**Sentiment Analysis of Airline Tweets**

**Abstract:**

The project performs sentiment analysis on the tweets mentioning each major U.S. airline. Twitter data was scraped from February of 2015 and contributors were asked to first classify positive, negative, and neutral tweets, followed by categorizing negative reasons (such as “late flight” or “rude service”). In our project we used our model, which we trained using naïve bayes classifier to predict if the tweets were either positive, negative or neutral.

We cross check our predictions with what the users have classified those tweets as, and using this knowledge we generate the classification report containing the precision, accuracy, F1 scores and other parameters. At the end of execution, the user can himself input some tweets and test out the model. The train dataset and test dataset are randomized during each execution which doesn’t decrease the probability of the code.

**Introduction:**

Sentiment is opinion or judgement prompted by feeling. Sentiment analysis involves extracting the opinion of the user on a scale based on the feedback given by the user. Sentiment Analysis can be applied on any type of events or categories the results can be used to make satisfactory changes to them.

In our project, we performed sentiment analysis on twitter data. This data was generated through the users of major airlines on Twitter. The keywords can be extracted from the negative tweets and the airlines can improve upon those aspects. For instance, if the major reason in the negative tweets is delay airlines can work on properly scheduling them and making sure that they arrive at due time. This can also be used for identifying the positive tweets also, and it helps in identifying the strengths of the airline company.

**Contributions:**

Work was split equally amongst us, each of us studied to research papers and wrote the report. Coding was done together on teams.

**Citations:**

# Sentiment Analysis of Review Datasets Using Naïve Bayes Classifier

Introduction Web 2.0 has led to dramatic increase in the amount of sentimental content available in the Web. Such content is often found in social media like movie reviews, product reviews and travel or place reviews, user comments, testimonials, messages in discussion forums etc. Discovery of the sentimental or opinionated web content has plenty of advantages. Understanding of the sentiments of human masses towards different entities and products enables better services for contextual advertisements, recommendation systems and analysis of market trends. This paper mainly focused on sentiment focused web crawling framework to facilitate the quick discovery of sentimental contents of movie reviews and hotel reviews and analysis of the same. Statistical methods to capture elements of subjective style and the sentence polarity were used. The paper elaborately discussed two supervised machine learning algorithms: K-Nearest Neighbor (KNN) and Naïve Bayes ‘and compares their overall accuracy, precisions as well as recall values. It was seen that in case of movie reviews Naïve Bayes ‘gave far better results than K-NN but for hotel reviews these algorithms gave lesser, almost same accuracies.

**Datasets** used in this paper: IMDB(Movie Reviews) and OpinRank(Hotels).

# Implementation of Sentiment Analysis on Twitter Using Naïve Bayes Algorithm to Know the People Responses to Debate of DKI Jakarta Governor Election

The widespread use of social media has resulted in the availability of a large textual information so that it appears the need of presenting data that will allow users to get accurate information. This research was conducted on sentiment analysis of Twitter users against election debates governor Jakarta to know the public response to some public figures. Data was collected from twitter before the elections, Many, of the people who responded to a public figure in the social media Twitter. Public figures referred to in the study was the candidate of the Jakarta governor which are Basuki Tjahaja Purnama, Anies Rasyid Baswedan, and Harimurti Agus Yudhoyono. Based on this research, it will use the keyword 'Basuki' or 'Ahok', 'Agus', 'Badwater', 'pilgub', 'election' and 'debat' on Twitter sentiment in Indonesian language. Tweet containing a sentiment that has been collected, was seen whether its profile location so it can be considered whether the user was included in the province of Jakarta or outside of Jakarta. Each tweet sentiments of these two categories were then pre-processed to make it easy to be processed and analyzed. After passing through the phase, it was calculated for each category with Naive Bayes algorithm in order to get the value of the sentiment of the people residing in the provinces of Jakarta and outside Jakarta to the three candidates for the Jakarta governor.

