Social Media Sentimental Analysis

Literature Review:

Sentiment Analysis on Social Media by Jyoti Yadav

Social Media Sentiment Analysis: A Comprehensive Analysis

Tools and Technology

data collection

Name	Type	Link	Description
Twitter API	API	https://developer.x.com/en/docs/x-api	using twitter API to fetch tweet but it has rate limit and only 3200 free per month
twint	Scrapping	https://github.com/twintproject/twint	scrapping tools to scrape tweet without limits from twitter
Twitter Kaggle	Kaggle dataset	https://www.kaggle.com/datasets/kazanova/sentiment140	This dataset contain 1.6m tweet from the use of API of twitter
Reddit Kaggle	Kaggle dataset	https://www.kaggle.com/datasets/prakharrathi25/reddit-data-huge	Reddit dataset that has sub-reddit to choose from a very huge selection and big, good for NLP
tweepy	python library for twitter api		
Praw	python library for reddits api		

data processing and storing

Name	Type

Spark	processing
Hadoop	storing

Pipeline v1:

- □ Extract data
  - □ fetch data from reddits, twitter and kaggle
- □ Processed data
  - □ Convert text to lowercase
  - □ remove most common stop words such as a, about, above…
  - □ remove non character texts such as punctuations and emojis from text
  - □ filter and remove repeated words, URLs, and number from texts
  - □ tokenization was done to convert texts into tokens, which is to split sentences into smaller units or words. So meaning can assign to word more easily
  - □ Stemming was done to extract base form of the words by removing affixes from them (EX: words such as “likes” , “likely” and “liked” returned as “like” after stemming)
  - □ Term Frequency-Inverse Document Frequency Vectorizer (TF-IDF) was pre-owned to assess how relevant a term is in the corpus/text data, where TF-IDF vectorization is process for calculating the TF-IDF score for every word.
- □ Sentiment Analysis
  - □ Perform sentimental analysis using VADAER, TextBlob, BERT
  - □ add sentiment scores to each post (positive, negative, neutral)
- □ Feature Extraction
  - □ use TF-IDF or Word2Vec to convert text into numerical features
- □ Clustering
  - □ apply K-Means or DBSCAN to group base on sentiment or topic similarity
- □ Topic Modeling
  - □ apply LDA to find underlying topics within clusters
  - □ combine sentimental analysis and LDA to visualize how sentiment changes across topics
- □ Visualization
  - □ use t-SNE or PCA for visualize clusters
  - □ Visualize sentiment distribution within each topic using **pie charts** or **bar graphs**.