Twitter Analytics

Musarath J. Rahamathullah

Harrisburg University of Science and Technology, Harrisburg, PA.

# ABSTRACT

This paper studies the tweets attitude towards a brand or product, by analyzing the content of tweets from Twitter. It is important to emphasize that this research focuses on automated categorization of information for the purpose of various business requirements and examine abstract based sentiment analysis. The contributions of this research are: (1) To develop an automated text classifier for tweets. (2) A Twitter Analytics tool for brands with high level statistics tracked on Twitter data, uncover insights to help drive decisions.

*Keywords: Twitter, Tweets, Text Classification, Sentiment Analysis, brands.*

**Table of Contents**

ABSTRACT 2

INTRODUCTION 5

LITERATURE REVIEW 7

PROBLEM STATEMENT 12

DATASET 12

METHODOLOGY 13

TECHNICAL APPROACH & REPORTS 16

Data Preprocessing 16

Feature extraction 18

WordCloud 18

Topic Modelling 19

Model Building 20

Naïve Bayes 21

Linear Classifier 21

Support Vector Machine 22

Random Forest (Bagging) 22

Extreme Gradient Boosting 22

Text Classifier 23

Sentiment Analysis 24

TWITTER ANALYTICS TOOL 29

CONCLUSION 30

REFERENCES 31

# INTRODUCTION

Twitter is a buzzword that we hear a lot in the social networking media nowadays. social media allows people to share the information efficiently and expeditiously in real-time on the websites or applications designed for this purpose. Many people intellectualize social media as apps on their mobile devices (e.g., [smartphones](https://en.wikipedia.org/wiki/Smartphone) and [tablets](https://en.wikipedia.org/wiki/Tablet_computer)), but the fact is, these communication tools started with web based technologies on desktops and laptops. The truth that most users access these networking tools through apps is the reason for the misconception among individuals. “With over 313 million monthly active users and a younger demographic to boot, Twitter is a great platform for most marketers”(Jackson, 2019). By analyzing real-time tweets on twitter, we can extract user-generated content for any brand hashtag. With so much information available, this can be used in marketing, to identify unhappy customers of your competitors, resolve a crisis, and more.

Data science is a field of science that extracts insights or information underlying in the data, to solve complex problems not only by predicting into the future values but also by leveraging the process of analyzing and extracting information from textual data. It is not easy to appropriately detect the trends and predict the future, but according to some scientists, there are few things that basically must happen in the future based on experience with data science so far (Kristijan, 2017).

Illustrating an example would be the best way to enlighten how text analytics works and why is it so valued. For instance, a broadcasting company could apply text analysis to scan reviews on Twitter (or any social networking site) for customer feedback about their services. This help the companies to comprehend their customers better or any early alert about when consumers are annoyed with the performance of their service and help resolve the issue before it is notified by customer call, to officially complain or request contract cancellation. According to (Frank, 2018), any company that wishes to enhance one’s business utilizing more data-driven techniques, data science is the secret sauce.

Although unstructured data in the form of text is available everywhere in the form of reviews, survey responses, customer complaints, emails, chats, on social media like Facebook, LinkedIn, Instagram, etc.. and more but extracting insights from these extremely rich source of information can be intense work and time-consuming due to its unstructured nature. Hence, the businesses are turning towards text classification for structuring and analyzing the text in a fast and cost-efficient way to enhance decision-making and automate processes.

In this research, the focus is on one popular microblog, Twitter and build models for classifying tweets around a brand. Categories such as Customer Service, Price, Product, News, and Recommendation are used to classify. This is a multi-label classification problem, where the text of the tweets is tagged to one of the abstract classes. And subsequently build a Twitter Analytics tool, that track all tweets on a specific hashtag of any brand, analyze the tweets and create a report on the particular hashtag brand.

Although this question appears to be a traditional NLP (Natural Language Processing) question, there may be unique challenges when working with real scenarios and data. One problem with text as data is that it piles up quickly. The sheer volume and the problems of handling and sorting through so much information have made text analysis less popular in the past than it is now. And the other reason, because of the unstructured nature of text, understanding and analyzing, the data is tough. Nevertheless, with proper preprocessing, extracting the relevant features of the text, and choosing appropriate classifier to use, could make this research a real time application product. Data needed for this project will be processed in real-time, by using Twitter search API (Application Program Interface). The data collected is initially preprocessed before visualized, to gain insights and then various classification algorithms are applied, to build the best classification model.

# LITERATURE REVIEW

The history of social media backs to 1990s, in 1997 the first social networking site “Six Degrees”, named after social connection theory was born, which allowed users to create profiles and other users. It ceased after a few years in 2001. Next in 2002, two more apps, namely “Friendster” and “LinkedIn” were framed. While the first app was used to search and connect to friends, the later was business-focused social network. But in the next ten years there was a tremendous increase in networking apps, “MySpace”, “YouTube” to share videos, “Reddit”, “Facebook”, “Twitter”, “Instagram”, “Pinterest”, “ Snapchat” to name a few. Recently in 2019 “TikTok” was launched worldwide. Twitter, the concept of blogging and content feeds went live in 2006, by 2012 it achieved 140 million active users. Thus, in a relatively short span of time, social media apps have evolved from cool, sharing platforms, primarily for young demographics into a key platform where individuals can connect with products, people and services in a trending range of ways. Even though social media comes with some negative elements, it’s hard to imagine social platforms not existing.