# Sentiment Analysis using Naive Bayes and Complement Naive Bayes Classifier Algorithms on Hadoop Framework

Sentiments can be classified using machine learning methods, lexicon-based methods and hybrid which is combination of machine learning techniques and lexicon-based technique. In this study, sentiment analysis was conducted using machine learning techniques such as Naive Bayes and Complement Naive Bayes Algorithms using Hadoop software framework. Experiments were carried out using varying sizes of training datasets and about 8 million of reviews were classified as positive, negative and neutral. Performance of the algorithms were compared according to accuracy, precision, recall, and F-measure performance evaluation criterions.

In this study, a training set is created by combining positive words dataset and negative words dataset. While positive words are assigned to “1” class, negative words are assigned to “0” class. In this training set, there are 2005 positive and 4783 negative words. A new training is created by scaling up this dataset with multipliers of 2, 4 and 8. System is trained using Naive Bayes and complement Naive Bayes classification algorithms using the optimum parameters which are obtained empirically for all size of the training sets. To get optimum parameters which gives the best accuracy for Naive bayes and complement Naive bayes algorithms, “20 Newsgroups” dataset which include 18846 documents and 20 different news groups was used. This dataset was scaled up with multiplier of 2, 4 and 8.

Experiments were carried with different parameters of Naive Bayes and complement Naive bayes algorithms with different size of training sets. System was trained and tested on Hadoop framework and Mahout library. To test the models, “Amazon movie reviews” dataset was used. This dataset includes about 8 million reviews.

# Sentiment Analysis of Movie Reviews Using Hybrid Method of Naive Bayes and Genetic Algorithm

The area of sentiment mining (also called sentiment extraction, opinion mining, opinion extraction, sentiment analysis, etc.) has seen a large increase in academic interest in the last few years. Researchers in the areas of natural language processing, data mining, machine learning, and others have tested a variety of methods of automating the sentiment analysis process.

In this research work, new hybrid classification method is proposed based on coupling classification methods using arcing classifier and their performances are analyzed in terms of accuracy. A Classifier ensemble was designed using Naive Bayes (NB), Genetic Algorithm (GA). In the work, a comparative study of the effectiveness of ensemble technique is made for sentiment classification. The ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. The feasibility and the benefits of the proposed approaches are demonstrated by means of movie review that is widely used in the field of sentiment classification. A wide range of comparative experiments are conducted and finally, some in-depth discussion is presented and conclusions are drawn about the effectiveness of ensemble technique for sentiment classification.

# Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning

The growing expansion of contents, placed on the Web, provides a huge collection of textual resources. People share their experiences, opinions or simply talk just about whatever concerns them online. The large amount of available data attracts system developers, studying on automatic mining and analysis. In this paper, the primary and underlying idea is that the fact of knowing how people feel about certain topics can be considered as a classification task. People’s feelings can be positive, negative or neutral. A sentiment is often represented in subtle or complex ways in a text. An online user can use a diverse range of other techniques to express his or her emotions. Apart from that, s/he may mix objective and subjective information about a certain topic. On top of that, data gathered from the World Wide Web often contain a lot of noise. Indeed, the task of automatic sentiment recognition in online text becomes more difficult for all the aforementioned reasons. Hence, we present how sentiment analysis can assist language learning, by stimulating the educational process and experimental results on the Naive Bayes Classifier.

# Sentiment Analysis for Various SNS Media Using Naive Bayes Classifier and Its Application to Flaming Detection

SNS is one of the most effective communication tools and it has brought about drastic changes in our lives. Recently, however, a phenomenon called flaming or backlash becomes an imminent problem to private companies. A flaming incident is usually triggered by thoughtless comments/actions on SNS, and it sometimes ends up damaging to the company’s reputation seriously. In this paper, in order to prevent such unexpected damage to the company’s reputation, a new approach was proposed to sentiment analysis using a Naive Bayes classifier, in which the features of tweets/comments are selected based on entropy-based criteria and an empirical rule to capture negative expressions. In addition, a semi-supervised learning approach was proposed to re-labelling noisy training data, which come from various SNS media such as Twitter, Facebook, blogs and a Japanese textboard called ‘2-channel’. In the experiments, four data sets of users’ comments were used, which were posted to different SNS media of private companies. The experimental results show that the proposed Naive Bayes classifier model has good performance for different SNS media, and a semi supervised learning effectively works for the data consisting of long comments. In addition, the proposed method is applied to detect flaming incidents, and it’s shown that it is successfully detected.

**Objective:**

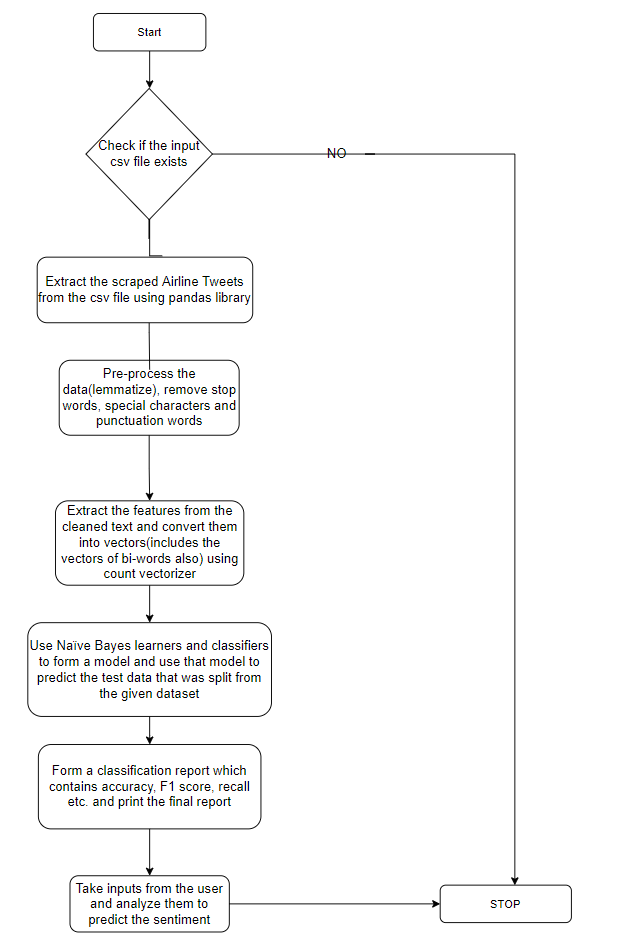
Objective is to identify the sentiment of tweets generated by the users on Twitter. Our program would train a model which would then perform sentiment analysis on the above-mentioned data. It would be able to identify the sentiment with as much accuracy as possible.

**Motivation:**

Many times, on twitter we see many trending hashtags, but in order to understand whether they are trending, either for negative or positive reason is hard. The tweets which pop up soon after we click the tag might be negative but not all of them are the same. Finding majority opinion is hard. Given the data frame our program will be able to identify the sentiment and it would be easy to identify why the tag is trending.

We can use the code for various data frames, it’s not limited to airline dataset.

**Proposed Model:**



**Methodology:**

**Data Extraction:**

Initially data is extracted from the csv file using pandas library. Unnecessary columns are removed and only the sentiment and text data are extracted from the csv file.

**Pre-processing:**

Data is pre-processed to remove unnecessary data. We start by removing stop words downloaded from nltk corpus. Then URLs which are frequently found in tweets are removed. Punctuation marks are removed from the tweets. HTML tags are also removed from the tweets. Usernames are also removed, they generally start with @ symbol, so they can be identified easily. Emojis can also be removed based on their unicodes. Contractions which are common in chat languages are also removed for instance won’t will be converted into will not have. continuous repeated characters are also removed. All the above-mentioned actions are done through functions which are called to perform their respective actions.

