and DR S ANUPAMA KUMAR's 'Real Time Sentiment Analysis of Twitter Posts' in 3$^{RD}$ IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018..
8. NIHARIKA KUMAR's 'Sentiment Analysis of Twitter Messages: Demonetization a Use Case' in 2$^{ND}$ IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017.

## AUTHORS PROFILE

**Sayan Ghosh**, Presently he is a student of computer science engineering in SRM Institute of Science and Technology at Ramapuram, Chennai. His email is sayandnsm@gmail.com

**Dipshikha Sarkar**, Presently she is a student of computer science engineering in SRM Institute of Science and Technology at Ramapuram, Chennai. Her email is sarkardipshikha@gmail.com

**Lokenath Basu**, Presently he is a student of computer science engineering in SRM Institute of Science and Technology at Ramapuram, Chennai. His email is lokenath.bose71@gmail.com

**Ms. S. R. Rajeswari**, Presently she is an assistant professor in SRM Institute of Science and Technology at Ramapuram, Chennai. She has completed her M.E. in CSE at Anna University. Her email is sr.rajicse@gmail.com

*Retrieval Number: B2500129219/2019©BEIESP*
*DOI: 10.35940/ijeat.B2500.129219*