[According to](https://www.slideshare.net/hootsuite/hootsuite-survey-highlights-importance-of-social-media-across-the-customer-journey/1) (Hootsuite, 2016), nearly half of Americans have interacted with companies or institutions at least on one of the social media networks. With 500 million tweets just on Twitter each day, there is a vast amount of data available, which can provide lots of actionable perceptions to business. Although the text is an extremely rich source of information, mining insights from it is not easy and takes the countless amount of time due to its unstructured nature. Hence the topic of classification has been widely studied in data mining, machine learning, rather data science in general, with applications in a number of diverse domains, such as marketing, financial trading, fraud detection, medical diagnosis, the voice of customers, etc..

Till 1975, for automatic indexing and content or text analysis, a quality of work had been done using statistical or probabilistic techniques. But regardless of efforts, these methods were very complex in nature and lacked in accuracy. Then a method known as discrimination value analysis was introduced which ranks the words in a text in the relevance of each document, from the other. “The value of a term is calculated as how much the average separation between individual document change when the given term is assigned for content identification and the best words are those which achieve greatest separation” (Salton, Yang, & Yu, 1975). This method was relatively simple, compared to older versions of text discrimination by assigning a specific role to each word in text analysis, comparing the words and phrases and word categories. The results were also appealing.

In 1998, (Russell Bernard & Ryan, 1998) did research on text analysis using the quantitative and qualitative methods on human thought and behavior. The research was on content analysis of surveys and in particular “how anthropologists have used those methods to look for meaning and pattern in the written text”(Russell Bernard & Ryan, 1998). Researchers focused on collecting and analyzing text data from political speeches, reviews, the editorial column in the newspaper, transcriptions, personal diaries and so on, with most of which deal with images. They covered two types of text analysis, one “linguistic tradition”, which considers the text as the object of analysis and discovered patterns and structures in them. And the other “sociological tradition”, which considers the text as a window into human experiences, using grounded theory, classical content analysis, semantic network analysis, cognitive mapping, and Boolean analysis.

Later in 2002, research was conducted on text classification using string kernels by (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002). A novel approach to categorizing text documents using kernel is discussed in detail. The term kernel is the product all features of subsequent length(k) i.e. any ordered sequence of k characters in the text. The subsequent characters are weighted by an exponentially decaying factor of the full text. Since the dimension of feature space grows with k exponentially, it involves a large amount of computation even for smaller values of k. The paper described how in spite of computation complexity; it can be calculated using dynamic programming methods. SVM (Support Vector Machine) mainly works on kernels.

Recently in 2016, the Facebook AI research team (Joulin, Grave, Bojanowski, & Mikolov, 2016) utilizes Bag of Words (BOW), to classify text. As now text classification plays an important role in NLP (Natural Language Processing), with many applications such as information extraction, sentiment analysis, search algorithms, and so on. Models based on neural networks have become increasingly popular (Kim, 2014). These models, though take relatively more time to train and test, yield high performance in real-time. Hence these are limited to large dataset usage. Meanwhile, linear classifiers are also considered a strong baselines for text classification problems(Chang, Hsieh, & Lin, 2008). An equilibrium exists between the performance of linear models and the selection of the correct features. According to Agarwal, they do have the potential to scale to the large corpus (Agarwal, Chapelle, Dudík, & Langford, 2014). Armand and the team evaluated the quality of their approach “FastText” on two different tasks, namely tag prediction and sentiment analysis, and found the accuracy to be good and is much faster.

Till now we considered text classification based on a collection of words, an alternative approach that was not much studied is context or word meanings, also known as senses. A research by (Kehagias, Petridis, Kaburlasos, & Fragkou, 2003) compared the classifier efficiency based on words to that of senses. After a series of experiments, the research concluded that the use of senses does not result in any significant improvement in categorization. However, it is not true, which we will reconsider in this research.

Text classification has become essential to classify and maintain the data, with the rapid expansion of information available online every day. Word2vec is an entirely new approach on content classification, by converting the words and phrases into a vector, giving a unique perspective to text mining problems. However, research conducted by (Lilleberg, Zhu, & Zhang, 2015) that word2vec can boost text classification by bringing extra semantic features, prove that the effectiveness of word2vec in combination with term frequency- inverse document frequency(TF-IDF) outperforms any one method individually. The reason being analyzed that word2vec provides complementary features (e.g. semantics that TF\_IDF cannot capture) to TF\_IDF.

The machine learning algorithms were also considered for text classification. Various studies and research compared the feature selection technique to feature space transformation and even compared the performance of different algorithms. With the rising interest in Support Vector Machine (SVM) algorithms, researches showed that SVM outperformed all other classifier models. Later a team(Colas & Brazdil, 2006) decided to investigate the truth beyond this myth and test SVM over Naive Bayes and K-NN on binary classification. Results could not prove that SVM outperforms other classifiers, but all the classifiers achieved comparable efficiency. If an appropriate preprocessing is done, even KNN can achieve good results. As for Naïve Bayes it also achieved good accuracy. Hence preprocessing is the key point in getting decent outcomes in classification problems.

