Twitter Sentiment Analysis

Group #6

Alekha Polavarapu 811189207 apolava1@kent.edu

Vanaja Karpurapu 811191666 vkarpura@kent.edu

Priyanka Yakkateela 811190147 pyakkate@kent.edu

# Abstract:

# "Sentiment Analysis" generally refers to the procedure of identifying and tagging the text's emotional tone. Tweets can provide a wealth of information about public sentiment when evaluated. The measurement of public sentiment is necessary for many purposes, such as company attempts to evaluate market reaction to their products, political outcome prediction, and the analysis of socioeconomic events like stock exchange. We may discover a lot about what the general public thinks and feels about a variety of topics. The major issue that needs to be identified is how the public perceives a specific topic on a social media platform like twitter. Therefore, the difficulties in resolving this issue will be addressed in our proposal. As a result, we will use machine learning to construct a sentiment analysis model in our project and solve the problems associated with solving this issue. The Twitter API was used to obtain tweets, and a method for sentiment analysis was created. This study provides a novel perspective on general sentiment patterns by carefully analyzing tweets from a range of demographic groups.

# Introduction:

# Nowadays, people's behavior has changed as a result of the Internet of how they communicate their ideas and beliefs. Today, the main method is online forums, product review sites, blogs, social news, etc. Millions of people use social media today. Social media networks such as Facebook, Twitter, Google Plus, etc. to convey their feelings, opinions, and viewpoints about daily routines. The online communities provide us with an interactive media that let users share information and influence others by way of forums. The amount of data generated by social media is Tweets, status updates, and blogs that provide emotional data postings, remarks, evaluations, etc. Additionally, social media gives businesses a chance by providing a forum for connecting with consumers for marketing.

# The field of computer science, artificial intelligence, and linguistics known as "natural language processing" (NLP) is focused with developing computational models that can process and comprehend natural language. These include text to speech, language translation, teaching the computer to recognize the semantic grouping of words (e.g., cat and dog are more semantically similar than cat and spoon). Sentiment analysis uses text analysis techniques to interpret and categorize emotions (positive, negative, and neutral) in text data. Organizations can determine how the general public feels about particular terms or topics by using sentiment analysis. Given that sentiments are the most fundamental characteristics to judge human behavior, sentiment analysis is the area that manages decisions, reactions, and emotions that are generated through messages. It is widely used in fields like social media analytics, web analytics, and data mining analytics. This particular topic is experiencing growth in both academic research and mechanical social structures. Positive, negative, neutral, or restrained quantitative scores can be used to determine whether a sentiment is appropriate.

Sentiments may be conveyed by calculating people's discernment of a certain point, approach, and disturbance to a unit, where this unit may be an event or a topic. Sentiment analysis is based on drawing conclusions, defining the attitude they convey, and finally classifying them into categories. Surveys are first collected, their sentiment is then perceived, their highlights are selected, their sentiments are then arranged, and finally, their sentiment polarization is decided or determined. While managing sentiment analysis, searching the appropriate dataset is of utmost importance. It can be useful for evaluating products for business, determining the highs and lows of financial exchanges, understanding the mindset of people reading the topics, and also understanding the viewpoints expressed by people in political dialogues.

# Problem statement:

# Textual information retrieval strategies are primarily concerned with processing, searching, or evaluating the available factual data. Although facts have an objective component, there are several other textual elements expressing subjective attributes. The majority of the contents are opinions, sentiments, assessments, and attitudes. Sentiment Analysis is based on feelings and emotions. It provides numerous challenging prospects for innovative product development. Applications, owing mostly to the massive increase in available data from internet sources such as blogs and social networks. For example, suggestions for things offered by a recommendation system can be anticipated by taking into account factors such as positive or negative opinions using sentiment analysis to learn more about those items. The aim of this project is to create a social media evaluation and analysis. Here we are considering text data extracted from twitter API which contains sentiments positive and negative (which is Twitter Sentiment Analysis).

# Data collection:

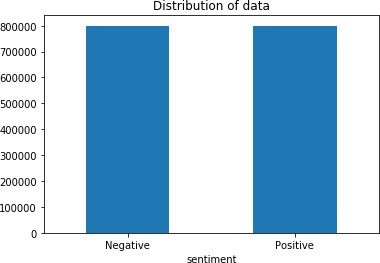
The potential methods for gathering data have been examined. Although there are only a few big sources for databases, there are numerous open source databases that are freely available to the general public. As a result, we decided to use the public open source database to meet the needs of our project because it is the greatest thing we could find.

We collected our dataset from open source data platform Kaggle. Link: <https://www.kaggle.com/datasets/kazanova/sentiment140>

It contains the following 6 fields:

* **target**: the polarity of the tweet (*0* = negative, *2* = neutral, *4* = positive)
* **ids**: The id of the tweet ( 3076)
* **date**: the date of the tweet (*Sat May 16 23:58:44 UTC 2009*)
* **flag**: The query (*lyx*). If there is no query, then this value is NO\_QUERY.
* **user**: the user who tweeted (*robotickilldozr*)
* **text**: the text of the tweet (*Lyx is cool*)

We ignore the other fields since we only need the sentiment and text fields. So we have removed those unwanted columns.



# Data Preprocessing and Analysis:

# Preprocessing of data is a crucial stage in Natural Language Processing (NLP) operations. It converts data into a more consumable format, allowing machine learning algorithms to perform better.

# Below are the steps followed for the preprocessing of data:

# Lower Casing: Each text is converted to lower case.

# 

# URLs Replacement: "URL" text is replaced with links that begin with "http," "https," or “www”.

# 

# Replacement of emojis: Emojis can be replaced with a pre-defined lexicon of emojis and their meanings.

# 

# Replacement of Usernames: Replace @Usernames with word “USER”.

# 

# Removal of Non Alphabets: Using a space to replace all characters except Digits and Alphabets.

# 

# Removal of consecutive letters: 3 or more consecutive letters are replaced by 2 letters.

# 

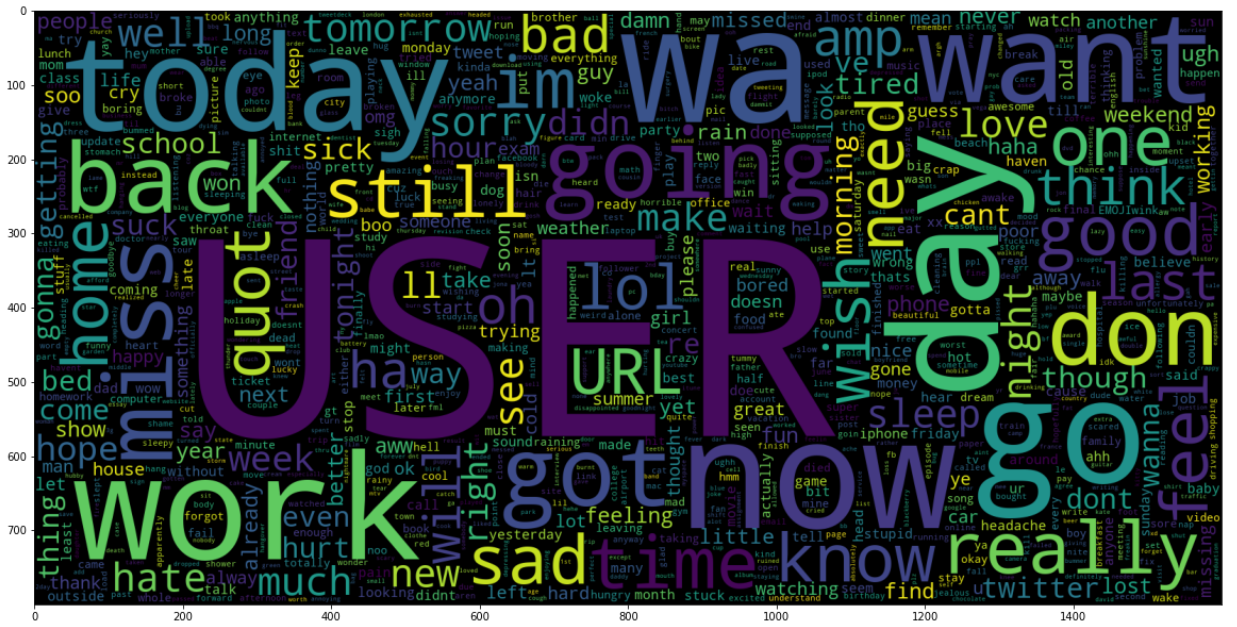
# Removal of Stop words: Stop words are English words that contribute little meaning to a statement. They can be safely ignored without affecting the sentence's meaning. (For example, "the," "he," or "have")