Final text is then stored into separate variables. Even the sentiments are converted into numerical formats so the program could understand them.

**Vectorization:**

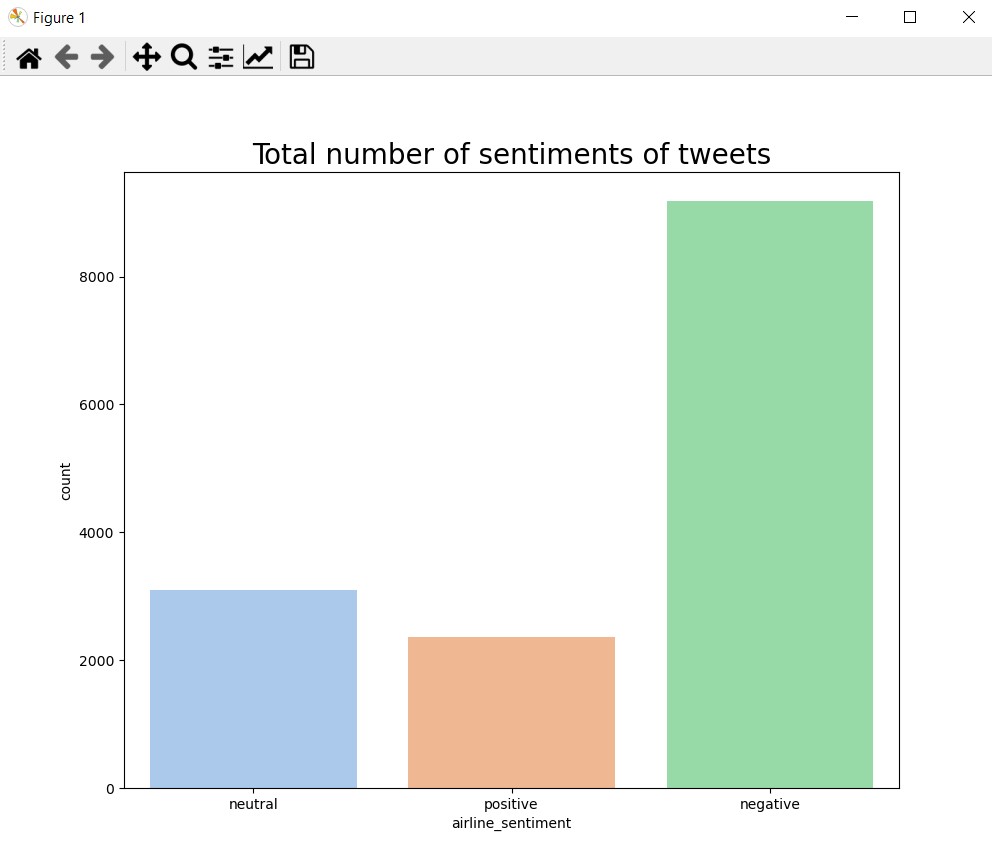
After pre-processing the data is then sent to TFIDF vectorizer which vectorizes the text into a matrix which contains the TFIDF weights of each word in the data frame. This is important because they cannot understand raw text so the text needs to be converted into vectors.