Nowadays there is a great attraction towards the crisis in social media by all demographic people and news media. A study investigated the possibilities of real-time and automated analysis of Twitter messages during such a crisis (Teun Terpstra, R. Stronkman,A. de Vries, 2012). This research analysis was performed on 97,000 tweets that were published in a crisis that happened in Belgium in 2011. A storm name “Pukkelpop” hit the region and there was a great amount of loss in the region. It was noticed that immediately after the storm struck, the tweets increased exponentially at the rate of 576 tweets per minute. And hence analysis was performed using an information extraction tool, which analyzed through geographical visualization, message filters like damage, causalities, etc. and tweet type filters (e.g., retweets). The important topics that emerged out of this analysis were “‘early warning tweets’, ‘rumors’ and the ‘self-organization of disaster relief’ on Twitter” (Teun Terpstra, R. Stronkman,A. de Vries, 2012). Results projected that automated filtering or text classification provides valuable information to society during a crisis.

“Currently Twitter receives about 190 million tweets (small text-based Web posts of 140 characters) a day, in which people share their comments regarding a wide range of topics” (Hao et al., 2011). Opinions about products and services contribute to a larger portion of tweets. Though Twitter being one of the influential social media platforms, being a new phenomenon these tweets are not utilized much as a source for evaluating customer sentiment.  Moreover much “research remains centered around isolated cases, for instance, events in political communication, crisis communication, or popular culture, often coordinated by shared hashtags”(Bruns & Stieglitz, 2013).

Another research on understanding professional athletes use of twitter to converse directly with their followers by (Hambrick, Simmons, Greenhalgh, & Greenwell, 2016) used content analysis of tweets. The research used text classification analysis to categorize 1,962 tweets by professional athletes into one of six categories namely interactivity, diversion, information sharing, content, promotional, and fanship. Interestingly most of the tweets fell into the interactivity category (34%). This motivated me to work on products or brands because people care about “What people are talking about a particular brand” before making their decision to buy or not. This research will analyze aspect-based sentiment analysis over a period of time.

# PROBLEM STATEMENT

Twitter is a microblogging and social networking service on which users post and interact with messages known as "tweets". And any brand and its competitors deeply influence its customers by these online discussions. It is needed to create a process that keeps a close watch on brand indications, to gain extra insights that help consumers to drive decisions. And through text classification, we can categorize brand text to get the answers more clearly and quickly, like “Products” category can give a view about what people are talking about a particular feature or service, “Price” category suggest whether the brand perceived has good value for money or not?

The objective is to create a text classifier that helps recognize topics when anyone tweets about a brand. Additionally, topic classifiers combined with sentiment analysis to get real-time measurement, to act when needed.

# DATASET

The data required for this research will be collected from Twitter social media. First, for retrieving data, the user should get Twitter developer permissions, and for that, we need Twitter account itself and create a new app. Once the developer account is approved, find the access credentials under the Keys and Access Tokens tab of the new app of the user account. Now that I got access, will next install Tweepy a python module using pip, which provides a front end for Twitter API. Using Tweepy.Cursor method searches the tweets to a particular hashtag. The tweets data thus obtained will be retrieved and stored in a csv/xlsx file. Though there are number of features in this data set, only date field and tweets text which is not more than 280 characters will be considered for this research.

# METHODOLOGY

The tweets data stored in the csv file are read and preprocessed before further analysis. The this is an essential step in text preprocessing, as it makes the tidy text prepared for data mining, i.e., it becomes easier to get abstract information from the text and apply machine learning algorithms to it. For this purpose first, removing Twitter handles (i.e. @user) in the tweet’s dataset, as they do not convey information needed for the analysis. Actually twitter handles are how the user acknowledges on Twitter. Next is removing punctuations, numbers or any special characters in the text. Later remove any short words, i.e. all the words having a length less than 3. For example terms like ‘hmm’ or ‘ok’. Finally tokenize all the cleaned tweet text. Tokenization is the process of splitting a string of text into individual terms or words. Lemmatize each term, i.e. normally aiming to remove inflectional endings and to return the base or dictionary form of a word, which is known as the *lemma*. Now these cleaned tweets are ready for analyzing.

For understanding the common words in the tweets, a WordCloud is built to accomplish this task. Here the words that occur most frequently appear in bigger size compared to words with less frequency. In order to analyze the data, it needs feature extraction from cleaned tweets. Text features can be extracted using assorted techniques like Bag of Words (BoW), Term Frequency Inverse Document Frequency (TF-IDF), Word Embeddings. This research considers TF-IDF method, because it takes into account, not just the occurrence of a word in a single tweet but in the entire corpus. TF-IDF gives importance to words which are not frequent in the entire corpus, but appears in more in a few documents, by penalizing the most frequent terms with lower weights. And the TF-IDF vectors are generated at different levels of tokens i.e. words, characters, and n-grams.

Using Latent Dirichlet Allocation for generating Topic Modelling Features, which is a technique to identify the group of words from a collection of tweets and bind them to a topic. LDA model is built and the probability distributions of words given by topics provide a sense of the different intuition contained in the documents.

As the last step in text classification, to train the classifier with the features created in the previous step. Machine learning has a wide range of models to train the final model. These models will be implemented one after the other and compared the accuracy score. Before training the models, first tagging of tweets is done, to one of the five categories viz Customer Service, Price, Products, News and Recommendation. And second the data is split into validation and testing set in the ratio of 75% to 25% respectively.

