Clustering and Opinion Mining on Tweets

A Project Report Submitted in Partial Fulfillment of theRequirements
for the degree of
Master of Computer Application
Of
Department of Computer Science and Engineering
Of
Jadavpur University

by,

Sougata Banerjee
Master of Computer Application – III
Roll Number: MCA226027
Registration Number: 149889 of 2019 – 2020

Under the supervision of

Dr. DIGANTA SAHA
Professor

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Jadavpur University, Kolkata –700032
India
June, 2022

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

I hereby recommend the project report entitled "**Clustering and Opinion mining on Tweets**" prepared by **Sougata Banerjee** (Reg. No: 149889 of 2019-2020, Roll No: MCA226027) under my supervision and be accepted in partial fulfilment of the requirement for the degree of **Master of Computer Application** in the **DEPARTMENT OF COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY** during the academic year 2021-2022.

Prof. Dr. Diganta Saha
Project Supervisor
Dept. of Comp. Sc & Engineering
Jadavpur university -700032

Prof. Dr. Anupam Sinha
Head of the Department
Dept. of Comp. Sc & Engineering
Jadavpur university -700032

Prof. Dr. Chandan Majumder
Dean, Faculty Council of Engineering & Technology
Jadavpur university -700032

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

This is to certify that the project entitled "**Clustering and Opinion mining on Tweets**" is a bona-fide record of work carried out by Sougata Banerjee in partial fulfilment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University during the period of January 2022 to June 2022.It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project for which it has been submitted.

_____

Signature of Examiner 1

Date:

_____

Signature of Examiner 2

Date:

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

This is to certify that the work in the project entitled "**Clustering and Opinion mining on Tweets**" submitted by **Sougata Banerjee** is a record of an original research work carried out by him under the supervision and guidance of **Prof. (Dr.) DIGANTA SAHA** for the awardof the degree of **Master of Computer Application** in the **Department of Computer Science and Engineering, Jadavpur University**, Kolkata-32.Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

NAME	**:**	**SOUGATA BANERJEE**
Examination Roll Number	**:**	MCA226027
Registration Number	**:**	149889 of 2019 - 2020
Project Title	**:**	**Clustering and Opinion mining on Tweets**
Signature	**:**	

ACKNOWLDEGEMENT

With my most sincere and gratitude, I would like to thank **Prof. (Dr.) Diganta Saha,** Department of Computer Science & Engineering,my supervisor, for his overwhelming support throughout the duration of theproject. His motivation always gave me the required inputs and momentum tocontinue with my work, without which the project work would not have come to current shape. His valuable suggestion and numerous discussions havealways inspired new ways of thinking. I feel deeply honored that I got thisopportunity to work under him.

I would like to express my sincere thanks to all my teachers for providingsound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of ComputerScience & Engineering of Jadavpur University for their continuous support.

Date: 27 / 06 / 2022

Sougata Banerjee

Master of Computer Application – III

Roll No. – MCA226027

Registration No:149889 of 2019 – 2020

Abstract

Faculty of Engineering and Technology, Jadavpur University

Department of Computer Science and Engineering

Master Of Computer Application

By Sougata Banerjee

Social media has become part and parcel of our lives in recent years and our involvement in social media is increasing rapidly. Especially in pandemic people became more tend to give their opinion in social media than in real world.

Also, when there is a war going on which has its affect every corner of the world;

It is important to know what people thinks about it and how war is impacting people.

Here we worked with real data retrieved from twitter and performed opinion mining.

Contents

USED FIGURES AND THEIR DESCRIPTION Page No

Abbreviations

1. **SA- sentiment analysis**
2. **ML-Machine Learning**
3. **NLP -Natural Language Processing**
4. **PoS- parts of speech tagging**
5. **API - Application Programming Interface**

Chapter 1 Introduction: -

Social media and its corresponding applications allow billions of users to express their opinion about a topic and show their sentiment/perspective by liking or disliking content. All these actions continuously gather textual data. This data refers to a massive set of opinions that can be processed, analyzed to obtain people tendencies in digital world. There are two types of such data – Objective texts and Subjective Texts. For past few decades Twitter have been an ideal platform to get an insight of trending topics that is what is going on worldwide at every moment and what does people feel about that topic. It depends on how people are reacting to a certain topic – generally people like to express their opinion by tweets about a topic by using the hashtag trending for that topic.

By doing Sentimental Analysis (SA) we can determine what do people have to say – is most of their opinions positive, negative or neutral.

But the data i.e. the tweets are not ready as NLP Data; as they consist of slangs, abbreviation, emoticons, sarcasm, hyperlinks, etc. These parameters can be hindrances while analyzing sentiment of a message. With help of NLP, we will try to clear noise from out data as much as possible and thereby classify them according to their sentiment of **social media**.

Sentiment Analysis (SA) is an intellectual process of extricating user's feeling and emotions. It is one of the pursued fields of NLP. SA – often termed as opinion mining or polarity detection, refers to set of AI algorithms and techniques used to extract the polarity of a given Dataset.

A lot of research work has already been done and still being held in the field of Sentiment analysis due to its relevance in real world problem such as- marketing strategies.

It is considered that SA started in early 2000's with the **article published by Bo Pang and Lilian Lee** and **by Peter Turney,** in the text subjectivity analysis performed by the computational linguistics community in 1990's [1]. However, the outbreak of computer-based sentiment analysis only occurred with the availability of subjective texts on the Web. Consequently, 99% of the papers have been published after 2004.In recent past companies like IBM, Twitter, Intel is using sentiment analysis software to understand customer needs, employee feedbacks-such as how likely is someone to stay in the job. SA is also used to shape company's sales plans, improve brand strength, translate digital PR into tangible actions. For instance, Intel uses

product from **Kanjoya Inc.** that uses language processing and ML algorithms to decipher emotion from a text. Besides all these, SA is used in health care-by improving response to specific emergency, understanding stock sentiment, Call Centers- to improve service, Banks such as–

- Understanding Customer Attitudes
- Equity Investing
- Monitoring Credit Markets
- Compliance Monitoring

So, it is quite evident that Sentiment Analysis aka Opinion mining is at its peak right now- and it is going to grow exponentially. In future sentiment analysis is going to continue to dig deeper, far past the surface of the number of likes, comments and shares, and aim to reach, and truly understand, the significance of social media interactions and what they tell us about the consumers behind the screens. This forecast also predicts broader applications for sentiment analysis – brands will continue to leverage this tool, but so will individuals in the public eye, governments, nonprofits, education centers and many other organizations.

Though SA has been revolting so much, researchers still face difficulties in opinion mining such as – irony and sarcasm, types of negations, word ambiguity, multilingual sentiment analysis, multipolarity etc. Hence there is always room for acquiring more accuracy using suitable techniques (preferably hybrid techniques) [5]. We are going to discuss all the nitty gritty of SA in detail in Chapter 3.1.

