

Twitter Analysis And Opinion Mining

Twitter contains lots of data in the form of tweets. People twitting, giving their reviews, opinions, positive and negative feedback in tweets with the help of emoji’s and expresses their emotions with the different hashtags.

Twitter Opinion Mining is the process in which try to find out what people were talking about particular keywords in tweets finding using NLP techniques.

# Problem Statement

The goal of project is using Natural language Processing and text analysis find out the sentiment in tweets and extract different topics present in it.

# Methods

1. Twitter API-Using twitter API with the help of tokens get all the tweets and collect dataset.
2. Elastic search- Storing large number of tweets use elastic search for storing documents.
3. Text Analysis – Use nltk python library for text cleaning and processing.
4. Sentiment Analysis – Classify tweets in positive, neutral and negative in3 different opinions using tweets polarity and subjectivity.
5. LDA Topic model – Explore and extract the different topics in tweets

# Architecture

Text preprcessing

punctuations

BERT ML model

Store index

Return source reading

Data Synchronization

Clean text

**Twitter Opinion mining**

Twtter API



Remove stopwords

Convert emoji to text

Lemmatization

stemming



Dashboard

Topic modeling

# 4. Data Collection

Twitter is a big data source. Millions of tweets sent per second. These tweets can be used to find out opinion and their sentiment in specific issues. Uses twitter API with the keywords coronavirus, covid-19.

For the backup purposes store tweets in elastic search with the index name tweets and query the text if needed.

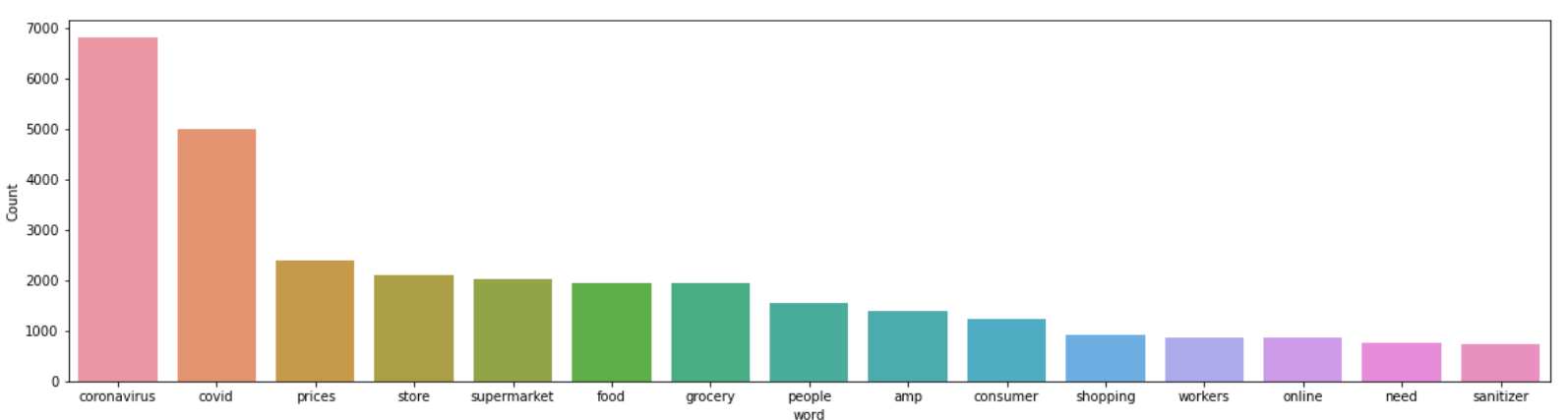
# 5. Text Preprocessing

Text Preprocessing is the major process for text cleaning. The main objective is to remove noise which is present in the form of urls, punctuation’s, special characters and symbols. Because they are meaningless. Remove stop words (using NLTK library) such as a, an and with having lesser meaning to tweets. Convert tweets to lowercase. Lemmatization which convert word such as plurals, ing and es words to its root word .Convert emoji’s to text which will add sentiment polarity.

# 6.Data Analysis and Data visualization

To express the emotions different new terms are used which gives bad impact on word set characteristics. The users who use most frequent words are extracted from the tweets indicates as top words. The top words are Covid, Prices, Supermarket, online, Shopping, workers, sanitizer and food with their respective counts.

Top 15 most frequent words in that mostly used in tweets are following



# Word cloud

One of the text data visualization method is Word cloud. In the word cloud each word represents word frequency with their size. The most frequent words with the larger word size whereas least frequent words having small word font size. As per choice mask word cloud into any shape, pattern, size as well as color.





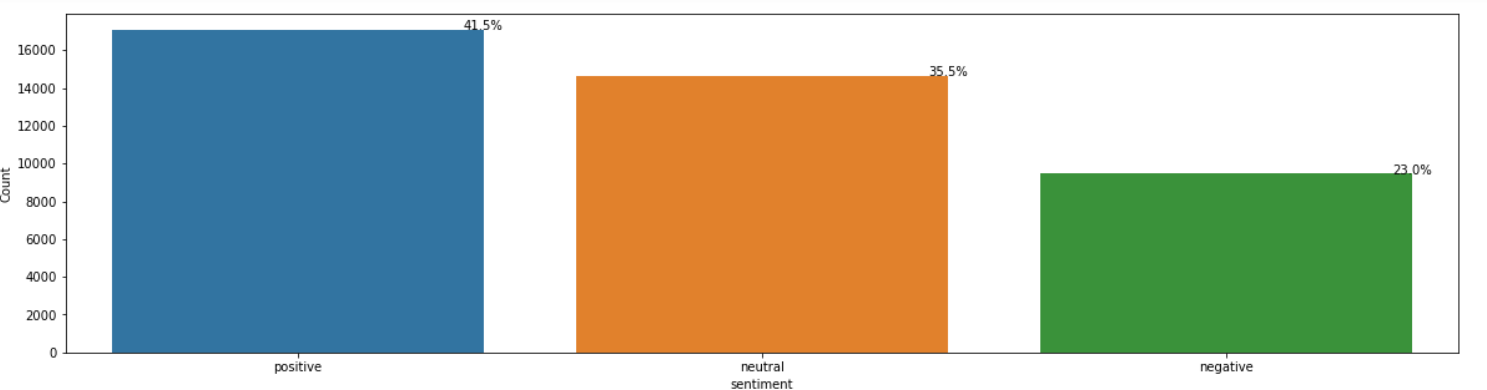
# 7. Sentiment Analysis

It is process in which classify tweets into 3 different categories which are positive, negative and neutral towards topic using text and words present in it.

# BERT model

Bidirectional Encoder Representations or Transformer is the full form of BERT model. It is the Sate Art of Model (NLP) developed by Google. Trains on large corpus and learn patters in data and generalizes on new data. Transformer based architecture consists of encoder and decoder. Bidirectional means learned information from both sides during training data. For that install Transformer library and use sentiment analysis pipeline from Hugging face.

To recognize significant topics, firstly the sentiment analysis is performed and each of review is labeled as the positive and negative. After sentiment analysis there were 16.9% positive and 83.1% Negative reviews in the dataset examined. For detailed analysis, the sentiment score is calculated in sentiment analysis. Further, part described as the sentiment analysis based on their scores.



# 8. Topic modeling

LDA stands for Latent Dirichlet Allocations. It is unsupervised topic modeling algorithms in which dividing words into groups so that at the end each group represent topic per document.

Latent: Hidden in the data. Topics to particular document are not known but considered as they present in text based on words that represent each topic.

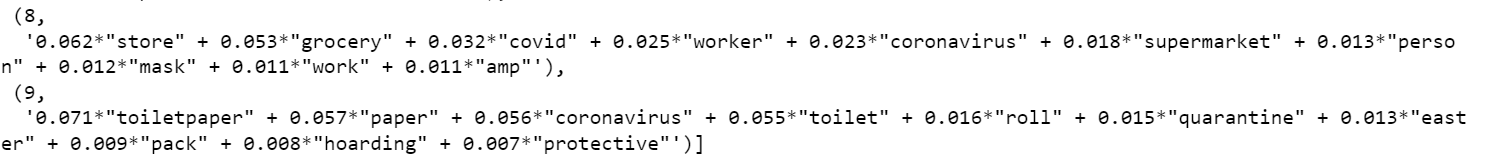
Dirichlet: It is the distribution of topics in documents and distribution of words in the topic.

Allocation: Allocate topics to the documents and words of the document to topics.

Firstly, the Latent Dirichlet Allocation (LDA) is used for acquiring the meaningful topic information from collected datasets. For training LDA model textual data is used. The textual data start with removing stop words, and tokenizing the text as reviews. From the tokenized object creates collection of texts. Building of LDA model with number of 10 topics as an input where each topic represents collection of different keywords having a certain weightage of keywords, assigned to that particular topic.

The LDA model give us 10 topics with different keywords and their respective weightage of words in specific topic. The words with the highest weightage is most important in that topic.