First implement Naïve Bayes model based on Bayer’s theorem, with an assumption of independence among features in the dataset. This is done using sklearn implementation on the features collected above. The trained model is tested over the testing dataset and the accuracy parameters are measured. Next Logistic regression will be built on the features extracted, which measures the probability of belonging to one group using sigmoid function. And another classification algorithm known as support vector machine (SVM) which extracts best possible hyper plane that separates the classes. Features are trained using SVM and the efficiency of the model is obtained and studied. Later Random forest model is trained. It is an ensemble technique that combines several decision tree models ran in parallel into one predictive model in order to decrease variance or bagging and the model tested with testing dataset. Finally Extreme Gradient Boosting (XGBOOST) Model is implemented. This model is similar to gradient boosting framework where it implements both linear model and tree learning algorithms. What makes XGBOOST fast, is its capacity to do parallel computation and primarily use for reducing bias problem which is caused due to less features. All these models will be implemented in the Scikit Learning package in python.

All the models built using different classification algorithms are then compared and studied for the best results. Classification accuracy, i.e. the rate of the test set for which the model predicted the correct output will be the most common metric. However, imbalanced dataset accuracy will not give the true performance of the model. Hence precision and recall will be calculated to interpret the best performing model. Precision will give the false positives or Type I error for each class, while recall gives the false negative or Type II error. Later the F1 score is calculated as the average of precision and recall. The performance score of all the models in descending order will be tabulated to illustrate the best model to predict the tag for the given tweet. This research is looking at building an app, that can run text classification on a file comprising of tweets.

In the second part of the research abstract based sentiment analysis is performed on the tagged dataset. Using SentimentIntensityAnalyzer in the nltk package the sentiment of tweets is obtained, which is then classified into positive , negative and neutral groups depending on the sentiment score is above or below zero value. With these tweets data, various plots are generated to visualize and analyze the tweets of a particular product or business.

Finally a Twitter Analytics tool will be developed, which will have the ability to measure hashtag performance. Association of topic classifier with sentiment analysis is used to get a real-time thermometer, about online information on a particular brand for getting actionable insights. This tool will create a report on the hashtag, which comprises of tweets distribution over time, Sentiments of tweets w.r.t different aspects and much more.

# TECHNICAL APPROACH & REPORTS

## Data Preprocessing

Using the developer access credentials obtained the tweets data on “Samsung” hashtag and write the tweets to the corresponding csv file, named as *tweets\_with\_hashtag\_(#name).csv*. Approximately 5k records were retrieved for the given hashtag.



**Figure 1: Number of tweets**

The tweets obtained using API search are as follows.



Later read the csv file and preprocess all the tweets as follows: a) removing twitter handles (@user), b) removing the punctuations, numbers and special characters, c) remove all urls, d) remove all short words with length less than or equal to three characters. The tweets after preprocessed is as follows. 

Now I tokenized all the cleaned tweets in the dataset. Tokens are the individual terms in the tweets. 

**Figure 2: Tokens of tweets**

The tokens are lemmatized to its base or dictionary form of form using nltk package. For instance terms like “studies”, “studying” to “study”.



**Figure 3: Lemmatize Tokens**

## Feature extraction

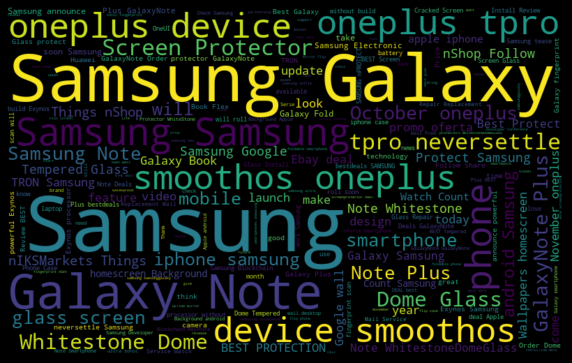
To represent the text into numerical features, TF-IDF method is used, which calculates as given.

* TF = (Number of times token t appears in a tweet)/(Number of tokens in the tweet)
* IDF = log(N/n), where, N is the number of tweets and n is the number of tweet corpus a token t has appeared in.
* TF-IDF = TF\*IDF

TF-IDF Vectors is generated at different levels of tokens like words, characters, and n-grams.

* Word Level TF-IDF : Matrix representing tf-idf scores of every token in different tweets.
* N-gram Level TF-IDF : N-grams are the combination of N tokens together. This Matrix representing tf-idf scores of N-grams.
* Character Level TF-IDF : Matrix representing tf-idf scores of character level n-grams in the corpus of tweets.

## WordCloud

To understand the common words used in the tweets, is accomplished by plotting wordcloud. The most frequent words appear in big letters and less frequent in smaller size shown in figure(4) below.

**Figure 4: WORDCLOUD of cleaned tweets**

## Topic Modelling

I used Latent Dirichlet Allocation (LDA), for generating Topic Modelling Features. By means of Topic Modelling identified the groups of words that belong to one group called a topic. In this number of topics and number of words in each topic is considered as ten is generate

**Figure 5: Topic Model of tweets**

## Model Building

The dataset is manually annotated to five categories viz Ease of use, Products, Price, News and Recommendation, based on keywords. Before training the model, the dataset is divided into training and test set in the ratio of 75-25 respectively and stored in the train and test dataframe and models are trained one by one. The trained dataframe looks like the given table below.

**Table 1: Train Dataframe**



### Naïve Bayes

Implementing the Naïve Bayes supervised classification model, using sklearn package on different training features. The results of which are as shown below.

