SENTIMENT ANALYSIS ON TWEETS DISCUSSING CHATGPT

Data Mining – CS 573

Presented By

Akhil Prasad

&

Chandrika Mukherjee

PURDUE UNIVERSITY | Department of Computer Science

Problem Statement and Motivation

- Analyze various tweets about ChatGPT and to understand the overall public sentiment towards it using well-known data mining methods.

- **Motivation** -
  - ChatGPT has experienced significant usage since its public launch
    - Content generation, language translation, writing assistance, programming assistance, entertainment – creating fictional stories, jokes etc.
    - Online sources indicate that there are different perspectives about the use of ChatGPT
    - There has been significant discussion regarding its applications, capabilities, drawbacks and impacts on human in future.

Reviews About ChatGPT – Mix of Positives and Negatives

"What I liked most about ChatGPT was its ability to provide quick and accurate answers to a wide range of questions. It was incredibly helpful in getting information and explanations on various topics."

Positive

Negative

"...ChatGPT performs poorly in terms of helpfulness for the medical domain in both English and Chinese.

The ChatGPT often gives lengthy answers to medical consulting in our collected dataset, while human experts may directly give straightforward answers or suggestions, which may partly explain why volunteers consider human answers to be more helpful in the medical domain."

Overall: I love it. It's like having a personal assistant, business coach, content generator, and editor all in your pocket. If you have a new business, this is a goldmine b/c while you may not be able to afford a marketer right off that bat, you can use Chat to make doing it yourself easier!

Positive

Negative

ChatGPT has bias baked into the system

Negative

"Complex phrasal features are based on the frequency of specific words and phrases within the analyzed text that occur more frequently in human text.

...Of these complex phrasal features, idiom features retain the most predictive power in detection of current generative models."

Negative

Concerns over ChatGPT Training and Privacy Issues

Methodology

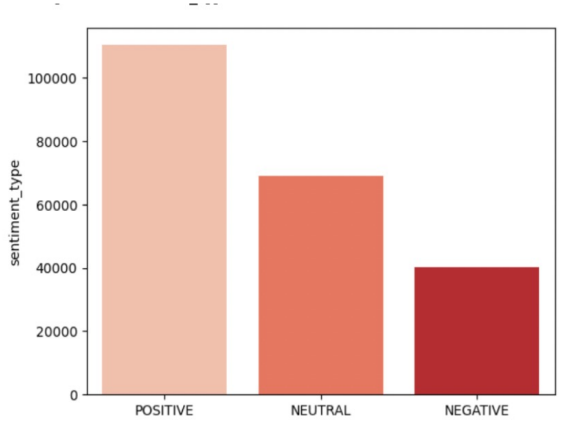
- Dataset Selection –
  - Initially identified four different datasets from Kaggle and Huggingface. With primary analysis, we identified following two more suitable datasets.
    - [https://www.kaggle.com/datasets/pcminh0505/chatgpt-twitter] [primary]
    - [https://huggingface.co/datasets/deberain/ChatGPT-Tweets] [secondary]
    - Selected primary and secondary both datasets have popular tweets about ChatGPT with primary dataset having more information about the tweet and the secondary dataset having more information about the user.

- Preprocessing –
  - Tweets are converted to lower case, stop-words, hyperlinks, emojis were removed from the tweets. Then lemmatization was applied to group together inflected forms of a word.

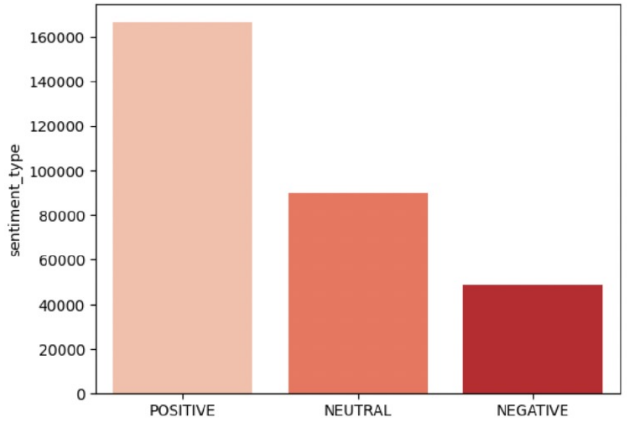
	tweet	final_tweet
0	ChatGPT: Optimizing Language Models for Dialog...	optimizing language model dialogue
1	Try talking with ChatGPT, our new AI system wh...	try talking new ai system optimized dialogue f...
2	ChatGPT: Optimizing Language Models for Dialog...	optimizing language model dialogue ai machinel...
3	THRILLED to share that ChatGPT, our new model ...	thrilled share new model optimized dialog publ...
4	As of 2 minutes ago, @OpenAI released their ne...	2 minute ago released new nnand use right
5	Just launched ChatGPT, our new AI system which...	launched new ai system optimized dialogue
6	As of 2 minutes ago, @OpenAI released their ne...	2 minute ago released new nnand use right n n
7	ChatGPT coming out strong refusing to help me ...	coming strong refusing help stalk someone agre...

