TWITTER Sentiment Analysis & Churn prediction

Mufaddal Kadibhai, Ankit Raj, Shreyash Shetty, Pruthvish Vyas,

University of South Florida

ISM 6190 Text Analytics Spring 2020

Professor Anol Bhattacharjee

Table of Contents

[**Executive Summary 1**](#_Toc39190635)

[**Problem Significance 2**](#_Toc39190636)

[**Prior Literature 2**](#_Toc39190637)

[**Data Source and Preparation 3**](#_Toc39190638)

[**Text Analytics Workflow 4**](#_Toc39190639)

[**Exploratory Data Analysis 5**](#_Toc39190640)

[**Representative words 6**](#_Toc39190641)

[**Churn Labeling 7**](#_Toc39190642)

[**Churn Detection 7**](#_Toc39190643)

[**Rule based Feature engineering 9**](#_Toc39190644)

[**Dashboard 10**](#_Toc39190645)

[**Key Actionable Insights 11**](#_Toc39190646)

[**References 12**](#_Toc39190647)

[**Appendix: 13**](#_Toc39190648)

# Executive Summary

We embark on previous experiments to test churn likeliness detection by using user remarks from twitter to identify characteristics of churners and then predict likelihood of churn and the changing sentiment over time. We seek to further develop a dashboard that shows the overall sentiment of the company and these features. Previous experiments have shown success in this task and we intend to confirm this and further test and develop a few models.

The availability and explosion in social media can play an instrumental role in the growth, development, and sustainability of companies. Thus, it is ever more important for companies to monitor these customer voice platforms and we explore Twitter in particular but keep in mind that the whole analysis needs to integrate data from a wide variety of social media sources such as Glassdoor, LinkedIn, and Facebook is left for future work. There is a corpus of constant data on Twitter and we think we can derive value from this analysis.

We choose to focus on the top 4 web streaming services (Amazon Prime Video, Hulu, Disney Plus, & Netflix). This is particularly due to current Covid-19 times causing increasing usage and engagement of these services.

Our studies will show the challenges in building machine learning models for twitter data analysis and extracting useful information. We conclude that an increased tweet count was attributed to higher positive sentiment and thus analyzing these time periods would be critical to get feedback. Especially being aware of competitor activity levels are helpful as an investigative starting point.

# Problem Significance

Churn prediction can help many companies identify factors that cause users to leave their service. Interpreting users feedback quickly will add positive momentum to product development and company growth which has significant positive ‘domino’ effects. The explosion of twitter usage has led to the availability of large streaming customer data which we can analyze in real time. We want to discover if these tweets can help inform a companies’ decision making with real insights. Although we analyze a small subset of tweets, we believe there are many companies and industries that could benefit by performing a similar analysis. Our analysis can also serve as a research source for investors looking to get real time competitive market insights into user remarks.

# Prior Literature

The literature in the field of opinion mining for sentiment mining is deep and evolving rapidly. Previous research has shown that valuable information can be derived from the twittershpere for sentiment purposes. That experiment was conducted with mobile carriers and their process and techniques have been used as a baseline for our research. We acknowledge the use of Naïve Bayes as a powerful model that has proved successful in text classification work. The use of deep learning model has also been recorded as a method building accurate classifiers. One study goes further by developing their own system using transfer knowledge and logic rules to outperform baseline models. There are other studies in which manual labeling using Amazon Mechanical Turk is performed. This research showed that the combination of unigrams and bigrams provide a good starting point for the model inputs. Further they show that demographic and content indicator information can be important factors in churn detection.

# Data Source and Preparation

We sourced 457,000 tweets with keywords of the top 4 web streaming platforms over the course of one month beginning March 23, 2020 from Twitter. After creating a developers account and getting access we used a python package called Tweepy to pull tweets. This was a long process requiring stoppage in between multiple data pulls.