**Training and testing:**

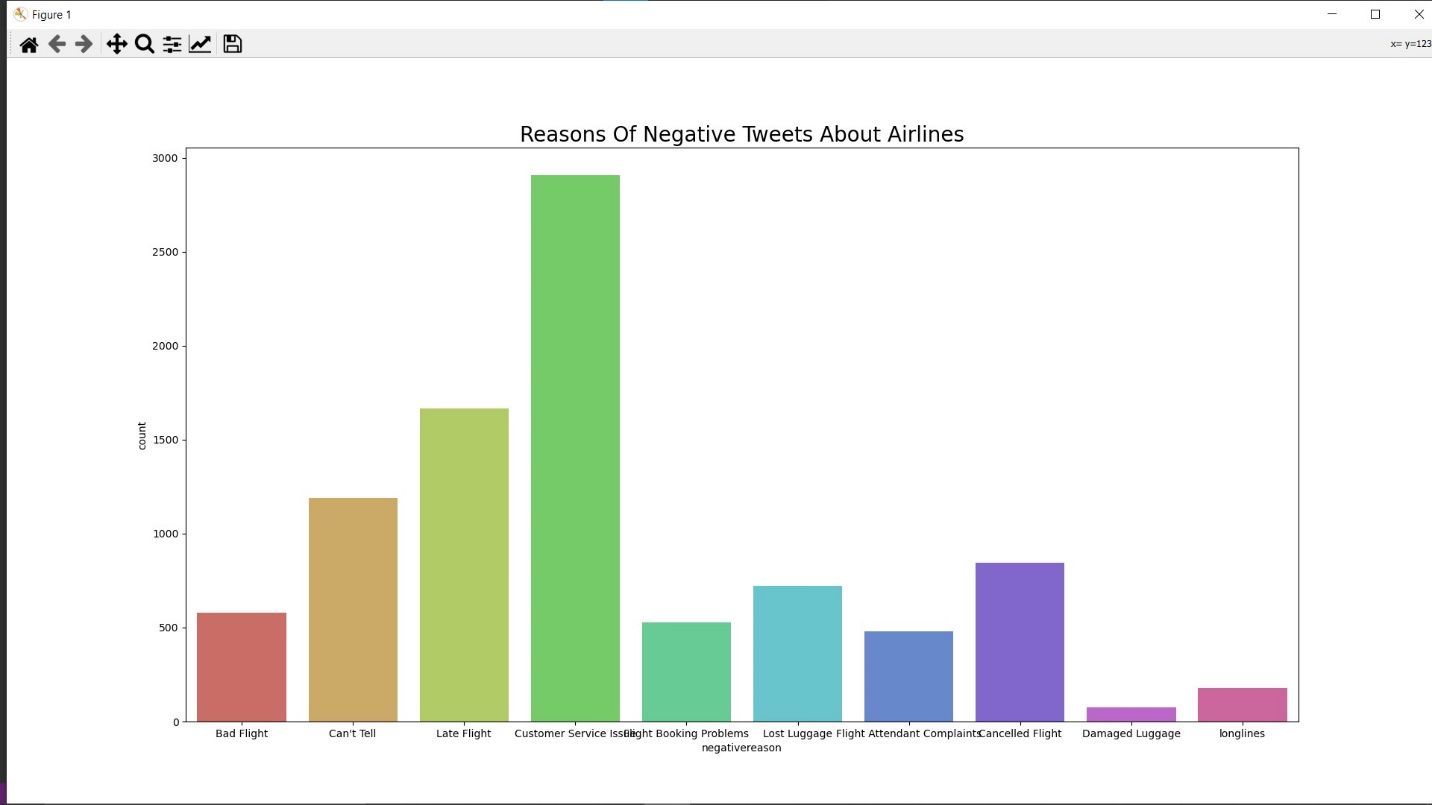
The vectorized data from above, and sentiments are sent to the classifier (in our case Naives Bayes) which reads the data and trains the model to identify the sentiment of the tweet.Now test data is tested with the classifier to measure it’s performance.