Here We intend to introduce some concepts of Sentiment Analysis, NLP, ML algorithms and some other parameters related to our work. In Chapter 3 under title “Theoretical Concepts” and “Methodology” we will describe respectively all the above-mentioned concepts and also the tools and libraries that are used in the proposed work. In our Work, our goal is to understand people’s feeling or sentiment – obtained from numerous tweets about an issue using Twitter API, thereby clustering our data into several clusters and then obtain polarity of each Cluster. Sentiment Analysis in this work is done in Unsupervised Learning Method.

Chapter 2 Related works:

Sentiment analysis is collection of methods and tools about detecting and extracting subjective information, such as opinion and attitudes, from human communication data which is language.[2]

Though first research paper was published in 1937, the number of research paper on sentiment analysis and opinion mining is exponentially increasing in recent years. It is noted that almost 7000 papers are published on this paper and 99% of them are after 2004.

The following figure shows the trend

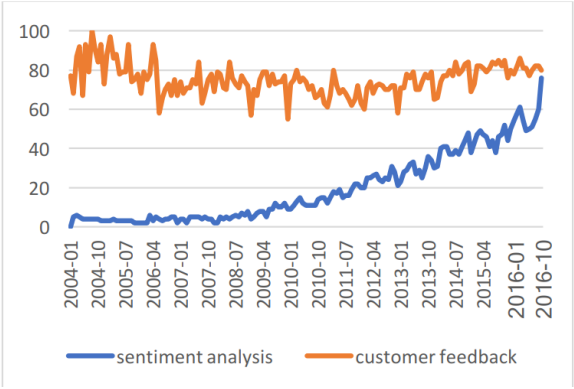

FIG:1

Some recent papers in this subject are listed below:[3]

1. In 2019, Saad and Yang have aimed for giving a complete tweet sentiment analysis on the basis of ordinal regression with machine learning algorithms. The suggested model included pre-processing tweets as first step and with the feature extraction model, an effective feature was generated.
2. In 2020, Kumar *et al*. have presented a hybrid deep learning approach named ConVNetSVMBoVW that dealt with the real-time data for predicting the fine-grained sentiment. In order to measure the hybrid polarity, an aggregation model was developed.
3. In 2019, Abdi *et al*. have suggested a deep-learning-based technique for categorizing the opinion of the user mentioned in reviews. Moreover, a deep learning model was a unified feature set that was representative of sentiment shifter rules, word embedding, sentiment knowledge.
4. In 2019, Ray and Chakrabarti introduced a deep learning algorithm for extracting the features from text and the user's sentiment analysis with respect to the feature. In opinionated sentences, a seven-layer Deep CNN was employed for tagging the features. In order to enhance the

performance of sentiment scoring and feature extraction models, the authors merged the deep learning methods using a set of rule-based models. Finally, it was seen that the suggested method achieved best accuracy.

CHALLENGES OF SENTIMENT ANALYSIS There are different challenges in sentiment analysis which is describe below.[4]

5.1 Implicit Sentiment and Sarcasm

There is a chance that a sentence may contain implicit sentiment even though it is not having any word that earns sentiments. For an example, two statements are taken; "One should question that the stability of mind of the writer who wrote this book." In this above sentence don't have negative sentiment bearing words and no negative words are seen, although both are negative sentences. Or this sentence " His acting is so good that this movie will be his first and last."Thus, identifying sentiment is important in Sentiment Analysis than syntax detection.

5.2 Domain Dependency

In this type of challenge words polarity changes from one domain to another domain in the domain dependency. For an example, two statements; "The story was unpredictable." and "The steering of car is unpredictable." In first statement, in Sentiment express that is positive whereas the second statement express sentiment is negative.

5.3 Language Problem in Opinion Mining

English language is mostly used because of its resource's availability means lexicons, dictionaries and corpora but User get attracted by using Opinion mining with German, Arabic ,etc. i.e., lexicons dictionaries and corpora for these languages.

5.4 Fake Opinion

Fake opinion is also called fake review and refers to bogus or fake reviews. The fake opinion is misguiding the users or readers by providing them untruthful positive or negative opinion related with any object. This is social challenge which is faces by OP.

Main motivation of this work was to know about the sentiment of people regarding Russia-Ukraine war.

Chapter 3 Methodology

Chapter 3.1.1 Natural Language Processing (NLP)

- *What is Natural Language Processing?*

There is a lot of information in the world about raw text- generally natural language. The term Natural Language is not limited to any particular language rather it covers any language humans speak in. Natural Language has many forms such as: -

- Text messages, emails
- Speech to Siri, Alexa
- Signs and gestures and many more.

As there is huge unstructured data of Natural Language, there must be some algorithms so that Computer can understand and reason with this data and extract accurate information from it.

NLP is a subfield of Machine Learning that works with interaction between machine and human language. The dataset of NLP can be of different forms; for instance – text from social media, Customer review/feedback about a product in an Online marketing site, web pages.

In 21'st Century NLP is present all the time in our daily life from getting an answer from google over speech-to-text technology to auto generated weather forecast scripts.

- *History Of NLP & why NLP was needed?*

Before getting into further details of NLP, we should know how and why NLP was introduced to the mankind.

In 1950, famous mathematician **Alan Turing** proposed in an article named "*Computing Machinery and Intelligence*" a test which is now famous with the name **Turing Test,** as a criterion to call a machine intelligence. The primary goal was to observe whether a computer can suffice mimicking a human and can carry on a conversation with other human -and till date no machine is able to do it; but 1950 is the time period that is considered to be starting period of NLP.

In 1954, scientist claimed in **Georgetown-IBM [7]** experiment that machine translation will be solved within three to five years. Though a machine translation system that can beat human efficacy has not been invented yet, machine translation has been evolving tremendously since concepts like Deep Learning is introduced.

Some applications are – social media-Facebook open sourced a Neural Machine Translation [8], Real Time conversation (Skype, FaceTime), Google translation.

But the hot topic among all these is **Chatbots -which completely changed the customer care and the way business contacts with clients.**

For instance, in 2016 **Microsoft's AI Chatbot-Tay** [9] was launched to mimic a teenage girl who converses with Twitter users in real time. She was programmed to learn from other users' tweets and comments on Twitter. But she was overwhelmed with tweet-trolls and started to post inappropriate things by learning other users' behaviour and she was terminated within 24 hours.

Now coming to the point why NLP was needed – we can see that machine is quite efficient reading data from spreadsheets with the help of Machine Learning just like humans read. But when it comes to unstructured data -such as texts (a huge form of data) in which human communicate, news articles, web page content, speech in languages other than English; machine cannot understand the words as well as humans do. So, introduction of NLP was quite inevitable to make machine smarter.

Chapter 3.1.2 Common tasks & techniques of NLP

NLP draws from several disciplines, including computational linguistics and computer science, as it attempts to close the gap between human and computer communications. Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the following:

- **Syntactic analysis**, also known as parsing or syntax analysis, identifies the syntactic structure of a text and the dependency relationships between words, represented on a diagram called a parse tree.[10]
- **Semantic analysis** focuses on identifying the meaning of language. However, since language is polysemic and ambiguous, semantics is considered one of the most challenging areas in NLP.[11]

Semantic tasks analyse the structure of sentences, word interactions, and related concepts, in an attempt to discover the meaning of words, as well as understand the topic of a text.[12]

some of the main sub-tasks of both semantic and syntactic analysis:

3.1.2.1 Text Analytics:

Text Analytics involves the use of unstructured text data, processing them into usable structured data. Text Analytics is an interesting application of Natural Language Processing. Text Analytics has various processes including cleaning of text, removing stopwords, word frequency calculation, and the list goes on.

