Sentithon Hackthon 2K22

Topic - Digital India

Team - Cluster

Sabbineni Lakshmi Gopi Koushik

Chunduru Jahnavi

Chinthapatla NavyaSri

GIT HUB - SLGkoushik/Sentithon-2k22-CLUSTER-

DIGITAL INDIA

We chose the theme "Digital India" because, in the modern world, every developed country is modernizing and placing a priority on cashless trade, and at this point in the growth of our economy, digital India promotes such trade. In this sentiment analysis, we investigate all positive, negative, and neutral opinions of people on digital India.

Data Collection:

- **Platforms used for extracting the data – Twitter**

  Twitter is a platform where people share their opinion or views, and find out what is happening in the world right now. This makes Twitter a perfect platform to source data for conducting research.

- **Tags are used to extract the data**

  #Digitalindia, #DigitalPayments

- **APIs or libraries used to perform data extraction – Tweepy, Snscrape**

  Twitter allows us to mine the data of any user using Twitter API or Tweepy. The data will be tweets extracted from the user. Snscrape is a scraper for social networking services (SNS). It scrapes details like user profiles, hashtags, or searches and returns the discovered items,

- **Statistics of the extracted data**

  - Length of data – 21352(unique)

  - Null values - No

  - 2272920 total words, with a vocabulary size of 209

  - Max tweet length is 285

  - Word cloud

- **Collected data is Noisy**

  The information gathered from tweets includes stop words, hashtags, non-English languages, and mentions.

Data Cleaning:

The various data cleaning steps are:

- **Removing # and @ words**
  Since # appears in all tweets and @ is used for account names, we remove them using Regex.
- **Stop words removal**
  Stop words are the most common words of a language like 'I', 'this', 'is', and 'in' which do not add much value to the meaning of a document. These values are removed to decrease the dataset size and increase focus on meaningful words.
- **Lower Casing and Removing Punctuations**
  - In NLP, models treat words like "Goat" and "goat" differently, even if they are the same. Therefore, to overcome this problem, we convert all the letters into the same case i.e; lower.
  - Punctuations are the marks in English like commas, hyphens, full stops, etc. These are important for English grammar but not for text analysis. Therefore, they need to be removed
- **Tokenization**
  The tokenizer simply splits words by whitespace, similar to the pythons .split() method.
- **Lemmatization**
  To further reduce noise in our text data, and get the most accurate frequency distributions possible, cue lemmatization. Lemmatization is the act of breaking a word token down to its root meaning.

Sentiment Analysis:

Modeling Approach – Text Blob

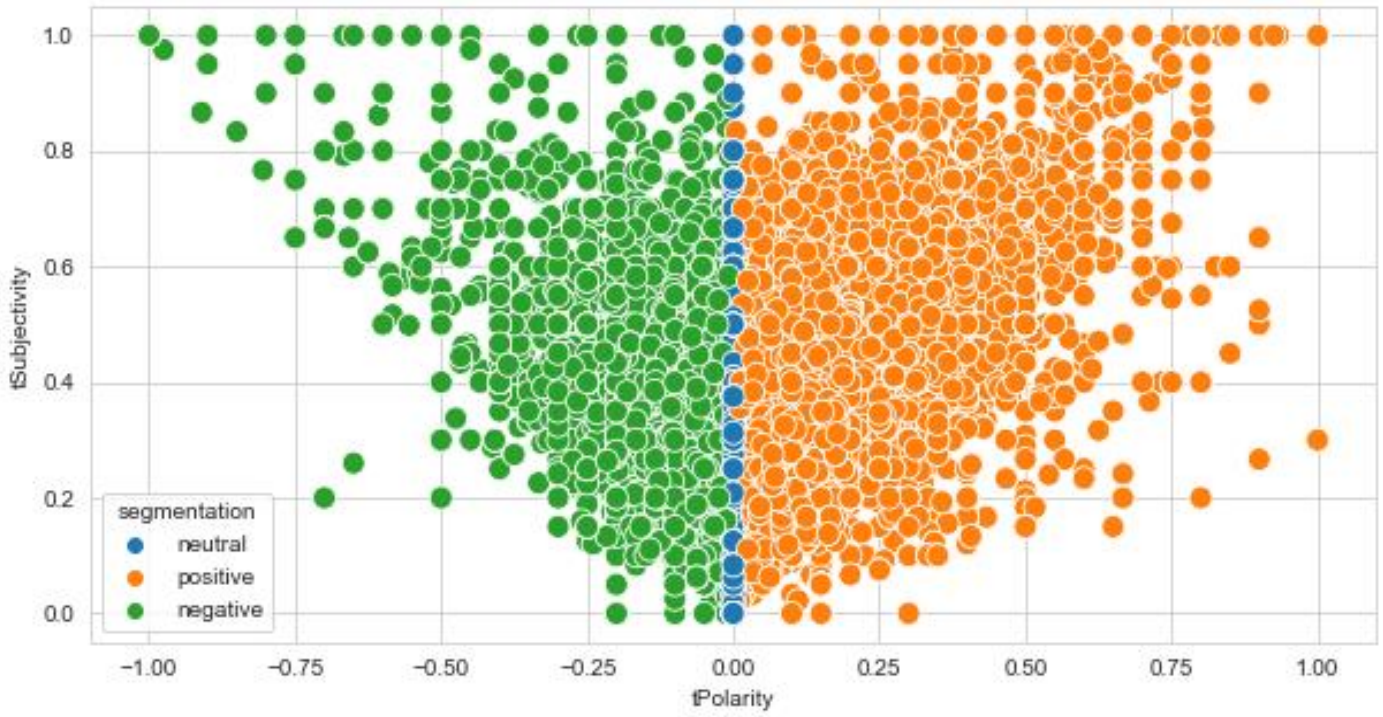
Text blob sentiment analyzer returns two properties for a given input sentence:

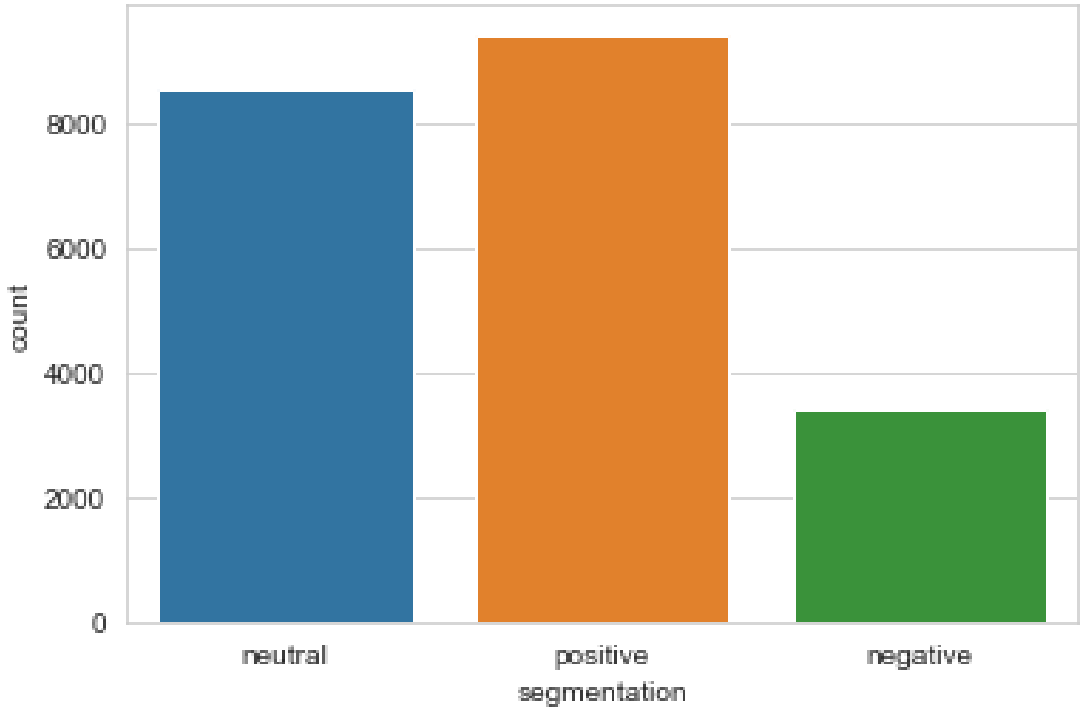
- Polarity is a float that lies between [-1,1], -1 indicates negative sentiment and +1 indicates positive sentiment.
- Subjectivity is also a float that lies in the range of [0,1]. Subjective sentences generally refer to opinion, emotion, or judgment.