**Table 2: Results of Naïve Bayes.**

|  |  |  |  |
| --- | --- | --- | --- |
| TFIDF Level | Precision | Recall | F1-Score |
| Word Level | 0.82 | 0.69 | 0.71 |
| N-Gram Level | 0.77 | 0.59 | 0.60 |
| Char Level | 0.82 | 0.62 | 0.65 |

### Linear Classifier

Implementing the logistic regression to measure the relation between categorical predictor and features. The results are as shown below.

**Table 3: Results of Logistic Regression.**

|  |  |  |  |
| --- | --- | --- | --- |
| TFIDF Level | Precision | Recall | F1-Score |
| Word Level | 0.84 | 0.78 | 0.79 |
| N-Gram Level | 0.80 | 0.55 | 0.58 |
| Char Level | 0.88 | 0.84 | 0.85 |

### Support Vector Machine

Implementing the Support Vector Machine algorithm of supervised learning, the results are as shown below in table(4).

**Table 4: Results of SVM**

|  |  |  |  |
| --- | --- | --- | --- |
| TFIDF Level | Precision | Recall | F1-Score |
| Word Level | 1.00 | 0.32 | 0.48 |
| N-Gram Level | 1.00 | 0.32 | 0.48 |
| Char Level | 1.00 | 0.32 | 0.48 |

Here the results indicate that SVM is not appropriate for the given set of tweets dataset.

### Random Forest (Bagging)

An ensemble bagging model, Random Forest is implemented on the feature levels and the accuracy scores are as shown below in table(5).

**Table 5: Results of Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
| TFIDF Level | Precision | Recall | F1-Score |
| Word Level | 0.84 | 0.81 | 0.82 |
| N-Gram Level | 0.70 | 0.52 | 0.53 |
| Char Level | 0.91 | 0.87 | 0.89 |

### Extreme Gradient Boosting

Boosting models are another type of ensemble models are implemented using Extreme Gradient Descent algorithm. The efficiency of model at different levels of features are as shared below in table(6).

**Table 6: Results of Extreme Gradient Boosting**

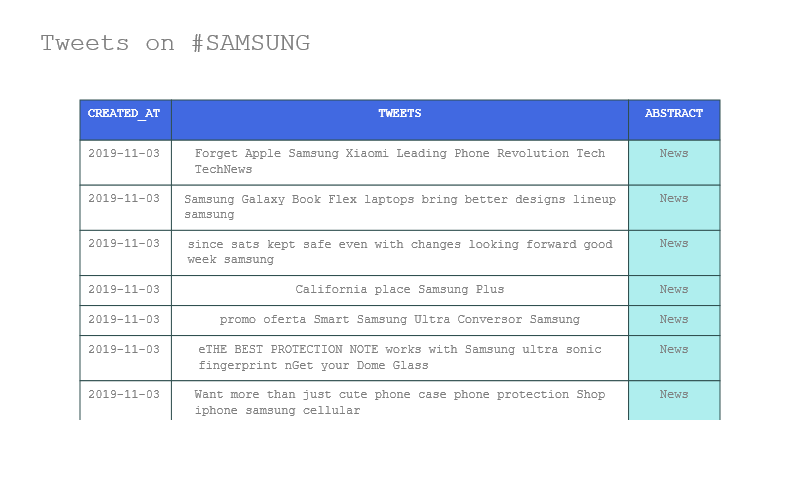
|  |  |  |  |
| --- | --- | --- | --- |
| TFIDF Level | Precision | Recall | F1-Score |
| Word Level | 0.87 | 0.83 | 0.83 |
| N-Gram Level | 0.71 | 0.52 | 0.54 |
| Char Level | 0.92 | 0.91 | 0.91 |

### Text Classifier

Comparison of all the results from different models, indicate that boosting method was good in classifying the text (tweets). Hence it was considered as the final classifier model for the tweets. Using Flask, a frontend user interface is developed, where user can upload a file of tweets and the application automatically annotates the tweets.



**Figure 6: Twitter Classification application.**



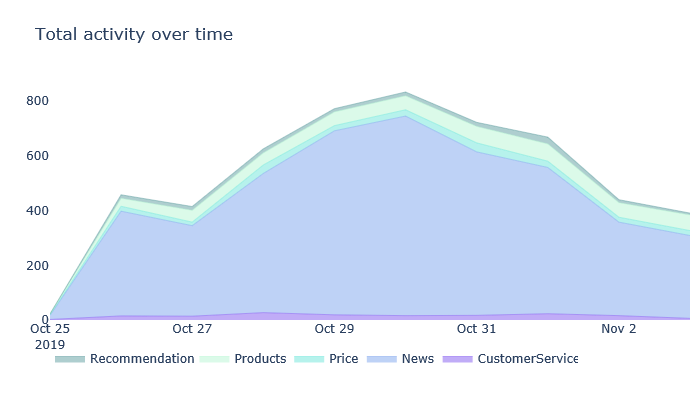
**Figure 7: Tweets classified**

## Sentiment Analysis

The sentiment of each tweet is retrieved using the SentimentIntensity Analyzer in Nltk package. And all the sentiment scores are grouped to Positive, Negative and Neutral categories, depending on the value of the score of each tweet. These values are stored in the tweets dataframe.

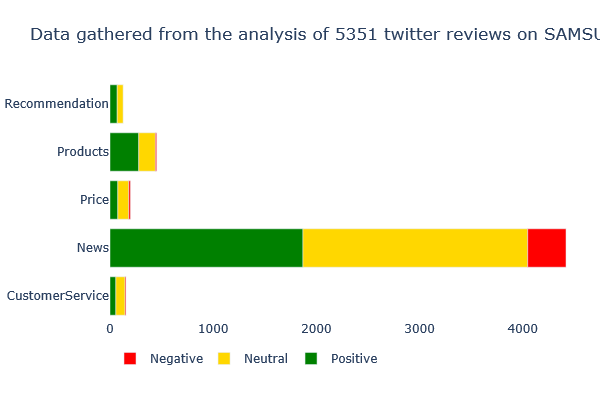


The tweets activity over a period of 10days is watched using the plot below. From the plot, we can say that October 30th was the day when a lot of interactions happened on twitter on “Samsung”, nearly 20% more and most of the tweets were of some sort of news.



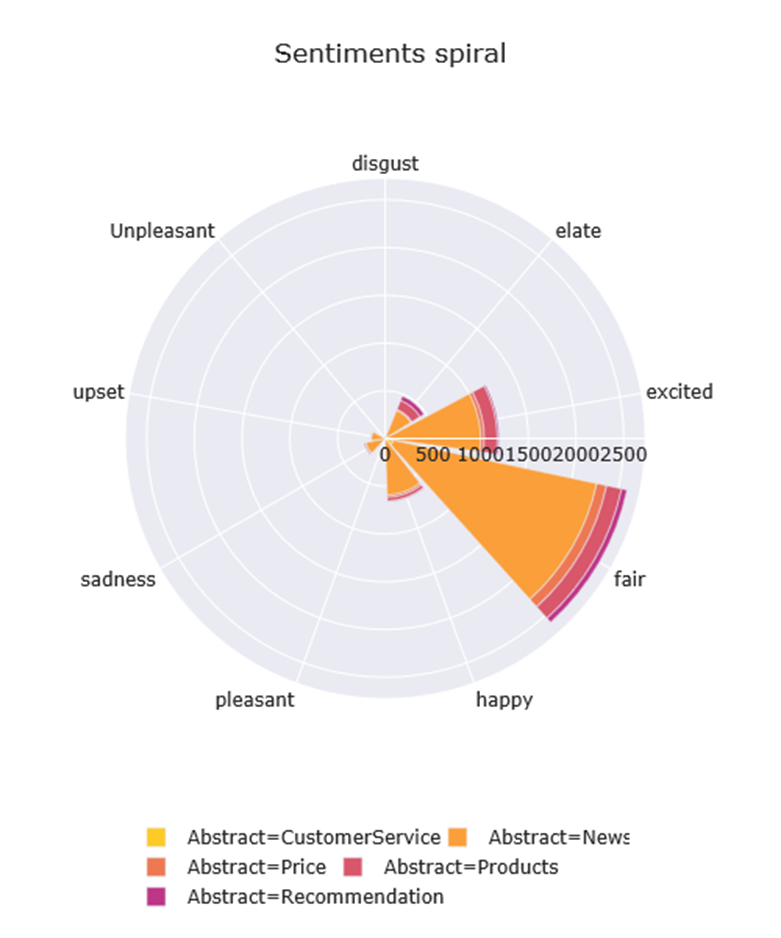
**Figure 8: Samsung Tweets over a period of time**

Next sentiment of tweets is studied using barplot w.r.t abstracts. As we see, people have been mostly positive or neutral about the Samsung products or brand, although there is slight (i.e. 8% approximately) negative comments about news releases and less than 5% about their prices, customer service and products . Hence we can conclude that Samsung has good reviews on Twitter.



**Figure 9: Tweets Sentiment bars**

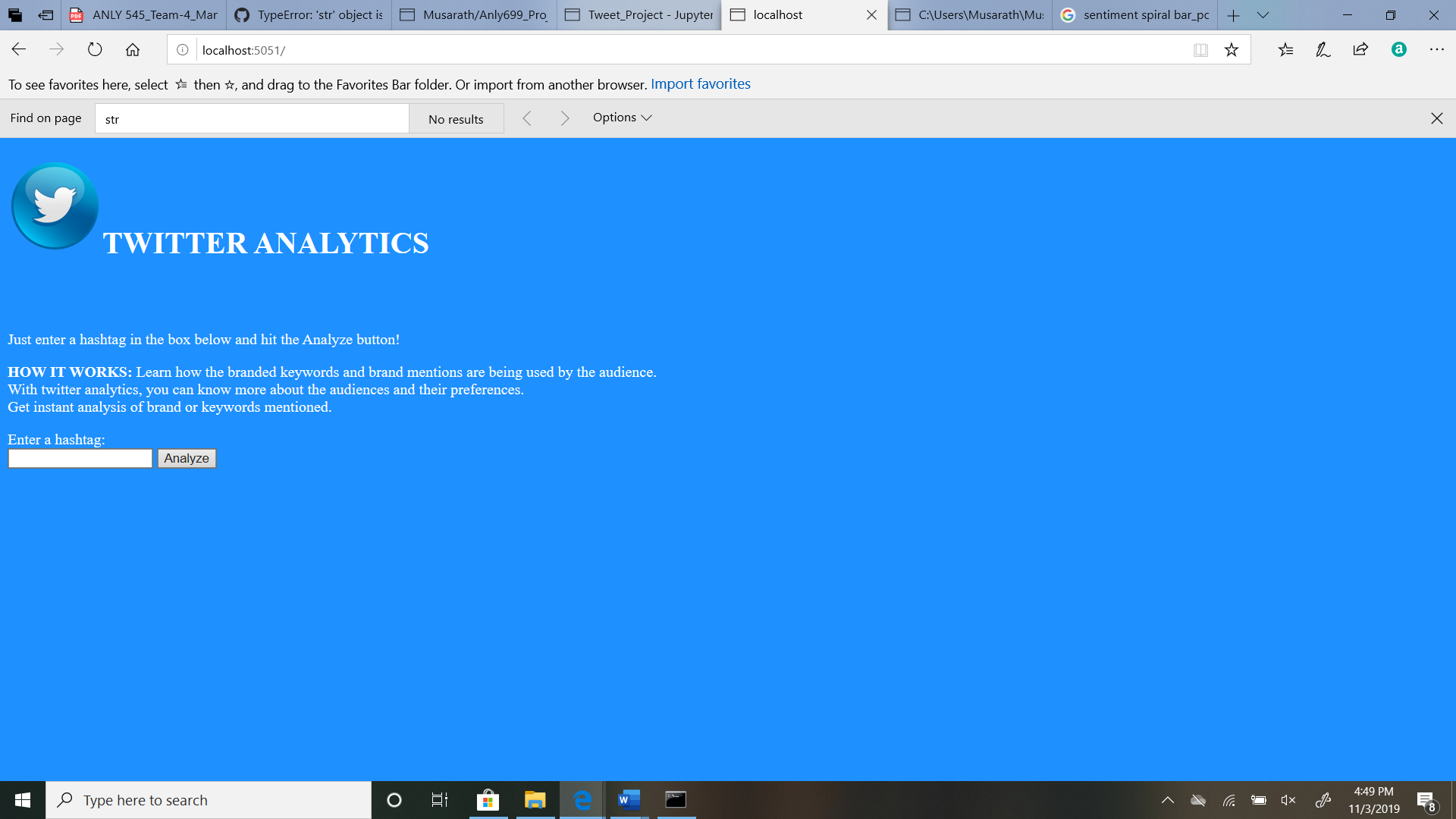
A [wind rose chart](https://en.wikipedia.org/wiki/Wind_rose), also known as a polar bar chart is a graphical tool used to visualize how abstract or categories of tweets and emotions are typically distributed in case of (here) Samsung brand. The plot below illustrates that there is good amount fairness regarding Samsung brand among people on Twitter.