The next step was to clean the data by removing stop words and punctuation. Then we parsed the tweets to also remove hashtags and handles. Once we had completed this step with cleaned tweets we could proceed to feature engineering using various methods as described in the text analytics workflow below:

# Text Analytics Workflow

Topic Modelling for positive and negative tweets

Trained model on labeled data and implemented classifier for predicting churn

Used representative vocabulary to determine key indicators for churn likeliness labeling

Extracted tweets for Prime Video, Netflix, Hulu and Disney Plus

Cleaned and preprocessed tweets

Grouped tweets and retweets for each carrier

Performed sentiment analysis

EDA: Tweets per platform, most frequent words, tweet count and polarity distribution over time, positive vs negative tweets

Rule based feature engineering to determine the better performing platform and customer remark

Aggregated the results and key findings into a dashboard and deployed it

# Exploratory Data Analysis

QUESTIONS

* How many numbers of tweets per service are there?
* How does the tweet count change over time, and does it affect the sentiment polarity?
* What are the most frequent words?
* Whether most of the tweets are positive or negative?
* What are the topics and churn indicators in the tweets?
* Which are the better performing platforms?
* What are the user complaints?

We can see that Netflix is the leading platform in terms of tweets followed by Hulu and Disney Plus a close second and third with PrimeVideo lagging the group. Interestingly we see a big spike in Disney for the first week that may be useful further study which brought its overall performance much higher. In terms of sentiment DisneyPlus seems to have the highest percentage of positive and lowest percent of negative tweets, with Prime, Hulu, and Netflix in that order. This means that although Netflix is currently holding the largest presence – this may be threatened by DisneyPlus surge were it to be a repeat threat as it did have the highest tweet count in a day for the period we studied. Thus, time is a critical factor and to properly understand the trend and anomalies and how we deal with them for machine learning purposes will be crucial for our analysis.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**A close up of text on a white background

Description automatically generated**

## Representative words

We performed topic modelling to identify key words that were in the positive and negative tweets after extracting the adjectives & nouns and derving the tfidf vectorizations. As we can see not all words provide equal value, however, these representative words help us find the churn indicators and the reason for user churning in subsequent rule-based feature engineering.