Why Text Blob?

Text Blob is built on top of NLTK and Pattern also it is very easy to use and can process the text in a few lines of code. Text Blob is an open-source and free-to-use python library for processing textual data. It offers a simple API to access its methods and perform basic NLP tasks.

Text Blob – Segmentation of our data





Topic Modeling:

Topic modeling is the method of extracting needed attributes from a bag of words. This is critical because each word in the corpus is treated as a feature in NLP.

Latent Dirichlet Allocation (LDA)

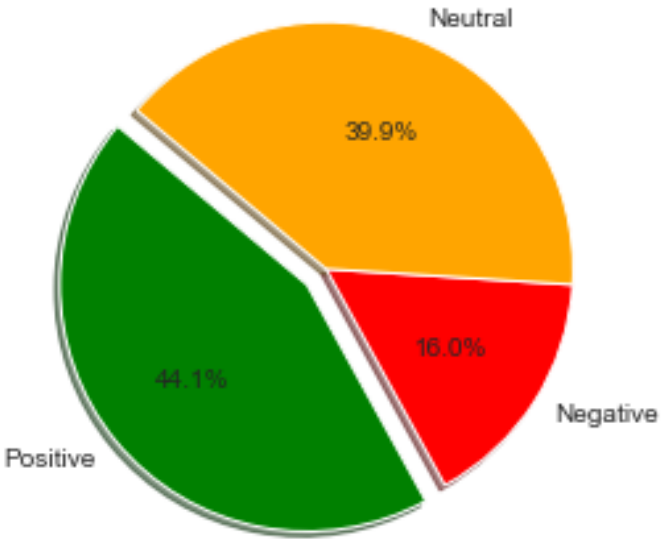
Latent Dirichlet Allocation (LDA) is an example of a topic model and is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modeled as Dirichlet distributions.

PyLDAvis

The PyLDAvis library is a great way to visualize topics from a topic model.

For the interactive dashboard run the pyLDAvis cell in the notebook.

Results:



According to the public's perceptions, 44.1% of people are favourable towards digital India, compared to only 16% who are unfavourable and the remaining who seem neutral.

Fitting an LDA model in Gensim

Here's the output:

```
[(0,
  '0.046*"india" + 0.036*"service" + 0.026*"app" + 0.025*"digital" + 0.021*"rt" + 0.019*"state" + 0.014*"this" + 0.011*"u
mang" + 0.011*"government" + 0.010*"amp"'),
 (1,
  '0.027*"digital" + 0.023*"india" + 0.013*"i" + 0.009*"payment" + 0.009*"bank" + 0.007*"day" + 0.007*"one" + 0.007*"tr
ansaction" + 0.006*"u" + 0.006*"amp"'),
 (2,
  '0.034*"we" + 0.026*"with" + 0.022*"plus" + 0.019*"india" + 0.014*"group" + 0.014*"of" + 0.013*"are" + 0.013*"based"
 + 0.013*"company" + 0.011*"govt"'),
 (3,
  '0.035*"india" + 0.029*"digital" + 0.020*"the" + 0.017*"amp" + 0.014*"ad" + 0.013*"from" + 0.011*"a" + 0.010*"to" + 0.
010*"and" + 0.010*"our"'),
 (4,
  '0.027*"digital" + 0.015*"vle" + 0.015*"csc" + 0.008*"million" + 0.007*"certificate" + 0.007*"marketing" + 0.007*"raised"
 + 0.007*"story" + 0.006*"district" + 0.005*"easy"')]
```

The output represents 5 topics, consisting of the top keywords and associated weightage contribution to the topic.

