Semester 1, 2024/2025

TTTP6234 UNSTRUCTURED DATA ANALYTICS

PROJECT A - Lexicon-based SA (domain lexicon vs public lexicon)

Prepared by:

Name	Matric Number
MOHAMED MOUBARAK MOHAMED MISBAHOU MKOUBOI	P139575

Lecturer:

Dr. Lailatul Qadri Zakaria

ABSTRACT

The performance of a customized lexicon-based sentiment analysis technique and the open-source VADER sentiment analysis tool are compared in this project. We built a custom lexicon using a dataset of 1,000 tweets and evaluated its performance against VADER's pre-trained lexicon. The accuracy of each technique's classification of tweets as neutral, negative, or positive was evaluated. The results indicate that 92.40% accuracy was achieved by both techniques. This illustrates how custom lexicons that adapt to certain datasets can perform on the same level as popular tools like VADER.

INTRODUCTION

A key component of natural language processing is sentiment analysis, which makes it possible to extract opinions, attitudes, and feelings from textual data. It can be used in a variety of businesses, from assessing social media trends to understanding consumer input. In this project, we focus on lexicon-based sentiment analysis, a simple and efficient method which ranks the polarity of words and combines these scores to classify the sentiment of texts. We have 1,000 tweets in our dataset that were created with Python modules. Three sentiments are annotated on each tweet: neutral, negative, or positive. The dataset's balance and diversity ensure evaluation reliability. The dataset contains metadata such as IDs, creation dates, and usernames along with text content. This dataset offers a good starting point for evaluating a customized lexicon's performance against the popular sentiment analysis tool VADER.

RESEARCH METHOD

- Dataset Preparation

The dataset consists of up to 1,000 tweets that were generated using Python. A balanced distribution of neutral, negative, and positive labels was achieved by using the textblob module to randomly assign sentiments. The preprocessing methods listed below were used:

  - Removing punctuation, URLs, and numbers.
  - Converting text to lowercase.
  - Tokenizing text into words.
  - Removing stop words and applying lemmatization.

- Custom Lexicon Development

The custom lexicon was created by:

1. Grouping words by sentiment labels in the dataset.

2. Counting the occurrences of each word within positive, negative, and neutral categories.
3. Assigning sentiment scores to words:
    - Positive: +1
    - Negative: -1
    - Neutral: 0

The foundation for classification of sentiments was the lexicon that resulted, which mapped each word to its sentiment score.

- Sentiment Analysis
    - Custom Lexicon-Based Analysis

    The sentiment score of each tweet was determined by adding up the word scores using the custom lexicon.

        - Sentiment was assigned as:
            - ✓ Positive: score > 0
            - ✓ Negative: score < 0
            - ✓ Neutral: score = 0
    - VADER-Based Analysis

    Using its pre-trained lexicon, VADER (Valence Aware Dictionary and Sentiment Reasoner) was used to classify tweets.

        - The compound score was used to assign sentiments:
            - ✓ Positive: compound > 0.05
            - ✓ Negative: compound < -0.05
            - ✓ Neutral: compound between -0.05 and 0.05.

EVALUATION AND RESULT

- Accuracy
    - Custom Lexicon Accuracy: 92.40%
    - VADER Accuracy: 92.40%

The efficacy of a dataset-specific custom lexicon was demonstrated by the fact that both approaches obtained the same accuracy on the dataset.

- Observations
    - ✓ Due to its customized nature, which captured features unique to the dataset, the custom lexicon performed well.
    - ✓ Although VADER lacked dataset-specific optimizations, its general-purpose lexicon was able to equal the performance of the custom technique.

- ✓ The overall accuracy and misclassification rates were impacted by the ineffectiveness of both approaches in capturing neutral sentiments.

CONCLUSION AND FUTURE WORK

This project shows that performance comparable to well-known tools such as VADER may be obtained with a well-designed custom lexicon. The findings highlight how crucial it is to customize sentiment analysis techniques for certain datasets in order to achieve the best results.

For future work:

- Contextual Information Incorporation: Improve the custom lexicon by taking contextual polarity shifts and multi-word phrases into account.
- Managing Intensifiers and Negatives: To improve sentiment grading, handle intensifiers (like "very good") and negation words (like "not good") better.
- Improving Neutral Sentiment Detection: Since both systems had trouble correctly recognizing and classifying neutral sentiments, more reliable techniques should be developed.

REFERENCES

- Prabu Palanisamy, Vineet Yadav, and Harsha Elchuri, "Serendio: Simple and Practical Lexicon-Based Approach to Sentiment Analysis," SemEval 2013.
- NLTK Documentation: https://www.nltk.org
- VADER Sentiment Analysis: https://github.com/cjhutto/vaderSentiment
- TextBlob Documentation: https://textblob.readthedocs.io/
- Dataset can be accessed at: https://www.kaggle.com/datasets/jocelyndumlao/twitter-sentiment-analysis-using-roberta-and-vader/data