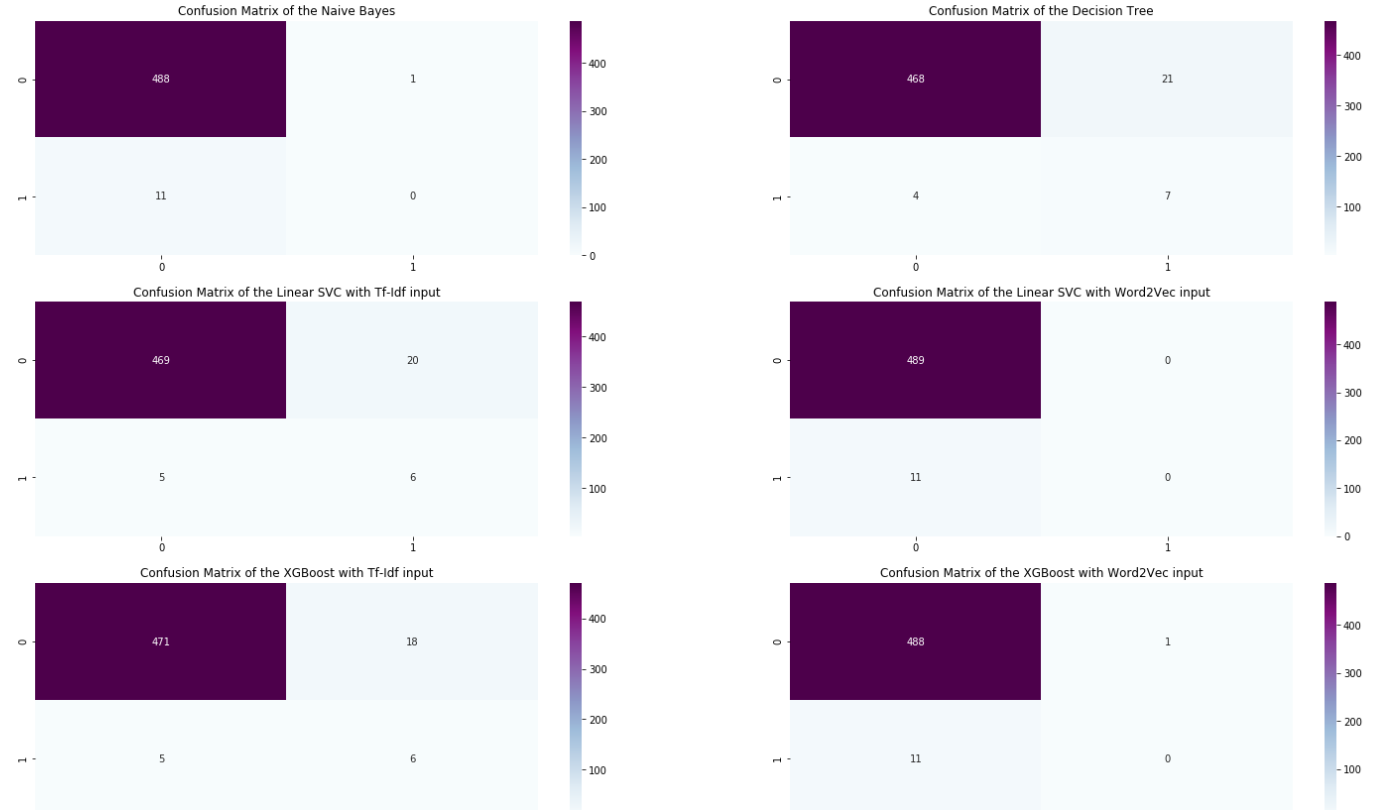


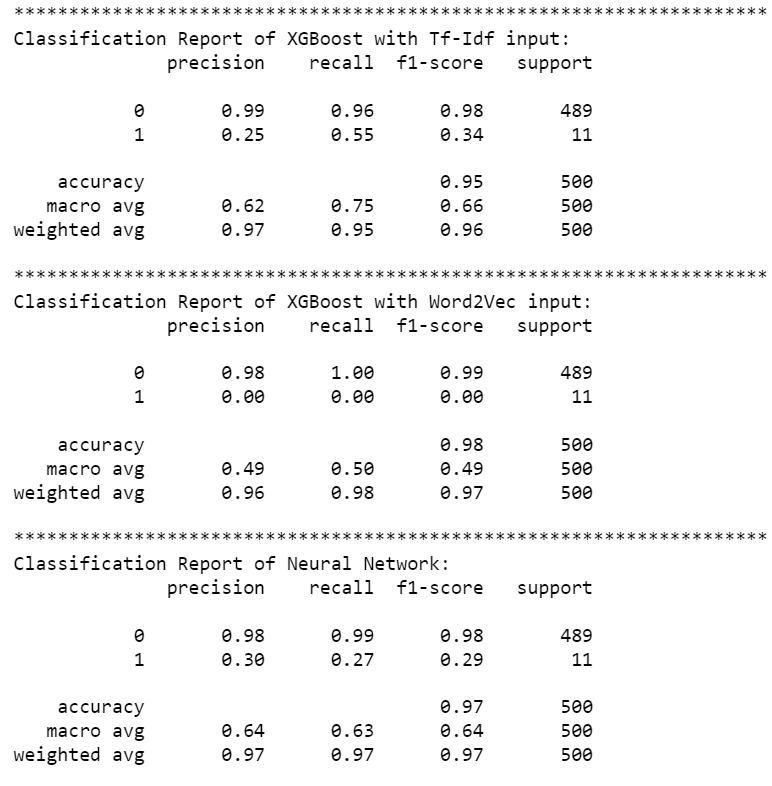
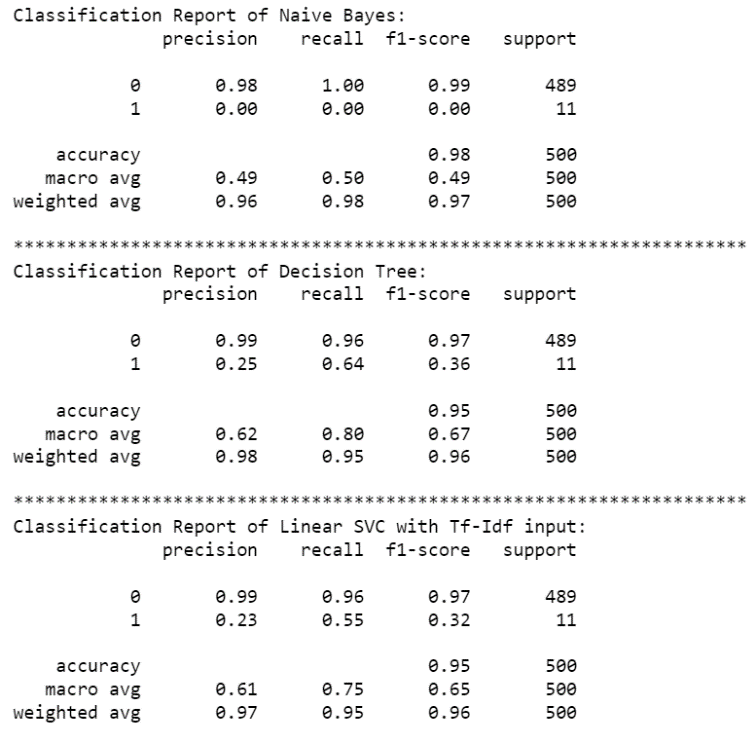
## Churn Labeling