# Lemmatizing of text: The process of reducing a word to its base form is known as lemmatization (For example, "Great" to "Good").

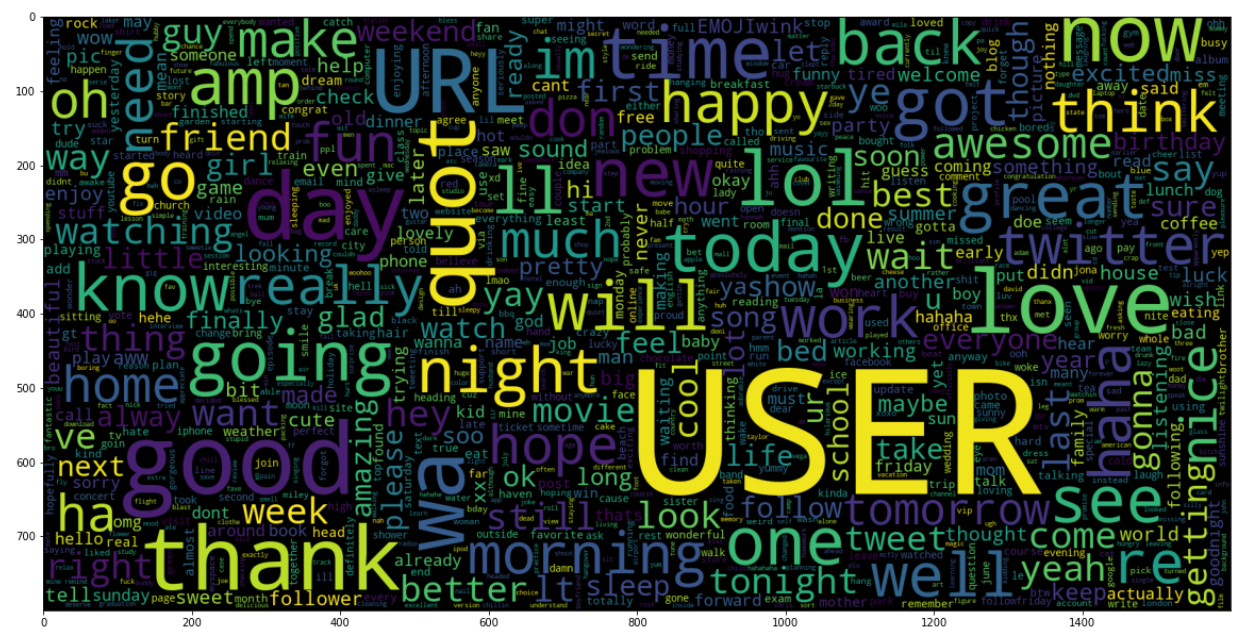
# 

To gain a grasp of the preprocessed data, we will be analyzing it. We'll create Word Clouds from our dataset of Positive and Negative tweets to discover which words appear the most frequently.