Text Analytics has gained much importance these days. As millions of people engage in online platforms and communicate with each other, a large amount of text data is generated. Text data can be blogs, social media posts, tweets, product reviews, surveys, forum discussions, and much more. Such huge amounts of data create huge text data for organizations to use. Most of the text data available are unstructured and scattered. Text analytics is used to gather and process this vast amount of information to gain insights.

Text Analytics serves as the foundation of many advanced NLP tasks like Classification, Categorization, Sentiment Analysis, and much more. Text Analytics is used to understand patterns and trends

in text data. Keywords, topics and important features of Text are found using Text Analytics.[13]

3.1.2.2 Parts-of Speech-Tagging (POS) –

PoS involves adding a part of speech category to each token within a text. Some common PoS tags- are *verb*, *adjective*, *noun*, *pronoun*, *conjunction*, *preposition*, *interjection*, among others. For example-

Part Of Speech	examples
Noun	John , Canada
Pronoun	They,her,him
Adjective	Beautiful,Awesome
Interjection	Phew,oops

3.1.2.3 Tokenization -

Given a sentence Tokenization is an essential task in natural language processing to breaking the text into fragments -called *tokens*. But First certain Characters are removed such as punctuations, emojis, slangs, URLs, digits.

Sentence tokenization splits sentences within a text, and word tokenization splits words within a sentence. Generally, word tokens are separated by blank spaces, and sentence tokens by stops.

Tokens composed of one word is known as unigrams, bigrams are of two consecutive words and n-grams are of n consecutive words.

Learning new things is awesome.

Learning new	things is	awesome

Above-mentioned is example of bigram.

3.1.2.4 Name Entity Recognition (NER) - Named entities are phrases of different categories -such as person, place, company date. NER is an important subtask of extraction of information and identify such entities. e.g- Identify John as Name of Person and Canada as place.

3.1.2.5 Sentiment Analysis –

It's a type of text classification where the NLP algorithms determine the text's positive, negative, or neutral connotation. Use cases include analysing customers' feedback, detecting trends, conducting market research, etc., via an analysis of tweets, posts, reviews and other reactions. Sentiment analysis can encompass everything from the release of a new game on the App Store to political speeches and regulation changes.

Example:-

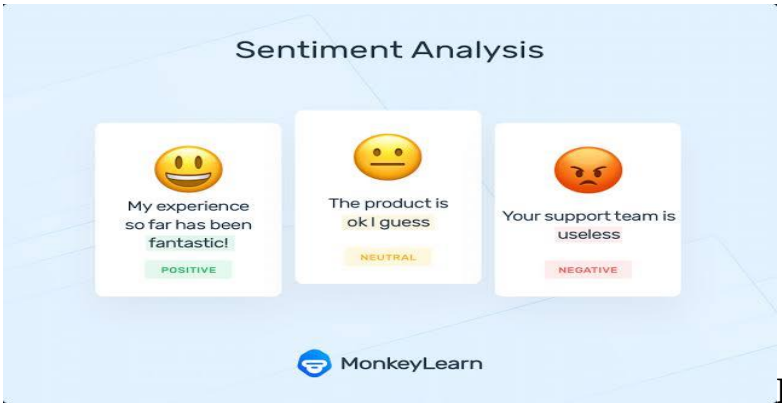

FIG 2

3.1.2.6 **Stop words Removal**- the most frequent words in text are common stop words. Generally, stop words do not carry any meaning that contributes to NLP or text analysis; therefore, most of the times Stopwords are removed before analysis or model training.

PoS tags are also an excellent alternative for stopword filters. But we will work with NLP library (refer to chapter 2.2) as almost all the libraries are integrated with Stop word lists in almost every language.

Removal of stop words results in less memory consumption, faster calculations.

Example - Text: - here is an example to show how removing stop words function works.

After removing stop words ; text reduces to

Example show removing stop words function works.

3.1.3 Vectorization

One of the most prominent differences between text data and structured data is that text is represented by words whereas structured data is mostly represented by numbers. As machine learning has advanced a lot, we need to find a mapping of text to numbers. Keeping in mind, that text is very complex in nature it is impossible to represent meaning of a document with a single digit.

The natural extension of real numbers is tuple of real numbers -vectors. Almost all text representations in text analysis use vectors.

The initial step towards making the text documents machine-readable is vectorization. Transforming textual data to meaningful vectors is a way to communicate with the machines for performing any Natural Language Processing tasks and solve problems mathematically. Researchers in the domain had proposed different vectorization models that range from a very simple to sophisticated ways helpful in solving NLP problems.[14]

Here, we will implement two popular models: Bag of Words, TD-IDF implementation.

Bag of words

The idea behind this method is straightforward, though very powerful. First, a fixed length vector is defined where each entry corresponds to a word in a pre-defined dictionary of words. The size of the vector equals the size of the dictionary. Then, for representing a text using this vector, each word is counted how many times they occurred in dictionary appears in the text and this number is put in the corresponding vector entry.

For example, if a dictionary contains the words {Machine, Learning, is, difficult, not, great}, and we want to vectorize the text “Machine Learning is great”, we would have the following vector: (1, 1, 1,0, 0, 1).

TF-IDF:

Unlike Bag of Words model TF-IDF uses a more generic approach by “punishing” words that appear frequently in a text document.TF-IDF takes care

of that by counting the number of total word occurrences. it will reduce the weight of frequent words and at the same time increase weights of uncommon words

Term Frequency-Inverse Document Frequency [15] is the most commonly used method in NLP for converting text documents into matrix representation of vectors. Tf-idf representation reflects the prominence a word in a collection of documents to the individual document.

Chapter 3.1.4 NLP libraries

Some NLP libraries we used in implementation: -

1. NLTK -**NLTK (Natural Language Toolkit)** Library is a suite that contains libraries and programs for statistical language processing. It contains packages to make machines understand human language and reply to it with an appropriate response. NLTK provides tokenization, PoS tagging Stop words, Corpora etc. NLTK provides support to research works in NLP also artificial intelligence, information retrieval, and machine learning. For Years NLTK is being used as a tool for prototyping and building research systems.

2. spaCy- **spaCy** is more powerful toolkit than NLTK.SpaCy is written in Cython which is more memory optimize, So it is faster than NLTK. This is a free and open-source library for NLP in Python with a lot of in-built capabilities. Unstructured textual data is produced at a large scale, and it's important to process and derive insights from unstructured data.

   spaCy features convolutional neural network models for part-of-speech tagging, text categorization and named entity recognition. It also provides support for tokenization, stopword removal for more than 65 languages.

3. TextBlob – This library is relatively new one built on top of NLTK.It simplifies NLP and text analysis with built in functions and methods. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.[18]

TextBlob provides a Subjectivity lexicon for English adjectives. The polarity are -1.0(negative) +1.0 (positive) 0.0 (neutral). The polarity is used to check if a text has positive sentiment or negative sentiment and subjectivity is used to check if the text is objective or subjective.

4. Gensim – Currently the most popular topic modelling toolkit is Gensim. It is available in Python. This is a library for topic modelling, document indexing, similarity retrieval with large corpora. This library includes streamed parallelized implementations algorithms like word2vec and doc2vec, latent semantic analysis (LSA, LSI, SVD), non-negative matrix factorization (NMF), latent Dirichlet allocation (LDA), tf-idf and random projections.

Chapter 3.1.5 Machine Learning Introduction

What is machine learning?

Machine learning (ML) is a branch of artificial intelligence (AI) that enables computers to "self-learn" from training data and improve over time, without being explicitly programmed. Machine learning algorithms are able to detect patterns in data and learn from them, in order to make their own predictions.

According to Tom Mitchell, A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

The main task of machine learning is to explore and construct algorithms that can learn from historical data and make prediction on new data.

Depending on the nature of learning data; machine learning can be categorized into three categories: -

1. Supervised Learning: - When data comes with description .targets or desired outputs. Goal is to find a rule that maps input to output.

2. Unsupervised Learning- data without description. Refer to chapter 2.1.5.1

3. Reinforcement Learning – Learning data provided feedback so that system adapts to dynamic conditions to achieve a certain goal.

Some examples of Machine Learning are NLP, Face recognition, recommendation system, biometric etc.

3.1.5.1 Unsupervised Learning

When Learning data contains only indicative signals without any description attached or labelled, the goal is to find structure of data underneath, to discover hidden information, or to determine how to describe the data.

This kind of Learning data is called **unlabelled data.** Unsupervised Learning can be used to detect anomalies, or to group customers with similar online behaviour for a marketing campaign.

Text analysis is also an example of Unsupervised Learning. Unsupervised Learning is highly rich in the area of NLP. Unsupervised Learning Algorithms comes to play when it is to mining text data.

3.1.5.2 Clustering :K-Means Algorithm

Clustering is an important branch of unsupervised learning, which identifies different groups of observations of data. This process ensures that similar data points are identified and grouped. Clustering algorithms is key in the processing of data and identification of groups.

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

- market segmentation
- social network analysis
- search result grouping
- medical imaging
- image segmentation
- anomaly detection

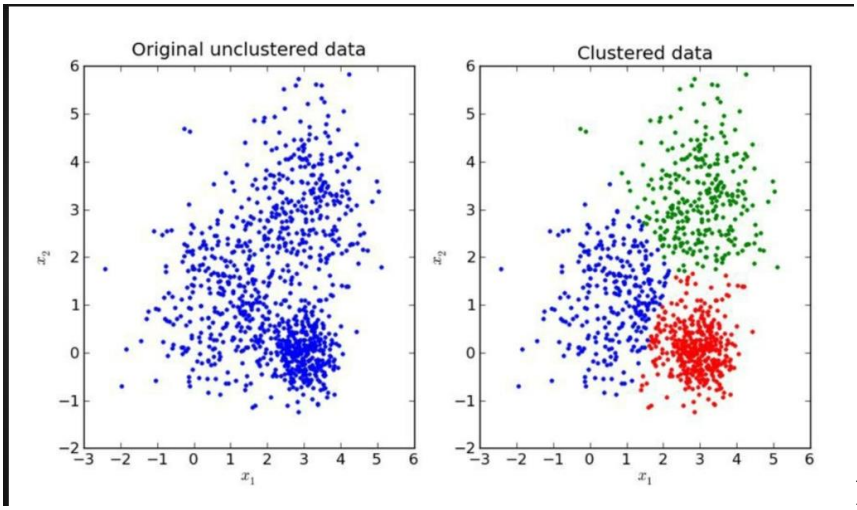

FIG3

K-Means Algorithms-

The goal of k-means algorithm is to partition the data into K groups based on feature similarities k-is predefined (determined by elbow method) property of this model. K being no. of Clusters. Each of K clusters is specified by a centroid and each data sample belongs to the cluster with the nearest centroid. During training , the algorithm iteratively updates the k centroids based on data provided.

Steps of the K-Means algorithm:

1. specifying K: algorithm needs to know how many clusters will be generated at the end.
2. Initializing Centroids: Randomly selects K samples as centroids from the data set.
3. Assigning Clusters: with K centroids samples sharing the same closest centroid creates one cluster. Eventually K clusters are created. This closeness is measured by Euclidean or Manhattan distance.
4. Update: For every Cluster recalculate mean of all samples in the cluster. K centroids are updated to be means of corresponding clusters.
5. Repeat 3 and 4 till the model converges when a small enough update of centroids can be done.

Getting value of K by Elbow Method: -

Initially, the value of K is not known and K -means algorithm needs a specific value of K to start with. So, to determine the value of K there is an approach Elbow Method.

In this method, different values of k are chosen and corresponding models are trained; for every trained model the **sum of squared errors, or SSE (also known as WCSS sum of within cluster distances) of** centroids are calculated and plotted against K. The optimal K is chosen where the marginal drop of SSE starts to decrease rapidly. That implies that further clustering does not provide any substantial gain.

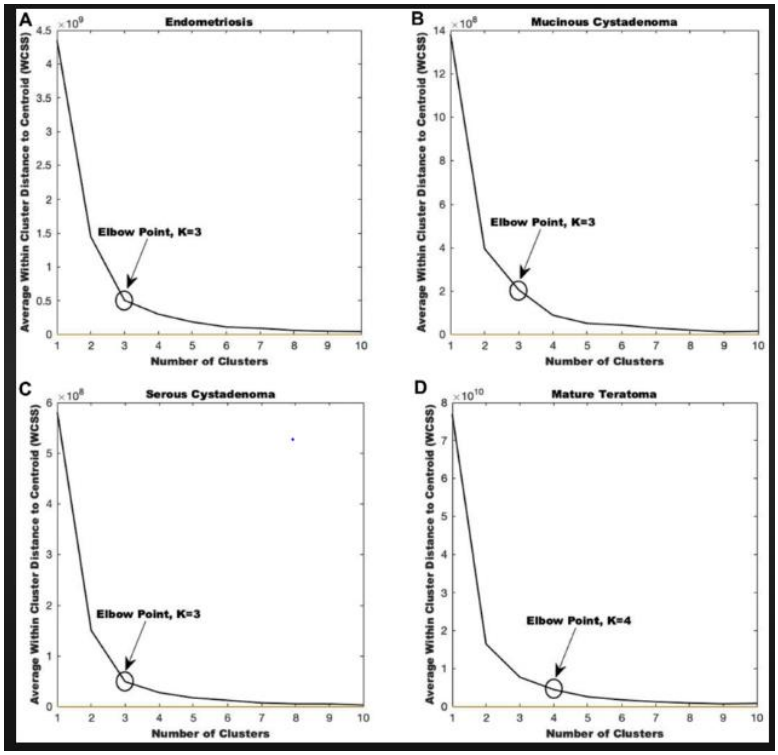

FIG 4

$$WCSS= \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$

WCSS/SSE: $\sum_{Pi \text{ in Cluster1}}$ distance $(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroid, we can use any method such as Euclidean distance or Manhattan distance.

Euclidean Distance	$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
Manhattan Distance	$\lvert x_1\text{-}x_2\rvert + \lvert y_1\text{-}y_2\rvert$

3.1.5.3. Topic Modelling -LDA

Topic modelling is an unsupervised machine learning approach that can scan a series of documents, find word and phrase patterns within them, and automatically cluster word groupings and related expressions that best represent the set.

Because it doesn't require a pre-existing list of tags or training data that has been previously categorised by humans, this type of machine learning is known as 'unsupervised' machine learning.

There are two popular topic modelling algorithms: - non-Negative matrix Factorization (NMF), Latent Dirichlet Allocation (LDA).

LDA:-

DA was developed in 2003 by researchers David Blei, Andrew Ng and Michael Jordan. Its simplicity, intuitive appeal and effectiveness have led to strong support for its use.

LDA is implemented by the Gensim library.

LDA topic modelling discovers topics that are hidden (latent) in a set of text documents.

It does this by inferring possible topics based on the words in the documents. LDA is a generative probabilistic model and Dirichlet distributions that explains each input document by means of mixture of topics with certain probability. **Topic** in topic modelling by means a collection of words with certain connection.

LDA is trained in generative manner where it tries to abstract from the documents a set of hidden topics that are likely to generate a certain collection of words.

Being a probabilistic graphical model LDA takes input data in terms of counts and term matrix derived by tf-idf vectorization is used as input data.

Now fit the LDA model on the term matrix –

```
>>> lda.fit(data)
```

Finally we can derived no. of topics (the no. of topic should be provided) Such as:

Topic 1		Topic 2		Topic 3	
term	weight	term	weight	term	weight
game	0.014	space	0.021	drive	0.021
team	0.011	nasa	0.006	card	0.015
hockey	0.009	earth	0.006	system	0.013
play	0.008	henry	0.005	scsi	0.012
games	0.007	launch	0.004	hard	0.011

FIG 5

Chapter 3.1.6 Dimensionality Reduction using PCA

Dimension reduction is a ML technique that reduces no. of features and also retains as much information as possible. Generally, it is performed by obtaining a set of new principal features.

It has been always been difficult to visualize data of high dimension, especially when data is of 100 or 1000 dimensions. Sometimes, some features in high dimensional data that are related brings redundancy in result. This is one of the major reasons that Dimensionality reduction is important.

Dimensionality reduction is not simply considering only two features from the original space; it is transformation of original feature space to a new space of fewer dimensions. Here Principal Component Analysis (PCA) is implemented for data transformation.

PCA is linear data transformation technique that maps the data in a higher dimensional space to a lower dimensional space where the variance of data is maximized.

Chapter 3.2 Implementation

In this Chapter, we will be discussing all the technicalities of how work was done from how was data collected to what can be the output. Flowchart below describes all the steps that are implemented.

Here we worked with Twitter data that has been retrieved using Twitter API and the keyword or Hashtag that is used throughout the whole process is

#StandWithUkraine [refer chapter 3.2.1 for details**]**

One of various reason of choosing this Hashtag was because this was one of the top trending topics in Twitter a few months ago. Also, the Russia-Ukraine war is one of the most talked topics in a while as the war impacts the whole world directly or indirectly. So, we should know what people are feeling about this war and what is sentiment of those tweets and then clustering those tweets trying to find the nature of clusters. Here K-Means Method is used for Clustering and LDA for topic modeling [refer to Chapter 3.1.5.2-3.1.5.3].

Algorithm :

Input: Social Media posts of different persons about a chosen topic.

Output: Clusters of Similar words that shows similar sentiments.

1. Begin
2. Authenticate all tokens, secret key of Twitter API
3. Creating Twitter API object
4. Input: search item: "hashtag of a chosen subject"
5. Creating data-frame of retrieved tweets.
6. For every tweet remove noise –
    a. punctuation, multiple-spaces, emojis, URLs.
    b. Sentiment analysis of each tweet using clean tweets.
    c. Clean Tweets are tokenized
    d. Stop-words are removed from list of tokens for each tweet
    e. Tokens are vectorized
    f. Dimension reduced and formed clusters.
7. Plotting each Clusters.
8. By topic modelling getting topics
9. Visualize Top-30 Most Relevant Terms for each Topic and also token percentage of Tokens that is present in each token.

Flow-Chart of Algorithm Implemented

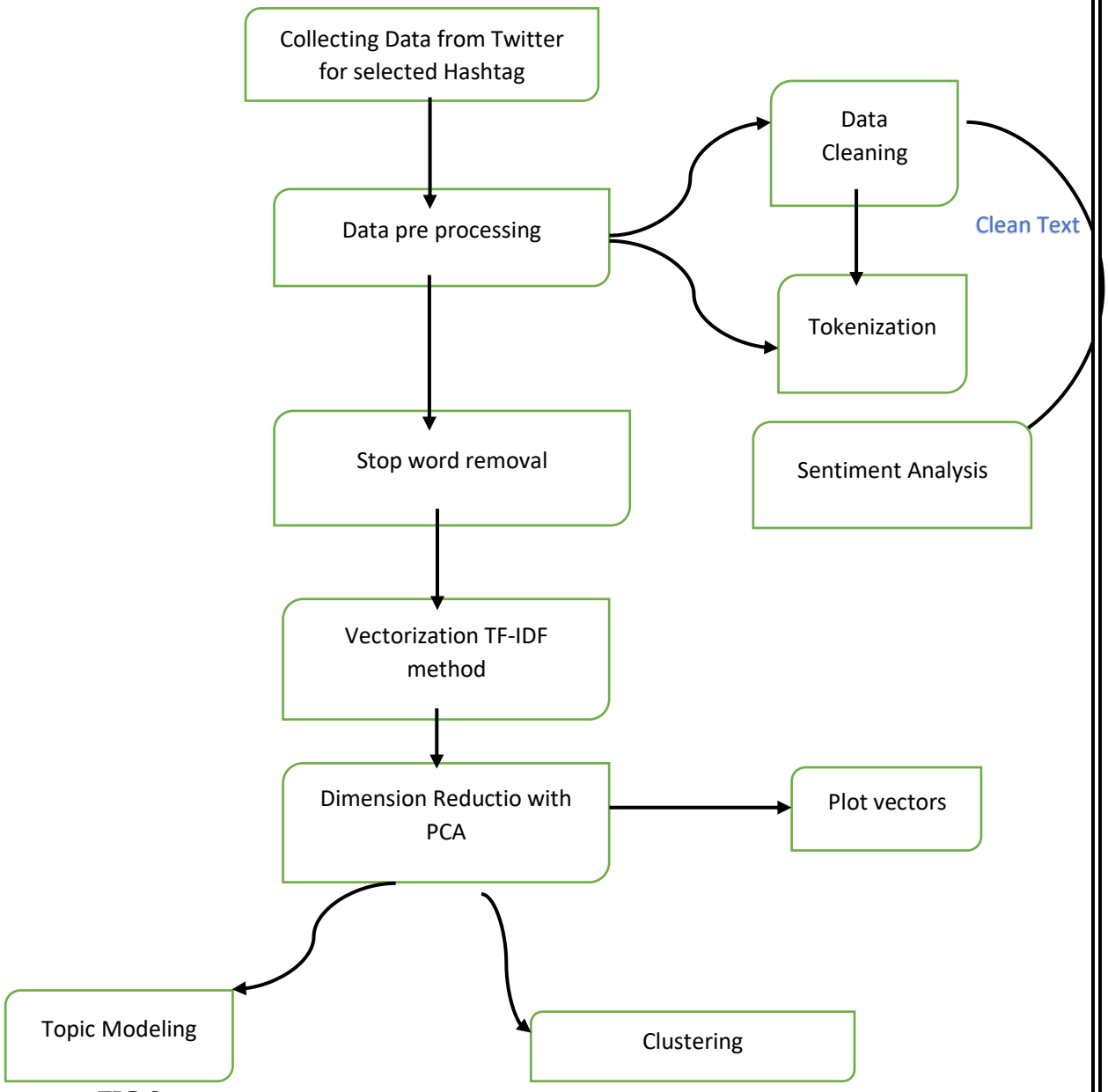

FIG 6

Chapter 3.2.1 Data Collection

Twitter data was retrieved using Twitter API ***tweepy*** by first registering as a developer to use the API. By authorizing access unique credentials (tabs and Secret Key, tokens) were provided to access the data from twitter.

Next,2000 tweets were retrieved about the chosen key word **#StandWithUkraine**

Extended tweet mode was used to retrieve 280 characters of every tweet and also all the retweets are filtered to get only distinct tweets for analyzing. With developer access we get tweets that are only from last week.

As, API specifies rate limit of 450 tweets in every 15 minutes, to avoid exceeding the rate limit wait_on_rate_limit parameter was enabled.

Once all the tweets are retrieved, to do further process dataset is transformed to data-frame using pandas library.This data frame is now the working dataset.

Next to do further analyze Data-Preprocessing is done.

Chapter 3.2.2 Data Preprocessing

Data Pre processing is done in Two steps 1. Cleaning noise -> Sentiment analysis.

2. Tokenization over Clean Data.

Chapter 3.2.2.1 Data Cleaning & sentiment analysis: -

All noises such as retweets, user-handle, multiple spaces, hashtags, hyperlinks, emojis are removed in order to achieve only tweet texts.

As clean texts are achieved now Sentiment analysis is done with this clean text by implementing TextBlob in order to determine how many of these tweets are positive, negative and neutral by calculating polarity and subjectivity.

Following graph shows the polarity percentages of each category:-

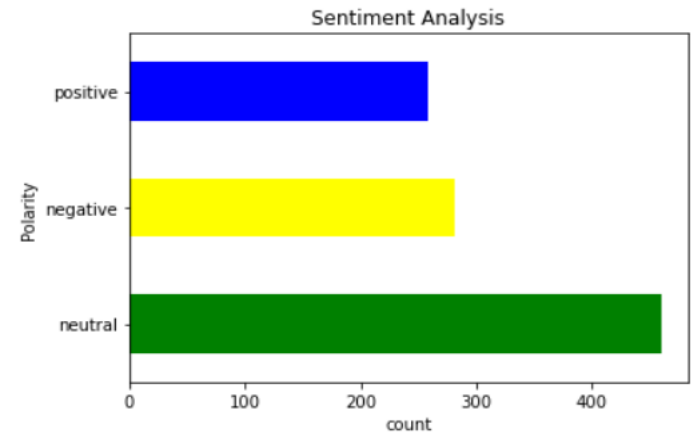

FIG 7

Chapter 3.2.2.2 Tokenization

In this step, Tokenization [refer chapter 3.1.2.3] over clean text from previous is done using *regular expression and regex library.* Each clean tweet is converted to tokens to extract words from a text.

Chapter 3.2.3 Stop Words Removal

Next words that are higher in frequency and do not carry much meaning i.e. stopwords are removed from tokens of each tweet

List of stopwords are downloaded from spaCy library. Now as processed data is ready we move further in process of clustering.

Chapter 3.2.4 Vectorization

In this step , cleaned tokens are mapped to numbers i.e vectorized with both BagOfWords and TF-IDF model. Here we get frequency of each tokens as well as by tf-idf less appearing tokens are associated with higher weights.

Chapter 3.2.5 Dimension Reduction with PCA & Plotting the vectors

In this step to visualize data better dimension of dataset are transformed using PCA linear transform that maps the data in a higher dimensional space to lower dimensional .

Then transformed data are plotted using heatmap.

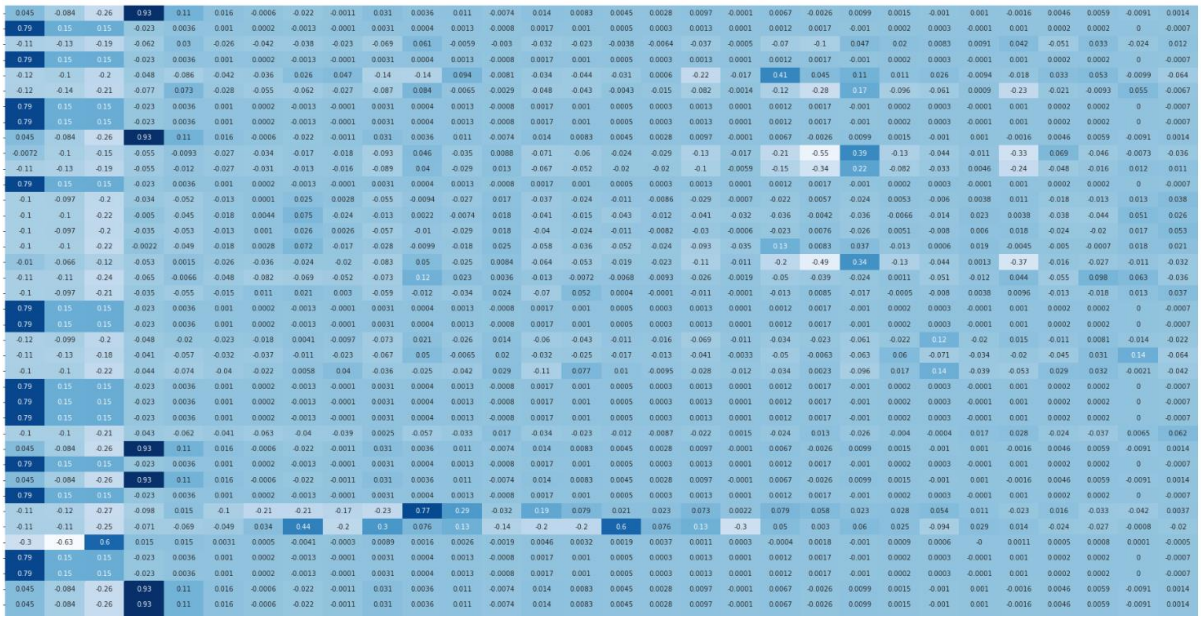

FIG8

3.2.6 Clustering

With this transformed data implementing K-Means algorithm clusters are formed with the help of elbow method.

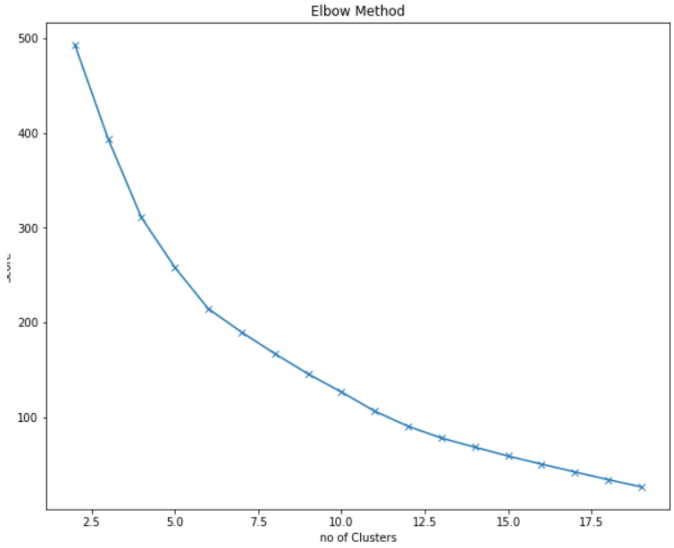

FIG9

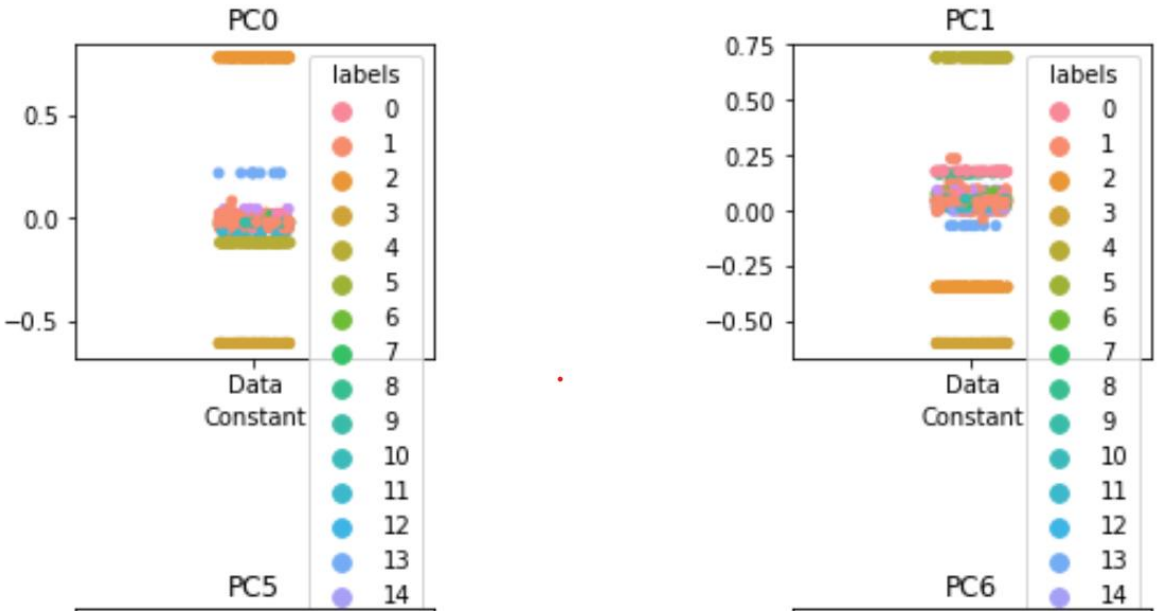

FIG 10

3.2.7 Topic Modelling

For topic modelling, by implementing LDA we get related words clustered in each topic whereas the no. of clusters is provided manually.

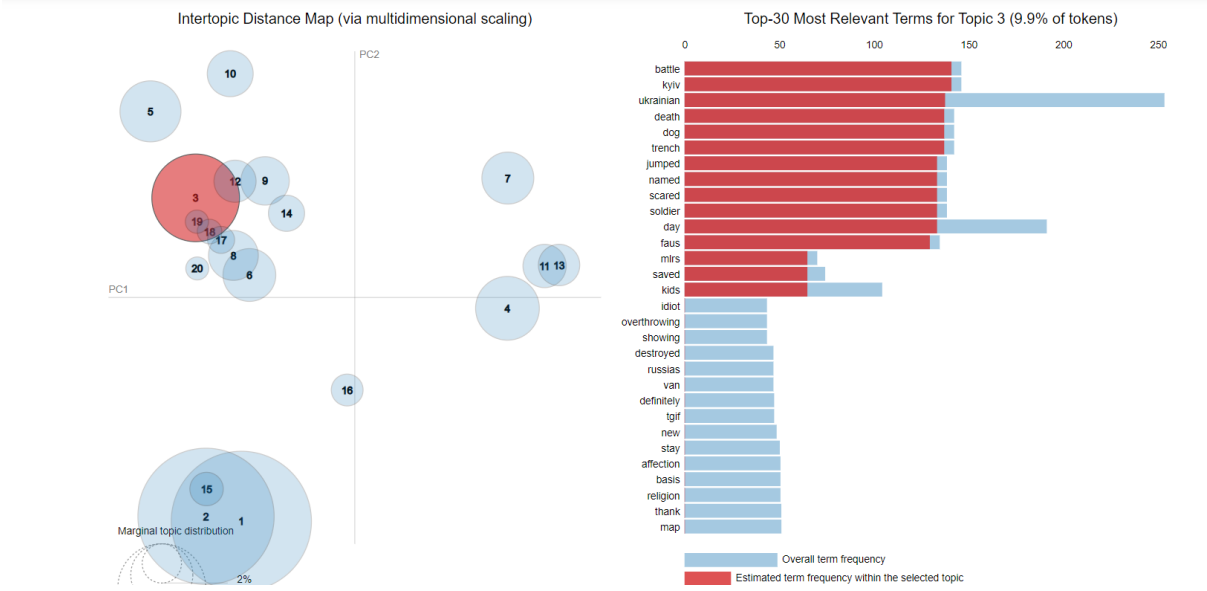

FIG11

4.Result and Performance Analysis

Software used: Jupyter Notebook

Language Used: Python

Dataset: Tweets of different persons about a chosen topic

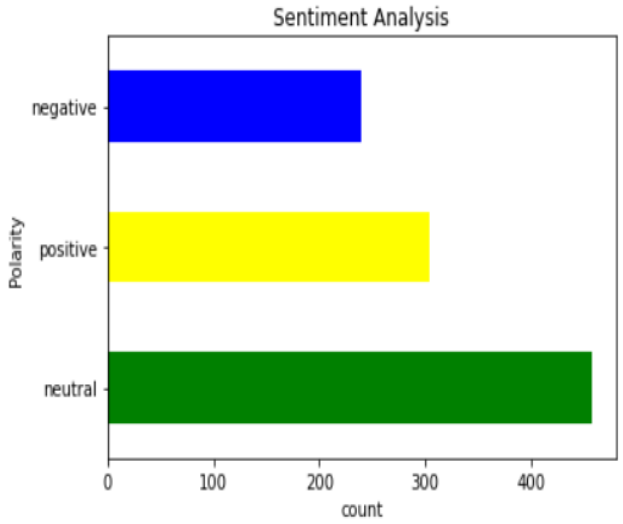

FIG12

```
The percentage of positive tweets is 30.4 %
The percentage of negative tweets is 23.9 %
The percentage of neutral tweets is 45.7 %
```

FIG 13

First, we see the sentiment analysis graph derived from tweets. It shows that more than 450 people are talking positively on the subject

And almost 300 and 220 people have neutral and negative opinion respectively.