Methodology

- Sentiment Analysis –
  - Relied on simple lexicon-based approach to determine the associated sentiment of the tweets. These rules are typically based on lexical and syntactic features of the text – such as presence of positive negative words and phrases.
  - We have incorporated Valence Aware Dictionary and Sentiment Reasoner (VADER Sentiment Analysis) to label the data.
- Feature Extraction from Tweets –
  - *To obtain dense vector representations of words that capture their semantic meanings, we have used tools to vectorize each tweet sentence to 100-dimensional vector.*
- Train and Test Split –
  - Divided the dataset into train and test set (80/20)
- Data Mining Methods –
  - Applied Logistic Regression, SVM, LSTM to train three different data mining models and tested the accuracy.



Sentiment Distribution in primary dataset



Sentiment Distribution in secondary dataset

Results and Findings

- LR -

  Test Accuracy – 64%
  Precision – 62%
  F1-score – 61%
  Recall – 64%

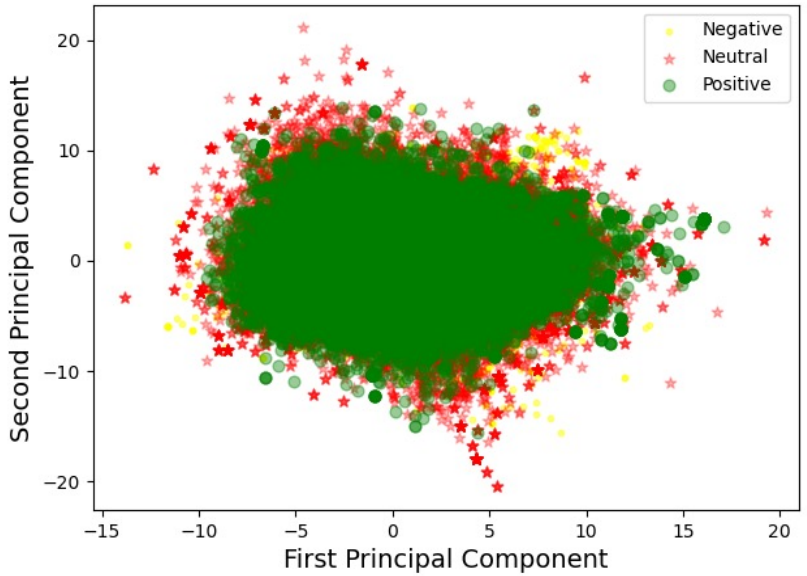
- SVM –

  Test Accuracy – 76%
  Precision – 74%
  F1-score – 72%
  Recall – 74%

- LSTM –

  Test Accuracy – 90%
  Precision – 90%
  F1-score – 90%
  Recall – 90%

Performance -
LSTM> SVM> LR



Analysis

- Performed PCA on the vectorized tweet data and projected it in 2D. Given that the data is no-linear, this might be one of the reasons for SVM to have better accuracy.

Insights and Future Work

- Analyzing the available data, we found that people have different positive, negative and neutral perspectives about the use of ChatGPT. Our selected dataset contains more positive data than the others.
- Instead of VADER based sentiment labelling, we can make use of pretrained transformer based deep-learning models.
- This project limits our work to sentiment analysis, as an extension of this, we want to work on the causality of the tweet data. Would be interesting to perform an extensive causal analysis of the tweets.
- With causal analysis on the tweets, would be able to understand the performance of ChatGPT in a better way and we can help in making the GenAI models better.
- Without relying on data from social media, we can arrange online, or offline surveys asking people specific questions about usage of ChatGPT and use that dataset to apply our data mining methods.

THANK YOU