The 10 topics with their 10 relevant words are shown

Topic 0 – About coronavirus

Topic 1- Tell us stay home

Topic 2- coronavirus impact on retail industry

Topic 3- high price and oil demand

Topc 4 – Drug selling company

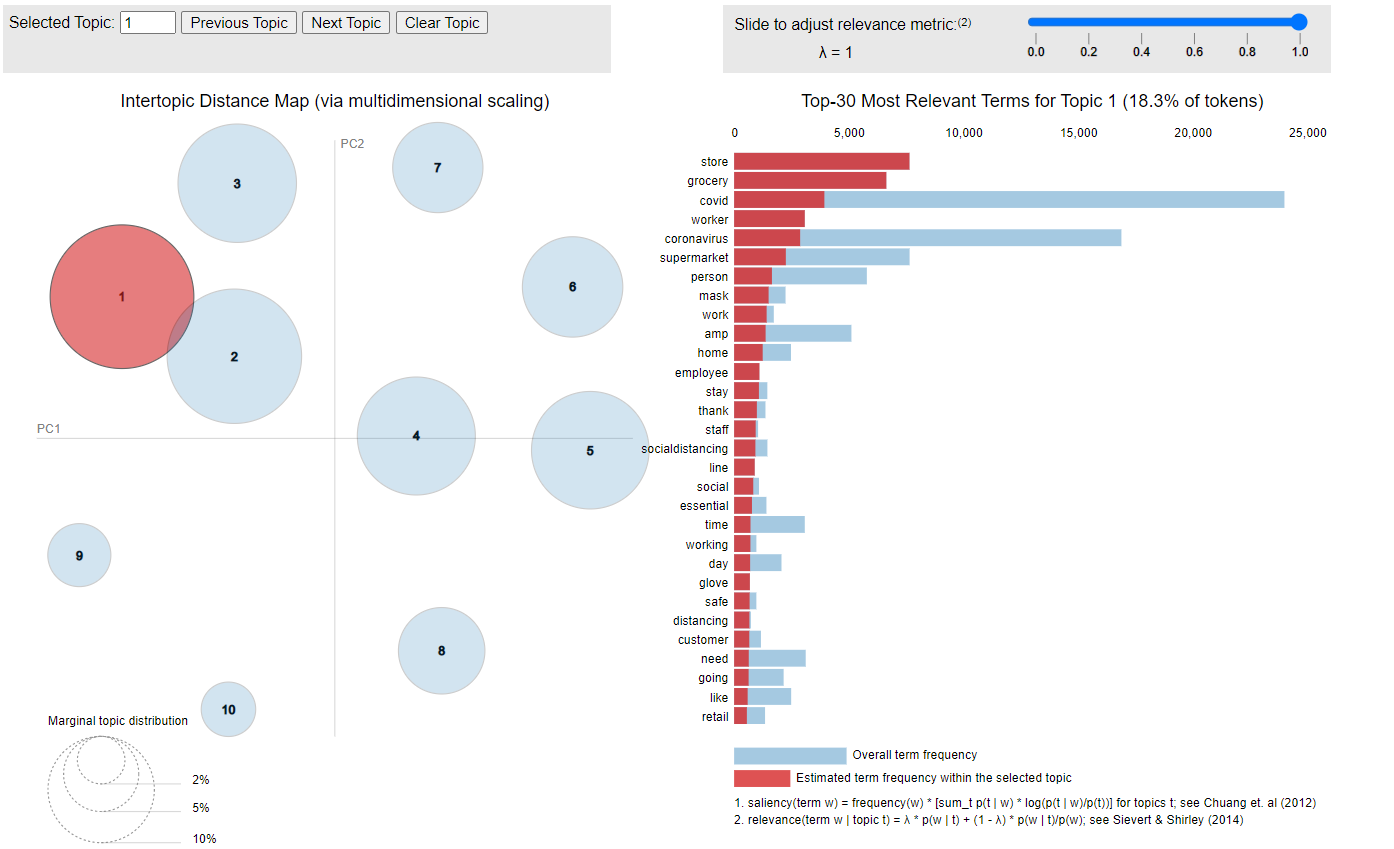
Topc 5 – Pandemic and online consumer shopping

Topc 6 – change in food supply and market

Topc 7 – Due to panic people start stock food

Topic 8- wear mask

Topc 9- Quarantine and toilet paper



The above diagram represent LDA visualization left part represents number of topics and right side is word used in that topic. Size of topic represents how much topic is prevalent .The bubbles overlapping are sign of too many topics. When hover over specific topic at the right side related to that topic keywords shown in saliency term. Using lambda function we can set the each topic with most similar or least frequent terms with the relevance parameters.

# 9. Summary

Natural Language processing field provided different machine learning models which work as a black box algorithms for topic modeling. With the help of LDA model (dimension reduction) extracting valuable information from large corpus. Using unsupervised technique find out the different topics hidden in text and from that concluded their sentiments. Model visualization gives an idea about what are the different terms or keywords used in topics. In further parts how to evaluate the best number of topics using LDA parameters and metrics.Tunning hyper parameters to achieve best and more native results. Distinguish between different topic modeling techniques.

# Refrences

Topic model — Wikipedia. <https://en.wikipedia.org/wiki/Topic_model>

<https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/>

word cloud - https://amueller.github.io/word\_cloud/auto\_examples/masked.html

LDA model - https://radimrehurek.com/gensim/models/ldamodel.html

Chau, M., Xu, J., 2012. Business intelligence in blogs: understanding consumer interactions and communities. MIS Quart. 36

Pak, A., Paroubek, P., 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/