International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 07 | July 2023 www.irjet.net p-ISSN: 2395-0072

Live Twitter Sentiment Analysis and Interactive Visualizations with PyLDAvis using Streamlit

Sabbineni Lakshmi Gopi Koushik [1], Chinthapatla Navyasri[2], Gurram Keerthana[3], Chunduru Jahnavi[4]

Abstract - *Sentiment analysis, often known as opinion mining, is a technique for determining the emotional tone or attitude indicated in a piece of text, such as a tweet, review, or news story. As Twitter has many influential users, it became an important tool for communication and information exchange on various platforms including politics, business, and entertainment. This paper provides a user-friendly online application for 'sentiment analysis of live Twitter data and interactive visualizations using pyLDAvis' built on Python's VADER module and the Streamlit framework. Our application asks users to enter a topic or phrase of interest and then uses the Twitter API to stream tweets in real-time. We use VADER to do sentiment analysis on incoming tweets and illustrate the results in real-time using various interactive charts and plots. Further, we implement PyLDAvis to do topic modeling on the tweets and display the topics and keywords connected with them. In a dynamic and interactive manner, our application allows users to explore the emotion and themes of Twitter conversations connected to their areas of interest.*

Key Words: Sentiment Analysis, pyLDAvis, Streamlit Framework, Vader library.

1.INTRODUCTION

Natural Language Processing (NLP) is a rapidly emerging computer science topic that has grown in significance over the past few years. NLP is focused on teaching machines how to read and process human language in the same way that humans do. This entails creating algorithms and computer processes capable of analyzing, interpreting, and producing human language data such as text or speech. NLP offers a wide range of practical applications, including text classification and sentiment analysis, as well as machine translation and chatbots. Sentiment analysis, commonly referred to as opinion mining, is an important method for assessing and analyzing sentiment represented in text data. Sentiment analysis uses computer techniques and algorithms to recognize and classify the emotional tone or attitude indicated in a piece of text, such as a tweet, review, or news story. Sentiment analysis is a tool that can analyze client feedback and evaluations, helping businesses in understanding the benefits and drawbacks of their products or services. It can also assist customer care personnel in promptly identifying and responding to bad feedback or complaints posted by customers on social media. Amongst the most difficult things of sentiment analysis is appropriately recognizing the sentiment expressed in text, particularly true for ambiguous sentences in which the sentiment may be unclear. Researchers and practitioners in sentiment analysis are always attempting to increase the accuracy and reliability of sentiment analysis approaches by employing modern machine learning algorithms and natural language processing techniques. Irrespective of challenges, sentiment analysis has grown in importance in the age of big data.

1.1 Problem in Existing System

The problem with the existing system is limited training data. To learn how to identify sentiment in text, sentiment analysis systems rely significantly on training data. However, if the training data is inadequate or polarised, the system may be unable to identify sentiment in new material effectively. Another problem of not having a display of sentiment analysis results is that it can be difficult for users to quickly and readily understand the sentiment of the text. Without a visual representation, consumers may have to go through huge amounts of text or data to understand the overall sentiment, which can be time-consuming and ineffective. Patterns and trends in sentiment analysis data may not always be seen from the text alone, but visualization can help. A visualization, for example, may illustrate the frequency of good, negative, and neutral sentiments across time, allowing viewers to observe changes in sentiment over time.

2. PROPOSED WORK

The proposed work provides an architecture for live Twitter sentiment analysis, which is based on a lexicon-based approach and includes classification issues in real-time. This model is entirely built on querying real-time tweets from the Snscrape and pre-processing them into a corpus of words using a Lexicon-based method, where the terms are specified.

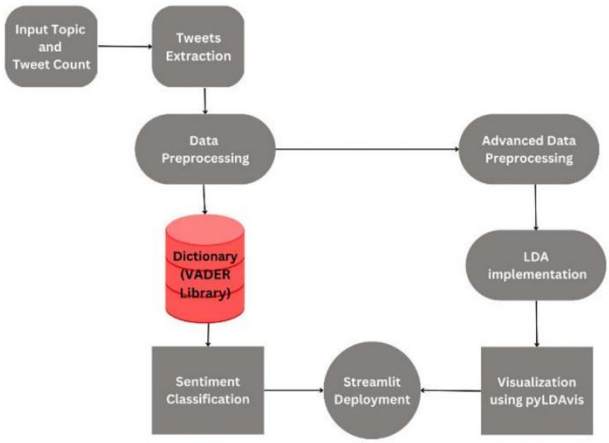


Fig -1: Proposed Work

3. PROPOSED METHOD

3.1 Data Collection

We must extract the data in order to comprehend emotions and carry out sentiment analysis. Twitter is used to collect the data as a consequence. The information from tweets, comments, and hashtags on Twitter is extracted.

Tags used to extract the data,

- Our task in this study is to deal with data tagging. Users, tags, and resources are the three entities that typically comprise a social tagging system.
- Label: #, date, User, Tweet
- Tags we used to extract the data: #Digitalindia, #DigitalPayments.

Libraries used to extract the data,

- Using the Twitter API or Tweepy, Twitter enables us to collect the data of any user. The user's tweets will be extracted as the data.
- Snscrape (SNS) is the acronym for social networking service scraper. It retrieves the stuff it has found by scraping information from searches, hashtags, and user profiles.
- The number of tweets collected is lower while using tweepy to extract the data. So, we extracted hashtags, comments, and related posts using snscrape.

3.2 Data Cleaning

The extracted data is inconsistent. Noisy data requires more storage and makes sentiment analysis less accurate. Stop words, hashtags, mentions, and non-English words can all be found in extracted tweets. Therefore, this data must be cleaned before performing sentiment analysis and topic modeling.

3.2.1 Removing special symbols and emojis.

- We use Regex to remove characters like #, and @ as collected tweets also contain mentions.
- Special characters present in the data make them unsuitable for sentiment analysis.

3.2.2 Removing stop words.

- Stop words are some English words that are not helpful in sentiment analysis, such as "I," "am," "and" and "an." Instead, they result in more storage and inaccurate analysis.
- Stop words can be eliminated to provide more attention to meaningful words.
- We use nltk (Natural Language Tool Kit) which is a suite that contains libraries to process the textual data.

3.2.3 Remove punctuations.

- In English, punctuation marks are symbols that provide the sentence additional meaning and are helpful for grammar. For sentiment analysis, however, these symbols are useless.
- '!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~' • The above punctuations are removed using string. punctuation library in nltk.
- In English, the words 'Digital' and 'digital' are the same. But, in NLP models, these are treated differently.
- To avoid this ambiguity, we convert the entire file into the same case either lower or uppercase.
- Data[]=Data[].apply(lambda x:x.lower())

3.2.4 Tokenization

- In Python, we use the .split() function to divide the contents of an array or list. In NLP, we tokenize words similarly by taking into account white spaces.
- We can use either the .split() method or the wordpunkttokenizer library of nltk.

3.2.5 Lemmatization

- Lemmatization is the process that reduces a word to its simple forms. It is used to eliminate additional tweet noise.

- For Example, the words like loving, and caring after lemmatization becomes love and care respectively.
- By lemmatization, we can get the most possible accurate frequency distributions.

4. IMPLEMENTATION

4.1 Sentiment Analysis

Now, the data is cleansed from all stop words, and punctuations, and it is in its base form, the data is perfect for sentiment analysis. It gives accurate results as the noise is removed from the data. We used VADER as a modeling approach for sentiment analysis which is a rule-based sentiment analysis tool designed for social media posts. It determines the sentiment of a piece of text by using a lexicon of words and emoticons that have been graded for their valence (positive, negative, or neutral) and intensity. VADER can also detect contextual cues such as capitalization and punctuation, which might influence the sentiment of a text. VADER's ability to handle rejection and sarcasm is one of its key advantages. Even though the word "happy" is positive, VADER would correctly detect the sentence "I'm not happy with my new dress" as negative. The VADER lexicon is a collection of words and phrases that have been graded on two dimensions: valence and intensity.

- Valence scores range from -1 (extremely negative) to +1 (extremely positive), with 0 signifying neutral sentiment.
- The intensity score runs from 0 (least intense) to 2. (Most intense).

4.2 Topic Modeling

The process of removing necessary properties from a Bag of words is called topic modeling. This is significant because NLP treats each word in the corpus as a feature.

4.2.1 Latent Semantic Analysis (LSA)

By developing a list of concepts that are statistically related to the documents and terms, distributional semantics, a subset of latent semantic analysis (LSA), is used in natural language processing to evaluate the relationships between a set of documents and the terms they contain. LSA is mostly used for concept searching and automatically classifying documents. In order to represent the text data in terms of these topics or latent qualities, the latent semantic analysis must be performed.

4.2.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a topic model type used to classify a document's text into different categories based on a certain topic. It constructs a topic-per-document model and a words-per-topic model using Dirichlet distributions as the modeling basis. Following pre-processing operations, the model was trained using (LSA, and LDA), and the results were then evaluated using the coherence value. These outcomes demonstrate that the LDA technique performed better than the LSA technique.

4.2.3 PyLDAvis

A useful tool for visualizing topics from a topic model is the PyLDAvis module. Use the pyLDAvis cell in the notebook to execute the interactive dashboard.

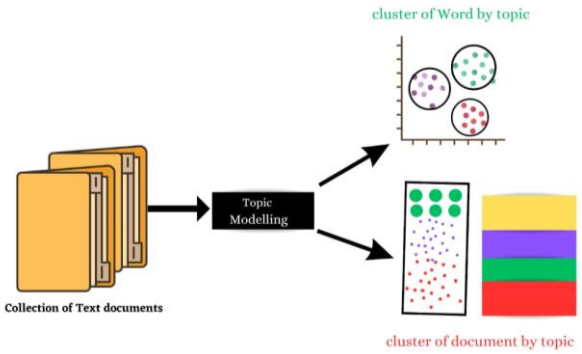


Fig -2: Topic Modeling

4.3 Streamlit Deployment

Streamlit is a commonly recognized Python framework for creating engaging online applications for data science and machine learning events. Deploying a Streamlit application entails hosting it on a server and making it accessible through the internet. Streamlit offers a free hosting platform called Streamlit Sharing, which allows you to quickly distribute your Streamlit app. Connecting your GitHub repository to Streamlit Sharing and delivering your app with a few clicks makes it simple to deploy your project.

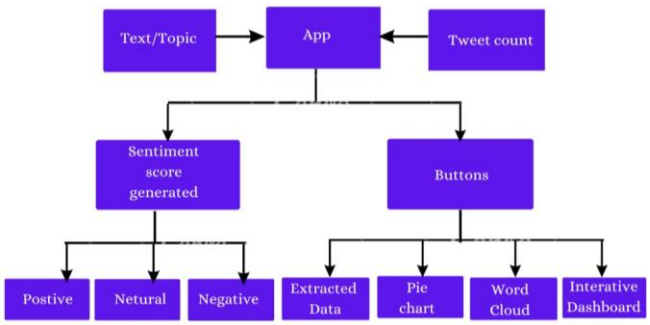


Fig -3: App overview

The deployed streamlit application receives the input topic/text and tweet count. If no tweet count is specified, the default tweet count is 50. It calculates the sentiment score for a given topic and categorizes it as negative or positive. It also has manual buttons for producing outputs like word clouds, pie charts, and interactive dashboards.

5. RESULTS

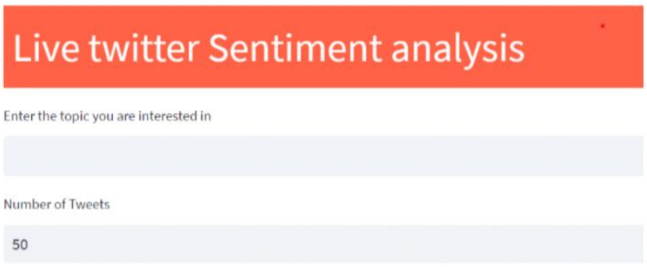


Fig -4: Firstly, enter the text/topic of the latest news or a product.

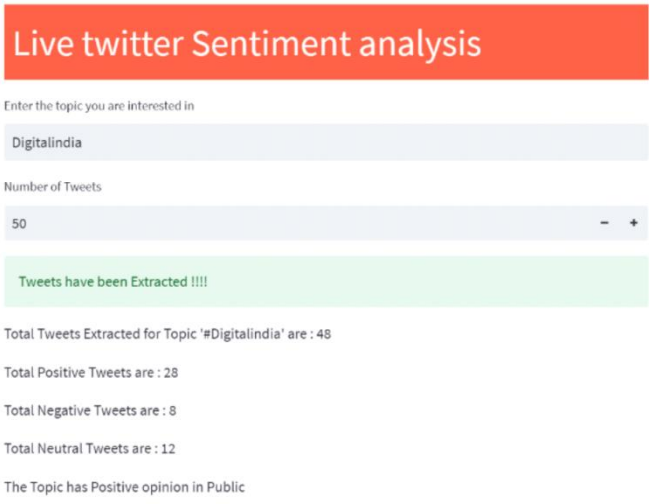
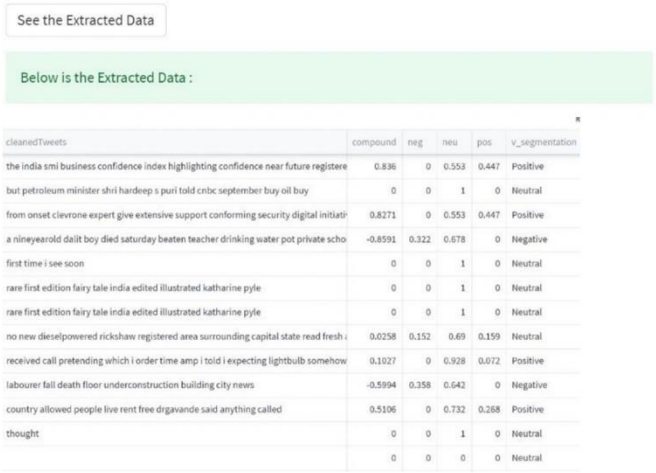


Fig -5: Analysis Performed

As you can see, the total number of tweets is divided into three categories: positive, negative, and neutral. These are combined to form the public's concluding view on the subject. You may explore the extracted text, word cloud, pie chart, and interactive dashboard by selecting one of the options listed below.



cleanedTweets	compound	neg	neu	pos	v_segmentation
the india smi business confidence index highlighting confidence near future registere	0.836	0	0.553	0.447	Positive
but petroleum minister shri hardeep s puri told cnbc september buy oil buy	0	0	1	0	Neutral
from onset clevrone expert give extensive support conforming security digital initiati	0.8271	0	0.553	0.447	Positive
a nineyearold dalit boy died saturday beaten teacher drinking water pot private scho	-0.8591	0.322	0.678	0	Negative
first time i see soon	0	0	1	0	Neutral
rare first edition fairy tale india edited illustrated katharine pyle	0	0	1	0	Neutral
rare first edition fairy tale india edited illustrated katharine pyle	0	0	1	0	Neutral
no new dieselpowered rickshaw registered area surrounding capital state read fresh	0.0258	0.152	0.69	0.159	Neutral
received call pretending which i order time amp i told i expecting lightbulb somehow	0.1027	0	0.928	0.072	Positive
labourer fall death floor underconstruction building city news	-0.5994	0.358	0.642	0	Negative
country allowed people live rent free drgavande said anything called	0.5106	0	0.732	0.268	Positive
thought	0	0	1	0	Neutral
	0	0	0	0	Neutral

Fig -6: Extracted data with the fields of cleaned tweets, positive, negative sentiment, etc. are displayed.