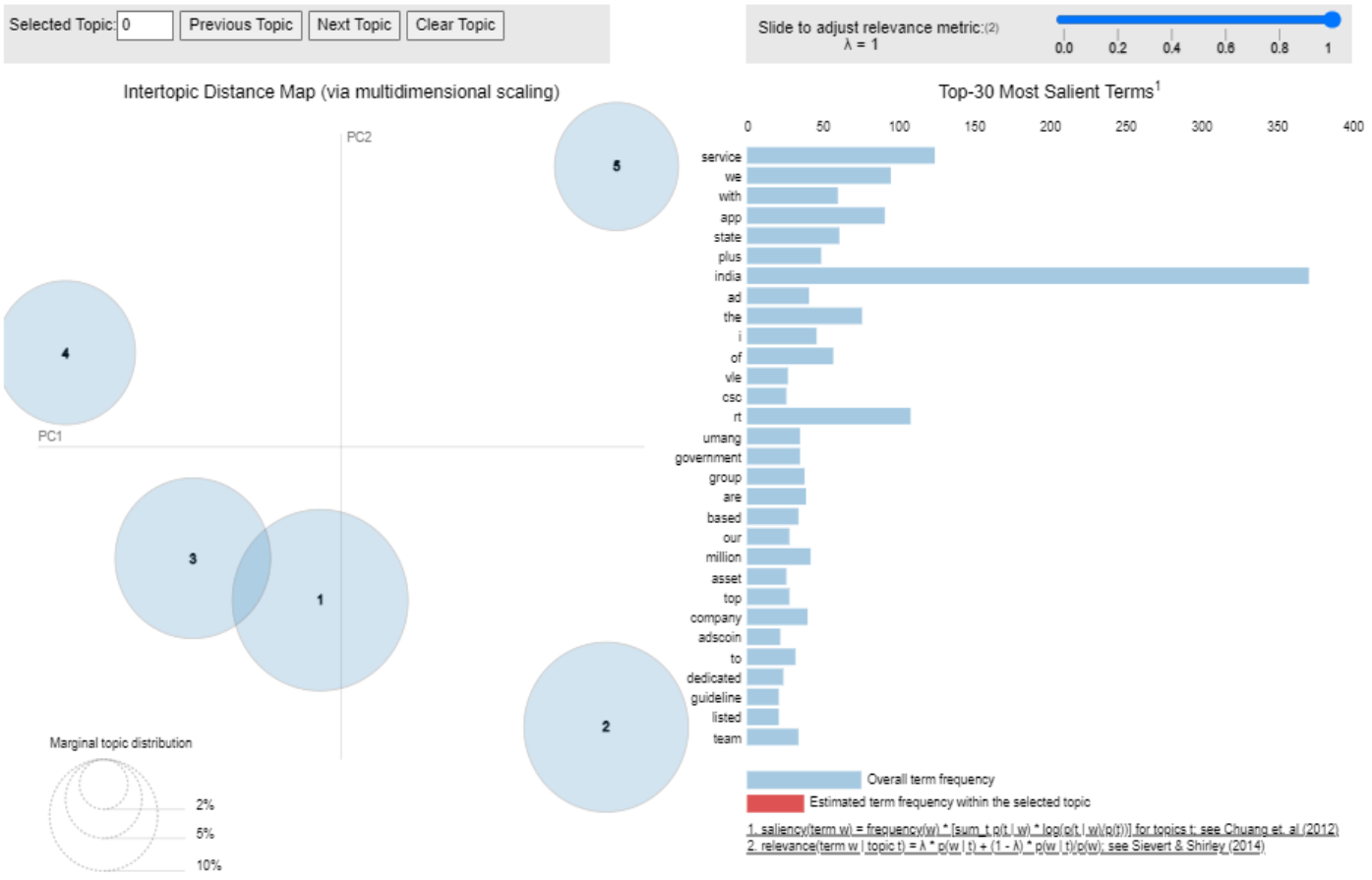
Ex:

From Eyeballing

1st topic may be related to the Umang app and its revolution.

2nd topic may be related to increase in the number of transactions due to the digital India initiative.

PyLDAvis Output:



**** For the interactive dashboard run the pyLDAvis cell in the notebook****

On the left, the topics are plotted on a 2-dimensional plane representing the distance between each topic. While the right horizontal bar chart represents the words most relevant to each topic. The chart is interactive, allowing you to select specific topics and view the related words for each topic, in hope of inferring meaning from each topic.

GitHub link:

https://github.com/SLGkoushik/Sentithon-2k22-CLUSTER-.git