Next, we see the words that are used frequently and therefore similar meaning bearing words are clustered under similar topic and also the weights that each keyword carries in descending order:

Topic 00	Topic 01	Topic 02	Topic 03	Topic 04
soldier (9.99)	thousands (8.28)	ukraine (8.82)	ukrainian (7.37)	amp (12.29)
came (9.99)	displaced (8.17)	latest (8.61)	russian (6.93)	ukrainians (10.54)
frontline (9.60)	hundreds (8.07)	defence (8.45)	maks (6.86)	war (9.14)
married (9.60)	killed (7.86)	update (8.37)	executed (6.33)	like (6.96)
brightest (9.60)	sociopathic (7.86)	intelligence (8.37)	levin (6.25)	ukraine (4.78)
dark (9.60)	authoritarian (7.86)	invasion (6.08)	photojournalist (6.01	days (4.41)
deepest (9.60)	madman (7.86)	map (5.90)	reporters (5.66)	support (4.29)
light (9.60)	tens (7.86)	unprovoked (5.82)	interrogated (5.66)	going (4.12)
shines (9.60)	millions (7.86)	illegal (5.82)	tortured (5.66)	today (4.02)
day (9.60)	vladi (7.76)	continuing (5.82)	troops (5.49)	good (3.83)

FIG14

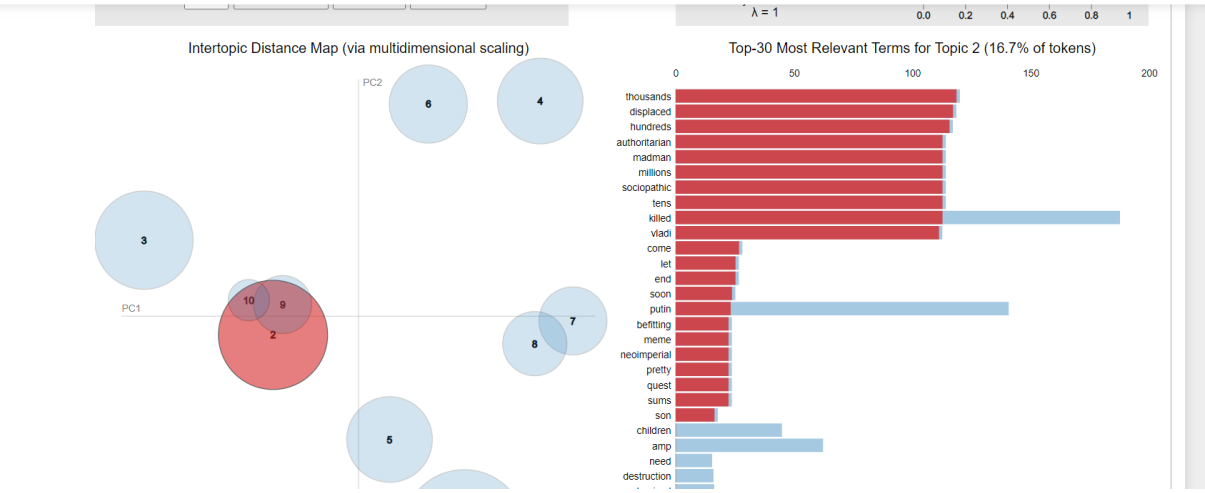

FIG15

Here in FIG15 we see that cluster 2 consist 16.7% tokens of all tokens. Also cluster 9 10 contains similar tokens as cluster 2. So we can say that cluster 2 9 and 10 are somehow similar in nature.

5.Conclusions

By the results, that are achieved in this study of opinion mining, we can say that there's no doubt that this field has great future ahead. Further works that can be done are as following:-

- Predicting what type of people are liking or disliking about a product and modifications can be done for that targeted audience.

- To prevent Cyber bullying, trolling prediction of nature of posts of a person analyzing his/her social media posts.
- Detecting emotion by identifying and analyzing the emotions /emojis from text messages, posts. Detecting several types of feelings such as fear, anger, happiness, sadness, love, vengeance.[16]
- Intent detection by analyzing text data to determine the author's intent underlying in texts whether it carries negative meaning hidden in positive wordplay.[17]

Our main agenda in this study is to detect sentiment of people about a topic that is trending on social media; and then clustering similar type of vectors(words) and analyse what type of sentiment each cluster carries.

Proposed study that we have proposed here can be achieved with higher accuracy if we overcome the challenges by utilizing algorithms:

1. Analysing images and videos.
2. Overcome the language barrier.
3. Retrieve files that is mentioned in hyperlinks.
4. Detection of negation.
5. Eliminate fake opinions of bots.
6. Analyzing emotions behind emojis that are used frequently.
7. Analyze actual sense of texts that are written in a hidden wordplay but carry completely different sentiment.

Bibliography

[1]https://www.kdnuggets.com/2018/04/understanding-behind-sentiment-analysis-part-1.html#:~:text=Historically%2C%20it%20is%20considered%20that,Lee%20and%20by%20Peter%20Turneyaces

[2] Mika V. Mäntylä Daniel Graziotin, Miikka Kuutila The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers

[3] G Dharani Devi Kamalakkannan SomasundaramJan 2019 Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications

[4] Purvi Prajapati Megha Joshi Ayesha Shaikh April 2017 A Survey on Sentiment Analysis

[5] Rudolf Eremyan Four Pitfalls of Sentiment Analysis Accuracy

[6] Keith D. Foote on May 22, 2019 A Brief History of Natural Language Processing (NLP)

[7]https://en.wikipedia.org/wiki/Georgetown%E2%80%93IBM_experiment

[8] https://ai.facebook.com/tools/translate/

[9] https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction

[10]-[12] https://monkeylearn.com/natural-language-processing/

[13]https://www.analyticsvidhya.com/blog/2021/06/text-analytics-of-resume-dataset-with-nlp/

[14] Anita Kumari, M.Sashi Aug 2019 Vectorization of Text Documents for Identifying unifiable News Articles.

[15] Sparck Jones, Karen. "A statistical interpretation of term specificity and its application in retrieval." Journal of documentation 28.1 (1972): 1121.

[16]-[17] Rachit Singh July 2021 Future of Sentiment Analysis.