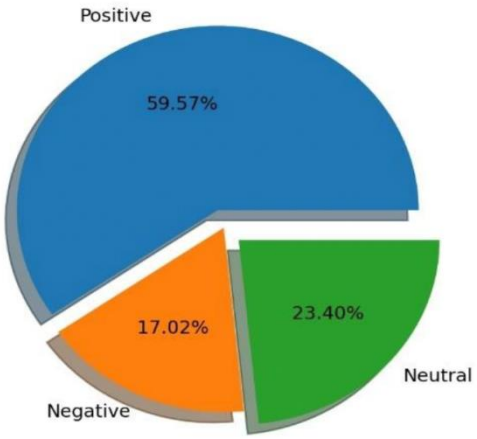


Fig -7: Pie chart of Sentiments

This is illustrated by the pie chart, for Digital India, where it is noticed that 59.57% of tweets are favorable, 17.02% are negative, and 23.40% are neutral. According to the research, the majority of the population is positive about this initiation.

Fig -8: Word cloud

Presenting the word cloud, which represents the most repeated terms in a text, with the size of each word representing its relative frequency or importance.

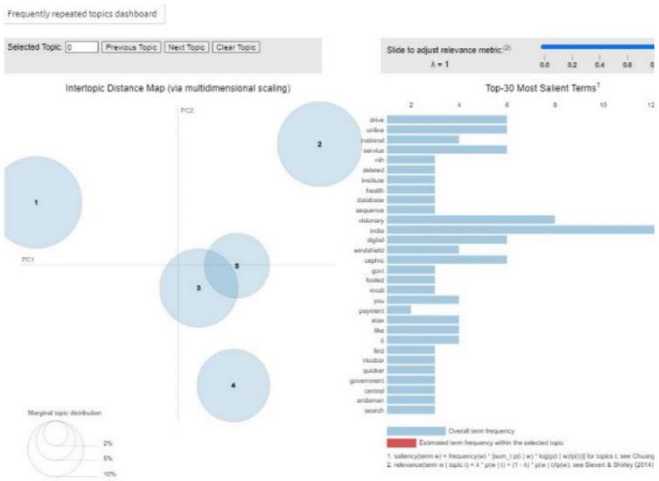


Fig -9: PyLDAvis Dashboard

Interactive dashboards provide Realtime sentiment trend visualization and can give insights into how sentiment evolves over time.

6. CONCLUSIONS

Finally, Live Twitter Sentiment Analysis and Interactive Visualizations with PyLDAvis Using Streamlit is a powerful tool for real-time sentiment analysis on social media. The application analyses the sentiment of tweets using natural language processing techniques and visualizes the findings via interactive visualizations.

Users do use PyLDAvis to investigate the subjects addressed in tweets and acquire insights into the most popular themes and attitudes. Streamlit offers an easy-to-use interface for engaging with data, making it simple to explore and analyze. Overall, Live Twitter Sentiment Analysis and Interactive Visualizations with PyLDAvis Using Streamlit might be a useful resource for organizations and people interested in monitoring and analyzing social media sentiment. This tool can assist by offering real-time insights into public sentiment.

REFERENCES

[1] Patil, Shilpa and Lokesha, V., Live Twitter Sentiment Analysis Using Streamlit Framework (May 25, 2022).

[2] Elbagir, S., & Yang, J. [2019, March]. Twitter sentiment analysis using natural language toolkit and VADER sentiment. In Proceedings of the international multiconference of engineers and computer scientists (Vol. 122, p. 16).

[3] Jagdale, R. S., Shirsat, V.S., & Deshmukh, S. N. [2016]. Sentiment analysis of events from Twitter using open-source tool. IJCSMC, 5(4), 475-485.

[4] Prabhakar Kaila, D. [2016]. An Empirical Text mining analysis of Fort Mcmurray wildfire disaster twitter communication using topic model. Disaster Advances, 9(7).

[5] Rai, S., S B, G., & Kumar, J. [2020]. Sentiment Analysis of Twitter Data. International Research Journal on Advanced Science Hub, 2, 56-61.