**Figure 10: Sentiment Spiral of tweets**

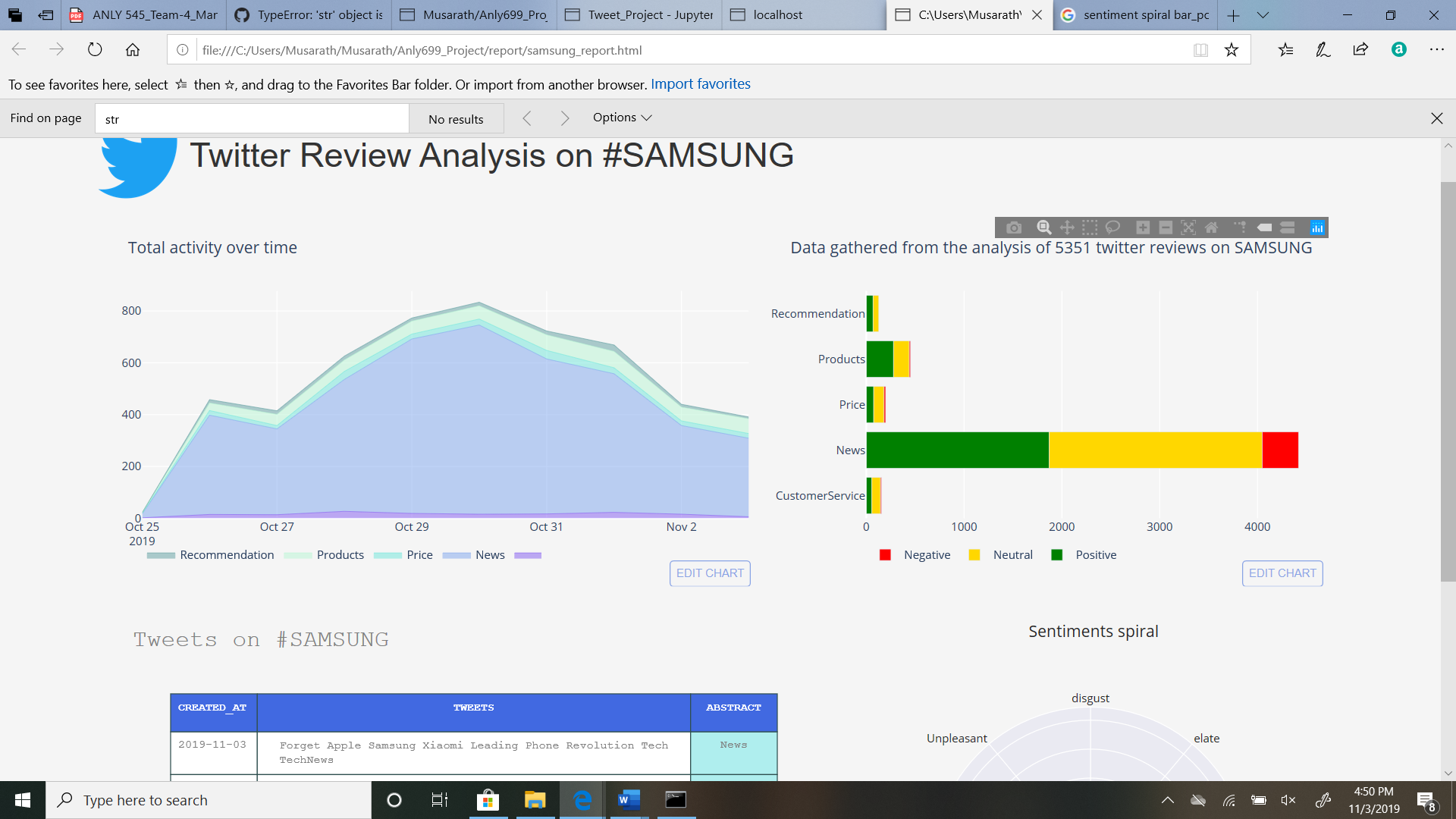
# TWITTER ANALYTICS TOOL

With the text classification using the trained text classifier model, Sentiment Analysis an application is developed in Flask (python). The User Interface(UI) screen appears like shown in the figure(11) preceding. Here user can enter keywords or brand names and click on “Analyze” button. The application will find the interactions corresponding to the keyword and first classify the text and then run sentiment analysis.



**Figure 11:Twitter Analytics Tool**

The report generated by the application is as published below in the figure(12).



**Figure 12: Report on Samsung Tweets**

# CONCLUSION

In conclusion, this research presented a collection of algorithms that addressed the problem of labelling the text and the experiments evaluated the best training classifier model, which is a significant task in text learning in the real world because of availability of huge volumes of unlabeled data and it represent the lot of actionable insights for business. Paper presented the results of text classifier from various approaches and report showed that Extreme Gradient descent method, had the highest accuracy(88%) and F1-score of 0.91.

The project developed two UI applications, one an automated text classification tool to classify tweets automatically. And second, the Twitter Analytics tool which analyzes the tweets and generate a report, to know what is being said about a particular brand or product across social media (here Twitter in particular).

I realized that people most commonly share the news interactions on Twitter, more than reviews. Although there is insight of information on a brand, the overall gain may be bleak. So in future work, will consider interaction content from all social media networks like Facebook, Instagram, blogs, customer complaints etc.. And attempt to increase the efficiency of text classification using deep learning neural network methods of AI.

# REFERENCES

Agarwal, A., Chapelle, O., Dudík, M., & Langford, J. (2014). A Reliable Effective Terascale Linear Learning System. In *Journal of Machine Learning Research* (Vol. 15). Retrieved from http://hadoop.apache.org/.

Bruns, A., & Stieglitz, S. (2013). Towards more systematic Twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, *16*(2), 91–108. https://doi.org/10.1080/13645579.2012.756095

Chang, K.-W., Hsieh, C.-J., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification Rong-En Fan Xiang-Rui Wang. In *Journal of Machine Learning Research* (Vol. 9). Retrieved from http://www.csie.ntu.edu.tw/

Colas, F., & Brazdil, P. (2006). Comparison of SVM and some older classification algorithms in text classification tasks. *IFIP International Federation for Information Processing*, *217*, 169–178. https://doi.org/10.1007/978-0-387-34747-9\_18

Frank, L. (2018). *What Is Data Science? What is a Data Scientist? What is Analytics?* Retrieved from https://datajobs.com/what-is-data-science

Hambrick, M. E., Simmons, J. M., Greenhalgh, G. P., & Greenwell, T. C. (2016). Understanding Professional Athletes’ Use of Twitter: A Content Analysis of Athlete Tweets. *International Journal of Sport Communication*, *3*(4), 454–471. https://doi.org/10.1123/ijsc.3.4.454

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L. E., & Hsu, M. C. (2011). Visual sentiment analysis on twitter data streams. *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings*, 277–278. https://doi.org/10.1109/VAST.2011.6102472

Hootsuite. (2016). Hootsuite Survey Highlights Importance of Social Media Across the Cus…. Retrieved October 29, 2019, from https://www.slideshare.net/hootsuite/hootsuite-survey-highlights-importance-of-social-media-across-the-customer-journey/1

Jackson, D. (2019). *Twitter Marketing: Strategies and Tips to Try Now | Sprout Social*. Retrieved from https://sproutsocial.com/insights/twitter-marketing/

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of Tricks for Efficient Text Classification*. Retrieved from https://github.com/facebookresearch/fastText

Kehagias, A., Petridis, V., Kaburlasos, V. G., & Fragkou, P. (2003). A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, *21*(3), 227–247. https://doi.org/10.1023/A:1025554732352

Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1746–1751. https://doi.org/10.3115/v1/d14-1181

Kristijan, J. (2017). *You Should Really Know About The Importance of Data Science and its Future | Welcome to my world!* Retrieved from http://www.kristijanjanusic.com/blog/you-should-really-know-the-importance-of-data-science-and-its-future/

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, ICCI\*CC 2015*, 136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text Classification using String Kernels. In *Journal of Machine Learning Research* (Vol. 2).

Russell Bernard, H., & Ryan, G. (1998). *Text Analysis: Qualitative and Quantitative Methods*.

Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, *26*(1), 33–44. https://doi.org/10.1002/asi.4630260106

Teun Terpstra, R. Stronkman,A. de Vries, G. L. P. (2012). *Towards a realtime Twitter analysis during crises for operational crisis management*. Retrieved from http://www.nrc.nl/nieuws/2011/08/18/pukkelpop-stilgelegd-na-noodweer/