We selected the topics from the representative vocabulary and used them as our churn indicators and based on the churn indicators we labeled a sample training dataset of 8000 records. We created another sample dataset and labeled that manually to use it as our test dataset.

## Churn Detection

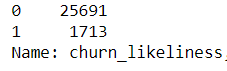
This we confirm by developing several models and cross validating their performance in a test set. We used both TFIDF and word vector inputs and the decision tree with TFIDF as vector inputs provides the highest churn detection accuracy with a good precision and recall and would be our top choice model. As we can see in the below classification reports and confusion matrix that the variation is slight between the models but the LinearSVC and XGBoost, with TFIDF input, perform well. The neural net has room to improve given more data and hyper-parameter tuning. However, the same models with a word vector input have a poor precision and recall.





These results as far as model accuracy and precision are relevant and there is future possibility to explore clustering algorithms to solve the problem from an unsupervised perspective.

## Rule based Feature engineering



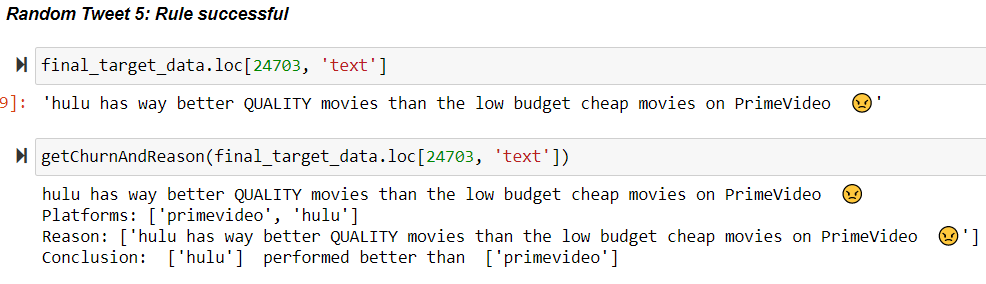
We identified over 1700 tweets that are indicative of churn out of over 27000 tweets giving an average churn rate of 6.25%.

Based on the classification model we selected, we were able to predict the churn likeliness for rest of the tweets, but only the churn likeliness can’t help in identifying the user complaint and the better performing service.

We, therefore, developed a rule-based algorithm that looks for the position of the streaming platform with respect to the churn indicator present in the tweet, and adds a positive/negative score to the platform based on its position. The platform with a high score is our better performing model.

To find out the user complaint, the algorithm divides the tweet into multiple segments and looks for the segment of tweet containing the churn indicator and identify that segment as the user complaint.

Our rule-based feature engineering model is able to predict which among the platforms mentioned is performing better and what were the user complaints. We will, however, improve our algorithm to segment the tweet based on clauses rather than sentences to identify the user complaint.