Word Cloud for negative tweet words:

****

Word Cloud for positive tweet words:

****

**Methodology:**

For this project, we aimed to create a social media evaluation and analysis. After acquiring information from the relevant sources, data preparation and attribute selection take place. After then, visualization strategies are used to interpret the data. Finally, we use TF-IDF, Naive Bernoulli Bayes, Linear Support Vector Classification, and Logistic Regression(LR).

## TF-IDF (Term Frequency-Inverse Document Frequency) Vectoriser:

Term-frequency is the number of times a term appears in a given document. When a specific "word" appears numerous times in a specific document, the text is deemed relevant for that word (or query), and the frequency value is known as term-frequency.

The issue with employing this term-frequency value alone is that some unimportant words, such as "the," "and," "or," and others, occur frequently in English text documents and are given more weight in terms of term-frequency value while being of little relevance in the context of a sentence or paragraph.

TF(t) = (Number of times term t appears in a document) (Total number of terms in the document)

On the other hand, inverse document frequency (IDF) examines if a search term appears in any documents. A term receives a higher IDF value if it only appears in a small number of documents, while a word that appears in most papers (i.e., is irrelevant) receives a lower IDF value.

IDF(t) = (Total number of document) (Number of documents with wordt in it)

In this method, infrequently occurring valuable words receive some spotlight, whereas frequently occurring useless terms suffer from an inverse document frequency value penalty. Thus, it addresses the problem of often occurring irrelevant terms like "the," albeit it is still not perfect because it does not order documents according to the frequency of a given search term. In other words, it is unconcerned with how frequently a word appears in a manuscript.

TF−IDF=TF∗IDF

**Implementation:**

The TF-IDF identifies a word's significance for comprehending a document or dataset. Let's use an example to better understand. Imagine of having a dataset with essays that students have written on the subject of "My House." The word "a" appears frequently in this dataset compared to other terms; it has a high frequency. Other words that appear less frequently in the dataset, such as home, house, rooms, and so forth, have a lower frequency and more informational content than the word itself. This is the underlying theory of TF-IDF.

A matrix of TF-IDF features is created by TF-IDF Vector from a set of input documents. Typically, only the X train dataset is used to train the vector. The range of words in a sequence is known as the ngram range. [For example, when you have an n-gram range of (1, 2), "very expensive" is a 2-gram that is considered as an extra characteristic separate from "very" and "expensive"] max features indicate how many features should be taken into account. For our work with sentiment analysis, we are developing three different types of models:

## Naive Bernoulli Bayes (BernoulliNB):

One more beneficial naive Bayes approach is the Bernoulli model. The model's underlying premise is that its features are binary (0s and 1s) in nature. Text categorization using the "bag of words" model is a Bernoulli Naive Bayes application. One of the Naive Bayes algorithms used in machine learning is the Bernoulli Naive Bayes variant. When the dataset is distributed in a binary way, with the output label either present or not, it is very helpful to apply this method. The key benefit of this approach is that only binary values for characteristics, like:

* + True or False
  + Spam or Ham
  + Yes or No
  + 0 or 1

## Linear Support Vector Classification (LinearSVC):

Linear SVC is the machine learning approach that is most suitable. This is mostly a method for sorting linear problems. A LinearSVC's (Support Vector Classifier) purpose is to correct the data and return the best hyperplane for data classification. For prophesy following the hyperplane, we must employ some characteristics. This enables the use of the particular methods and techniques in any manner. LinearSVC uses a linear kernel to execute SVC in a flexible manner.

Since there are an equal number of positive and negative predictions in our dataset, it is not skewed. Accuracy will serve as our evaluation metric. To further assess how well our model is doing on both categorization classes, we are charting the Confusion Matrix.

## Logistic Regression (LR):

In mathematics, the Logistic Regression Model (or Logit Regression) is used to simulate the likelihood of a particular class of event circumstances, such as passing or falling, winning or losing, alive or dead, healthy or unhealthy. This can also be used to mimic many tasks, such as determining when a picture of a cat, dog, lion, etc. is present. In its simplest version, the logistic regression model is mathematical and represents a logistic equation for a binary dependent variable model. However, there is also a more intricate expansion.