**Experimentation and results:**

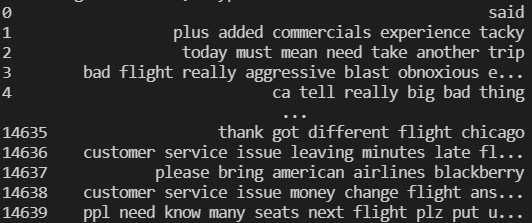
Our Dataset:



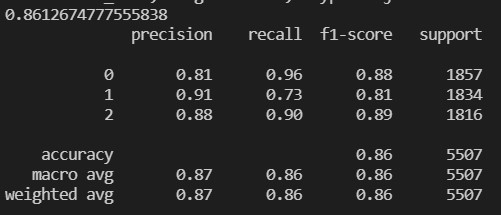
Representation of negative reasons mentioned in the dataset:



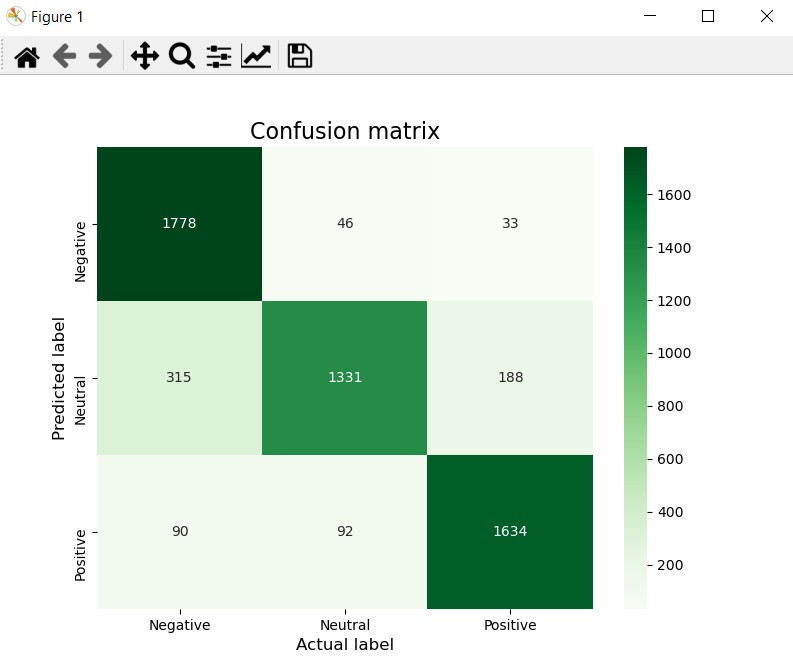
Pre-processed Data:



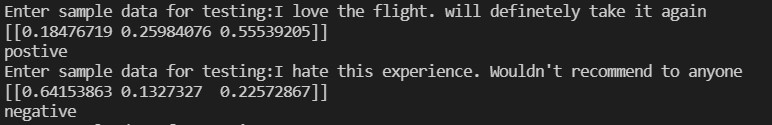
Scores:



Confusion Matrix:



Testing:





The probabilities in printed array represent probability being either negative, positive or neutral review respectively. 1st case is good statement so probability of being positive was very high, 2nd case was a bad statement so probability of being negative was high. In third case, me being a student at Shivnadar is irrelevant to the dataset or the model so it was marked neutral.

The best existing paper we saw had an accuracy of 93% compared to our 86%. But time for execution was reported as 510 seconds. I believe in terms of time complexity our paper is better.

**Conclusion and Limitations:**

In conclusion, our model has an 86% using Naives Bayes classifier and TFIDF vectorizer. In terms of accuracy and time complexity it is relatively good. There is still a 14% chance our prediction might be wrong. It can be improved using techniques such as linear regression but they take a lot of time to train and require much more processing power. It can also be improved by adding features such as sarcasm detection, All the above-mentioned add-ons will require higher times for execution.

Sentiment analysis is very broad and it will help companies to recognize the feedback given by the users. It is easy to curate massive amounts of data.

**References:**

1. Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, Sweta Tiwari,"Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.8, No.4, pp.54-62, 2016. DOI: 10.5815/ijieeb.2016.04.07
2. Yohanssen Pratama et al 2019 J. Phys.: Conf. Ser. 1175 012102
3. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8567243>
4. M.Govindarajan , December 2013, Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm
5. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6623713>
6. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7011523>