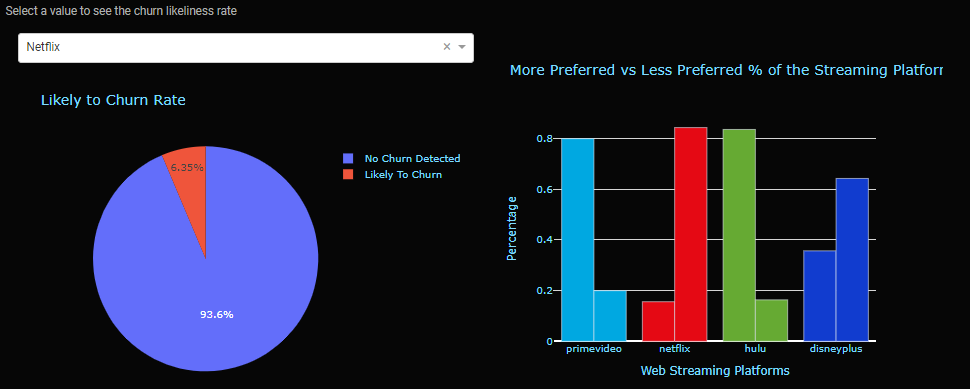
# Dashboard

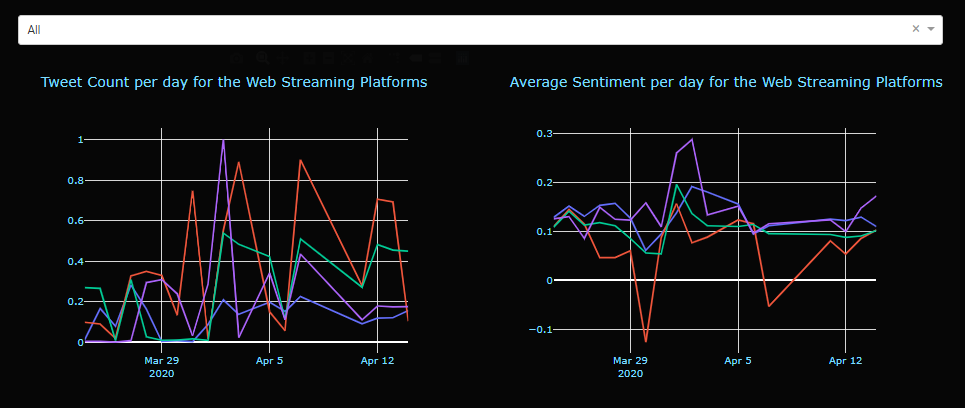
We created a dashboard using plotly and dash. Plotly is an open-source, interactive graphing library for Python.

Dash is written on top of Plotly, and React.js and uses Flask under the hood as the web framework. It is ideal for building visualization apps with interactive user interface using python.

We deployed this dashboard as a web app on a cloud service using Heroku and Git which is a 'platform as a service'.

Heroku has a stable ecosystem with many integrated data services which allows easy integration of different components into the app.





# Key Actionable Insights

* The findings can help the streaming companies to get an idea where they stand in the current market compared to the other streaming services according to the customer tweets. For example, it is visible from the more preferred vs less preferred graph that despite having 69 million US subscribers, out of all the tweets mentioned, Netflix is the less preferred service more than 80% of the time.
* The streaming companies can see from the tweet count and sentiment polarity distribution over time to observe if any recently added content causes any change in customer sentiment and thus, they can identify what kind of content is well received. For example, Disney Plus witnessed a surge in the tweet count on 2nd April, and thus a surge in the overall sentiment polarity. This is well justified as shows like Dr. Dolittle were added on April 2nd to Disney Plus.
* The findings of our analysis can also be helpful to the users to see which platform is more preferred and which is not might helping them decide on which services to subscribe.

# 

# References

<https://scholar.harvard.edu/files/hadi/files/amiri-aaai-16.pdf>

<https://pdfs.semanticscholar.org/f26e/3d935ec705428b292f630d4aae63c24443a5.pdf>

<https://acl-arc.comp.nus.edu.sg/~wing.nus/sig/papers_nlp/20120831_1.pdf>

Twitter as a Corpus for Sentiment Analysis and Opinion Mining

<https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pdf>

SemEval-2016 Task 4: Sentiment Analysis in Twitter

<https://arxiv.org/pdf/1912.01973.pdf>

Churn Identification in Microblogs using Convolutional Neural Networks with Structured Logical Knowledge

<https://www.aclweb.org/anthology/W17-4403.pdf>

Target-Dependent Churn Classification in Microblogs

<https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9849/9527>

# Appendix:

Dashboard Link: <https://twitter-data-analysis-test.herokuapp.com>

GitHub Link: <https://github.com/pruth93/TextAnalytics>