Regression analysis uses logistic regression, often known as logistic regression, which is a binary regression variant of the approximate logistic model parameter. A consistent

variable is represented by the same variable with two potential values in a binary logistics model. Values with the labels "0" and "1" designating two values, for as pass or fail.

Regression methods such as **binary or binomia**l deal with situations where there are only two possible outcomes forms of a dependent variable's observable effect are "0" and "1" (which stand for, for example, "living" or "dead" and "win" or "loss").

**Multinomial logistic regression** addresses situations where there are only three or more possible forms. These outcomes, which include "stage A" vs. "stage B" vs. "stage C," cannot be attained in any particular order.

**Ordinal logistic regression** works with always-ordered conditional variables.

The operational Regression understands the vector of variables, determines the coefficient for the input expressions, and then traces the results of the text class as a word vector.

Multiple linear functions are determined by the logistic regression function. As a part of building sentiment classifier using logistic regression, we train the model on twitter sample dataset. The dataset available is in its natural human format of tweets, which is not so easy for a model to understand. Thus we will have to do some data pre-processing and cleaning to break down the given text into an easily understood format for the model.

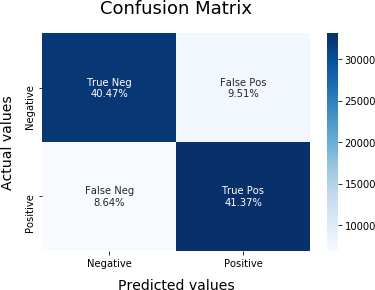
## BernoulliNB Model:

## 

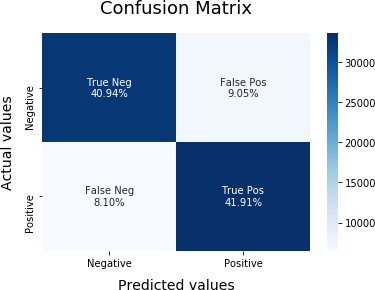
## 

## 

**LinearSVC Model:**

****

## Logistic Regression Model:



* It is obvious that, of all the models we tested, the Logistic Regression Model performs the best. When classifying the sentiment of a tweet, it obtains accuracy of about 82%.
* The BernoulliNB Model, although, is the quickest to train and predict on, it should also be mentioned. Additionally, it has 80% accuracy while calculating.
* Vectoriser and BernoulliNB, Logistic Regression Model are being saved for subsequent usage using PICKLE. We must import the Vectoriser and LR Model using Pickle in order to use the model for Sentiment Prediction. Data can be converted to a TF-IDF Features matrix using the vectoriser. While it is possible to forecast the sentiment of the converted Data using the model. However, pre-processing is necessary for the text whose sentiment needs to be anticipated.

**Challenges faced:**

* We faced difficulty while importing some of the libraries of required versions. Resolving those issues took us some time.
* Handling and working with this large amount of data took us more time.

**References:**

1. **Sentiment Analysis using Machine Learning and Deep Learning**

<https://ieeexplore.ieee.org/document/9083703>

1. ***Sentiment Analysis on Social Media for Emotion Classification***

<https://ieeexplore.ieee.org/document/9121057>

1. **Sentiment Analysis on Social Media**

[https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6425642](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp&arnumber=6425642)

1. **A Recommendation Method for Social Media Users based on a Sentiment Analysis Model** [**https://ieeexplore.ieee.org/document/9754863**](https://ieeexplore.ieee.org/document/9754863)
2. **Deep Learning Framework Study for Twitter Sentiment Analysis** [**https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9742522**](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp&arnumber=9742522)

**6,Sentiment Analysis of Twitter Data** [**https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8716464**](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp&arnumber=8716464)

# 