SENTIMENTAL ANALYSIS

BY

Andres Quintana, Joseph Cardenas, Kavitha Dasaratha

PROBLEM STATEMENT:

In today's digital age, text analysis and text mining have become essential parts of various industries. Text analysis refers to the process of analyzing and extracting meaningful insights from unstructured text data. One of the most important subfields of text analysis is sentiment analysis, which involves determining the emotional tone of the text.

Sentiment analysis is the practice of using algorithms to classify various samples of related text into overall positive, negative or neutral categories.

In this project we are performing the sentimental analysis for Amazon food review dataset.

GENERAL APPROACH:

We know that most of the datasets that is generated today are unstructured. To perform the sentimental analysis, we have to preprocess the data, classify and train the model.

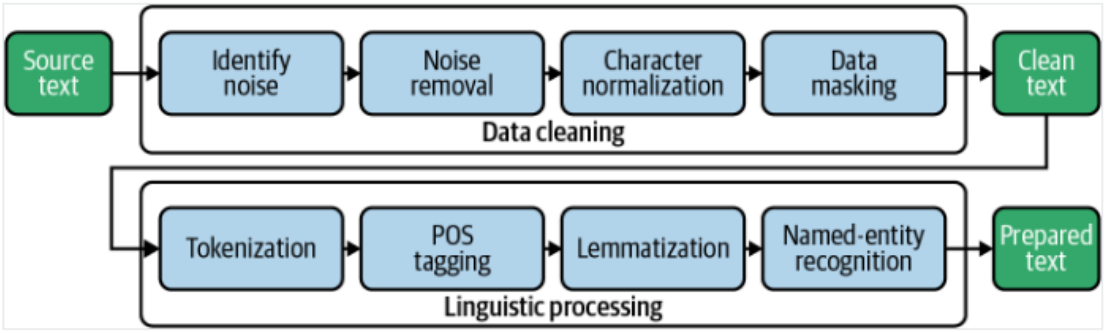
The project is broken down into preprocessing and classification.

- **NLP tool kit –** Natural Language Toolkit is a collection of libraries for natural language processing.
- **Naïve bayes algorithm -** The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

Preprocessing Text

Text preprocessing is a crucial step in performing sentiment analysis, as it helps to clean and normalize the text data, making it easier to analyze. The preprocessing step involves a series of techniques that help transform raw text data into a form you can use for analysis. Some common

text preprocessing techniques include tokenization, stop word removal, stemming, and lemmatization.



Tokenization

Tokenization is a text preprocessing step in sentiment analysis that involves breaking down the text into individual words or tokens. This is an essential step in analyzing text data as it helps to separate individual words from the raw text, making it easier to analyze and understand. Tokenization is typically performed using NLTK's built-in `word_tokenize` function, which can split the text into individual words and punctuation marks.

Stop words

Stop word removal is a crucial text preprocessing step in sentiment analysis that involves removing common and irrelevant words that are unlikely to convey much sentiment. Stop words are words that are very common in a language and do not carry much meaning, such as "and," "the," "of," and "it." These words can cause noise and skew the analysis if they are not removed.

By removing stop words, the remaining words in the text are more likely to indicate the sentiment being expressed. This can help to improve the accuracy of the sentiment analysis. NLTK provides a built-in list of stop words for several languages, which can be used to filter out these words from the text data.

Stemming and Lemmatization

Stemming and lemmatization are techniques used to reduce words to their root forms. Stemming involves removing the suffixes from words, such as "ing" or "ed," to reduce them to their base form. For example, the word "jumping" would be stemmed to "jump."

Lemmatization, however, involves reducing words to their base form based on their part of speech. For example, the word "jumped" would be lemmatized to "jump," but the word "jumping" would be lemmatized to "jumping" since it is a present participle.

Naïve Bayes Classifier

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Programming Language

In this project, we have used python code on co-lab IDE.

Performance

The model was 78.6 % accurate for the condensed dataset (included only 400 rows) used from Kaggle for Amazon Fine Food review. If we increase the number of rows, we can achieve higher efficiency.

Conclusion and future work

In this project, we were able to perform sentimental analysis for each review for the dataset. As a scope for expansion, we would like to use our model to take user input and classify the text.

Summarize the main lessons learned from the project.

We learnt the use of NLP tool kit used for data processing and apply the concepts of Naïve Bayes algorithm learnt in the class.

REFERENCES

https://www.datacamp.com/tutorial/text-analytics-beginners-nltk

https://www.ibm.com/topics/naive-bayes

dataset: https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews/