UNIVERSITI
KEBANGSAAN
MALAYSIA
The National University of Malaysia

NARATIF BAHARU UKM
UNIVERSITI
WATAN KITA

Semester 1, 2024/2025

TTTP6234 UNSTRUCTURED DATA ANALYTICS

PROJECT B - Machine Learning SA vs lexicon SA

Prepared by:

Name	Matric Number
MOHAMED MOUBARAK MOHAMED MISBAHOU MKOUBOI	P139575

Lecturer:

Dr. Lailatul Qadri Zakaria

ABSTRACT

This project uses a dataset of 1,000 synthetic tweets to evaluate the effectiveness of sentiment analysis techniques based on lexicon and machine learning. The VADER lexicon-based tool and a logistic regression model were used to examine the dataset, which was generated systematically with balanced sentiments. Although the results demonstrate that both approaches attain exceptional accuracy on the dataset, each has distinct benefits. This analysis offers insights into possible advancements for sentiment analysis approaches and emphasizes the significance of choosing the best technique based on requirements.

INTRODUCTION

Social media platforms with a lot of user-generated content, such as Twitter, are great resources for sentiment analysis. A dataset of 1,000 synthetic tweets generated by Python's random, faker, and textblob modules is used in this project. Reliable analysis is achieved by the dataset's balance of neutral, negative, and positive sentiments. This variety allows us to evaluate two popular sentiment analysis approaches: lexicon-based and machine learning (ML). This project compares their performance to identify the best approach for this dataset and offer useful information for further research.

RESEARCH METHOD

- Machine Learning-Based Sentiment Analysis
  - ■ Experiment Setup

  The machine learning pipeline involved the following steps:

  1. Data Preprocessing:
     - ❖ Text cleaning to remove special characters, hashtags, mentions, and URLs.
     - ❖ Text is converted to lowercase and extra whitespace is removed.
  2. Feature Extraction:

     Using unigram and bigram features, the Term Frequency-Inverse Document Frequency (TF-IDF) approach was used to vectorize the text.
  3. Model Training:
     - ❖ Three classifiers were trained: Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression.
     - ❖ Each model was evaluated using metrics for accuracy, precision, recall, and F1-score, as well as an 80-20 train-test split.

  - ■ Best Model Selection

Based on the test data, the model with the highest accuracy, precision, recall, and F1-score were found to be logistic regression.

- Lexicon-Based Sentiment Analysis

The VADER (Valence Aware Dictionary and Sentiment Reasoner) tool was used in the lexicon-based analysis. The actions listed below were taken:

1. Preprocessing:

The ML pipeline's approaches were also used to clean the text.

2. Sentiment Scoring:

Each tweet's sentiment polarity score was calculated using VADER.

Predetermined thresholds were used to assign sentiment labels (positive, negative, and neutral).

EVALUATION AND RESULT

- Machine Learning Results

➢ Logistic Regression:

✓ Accuracy: 100%

✓ Classification Report:

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	82
neutral	1.00	1.00	1.00	50
positive	1.00	1.00	1.00	68

accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

➢ SVM:

✓ Accuracy: 100%

✓ Classification Report:

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	82
neutral	1.00	1.00	1.00	50
positive	1.00	1.00	1.00	68

accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

➢ Naïve Bayes:

✓ Accuracy: 100%

✓ Classification Report:

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	82
neutral	1.00	1.00	1.00	50
positive	1.00	1.00	1.00	68

accuracy			1.00	200
macro avg	1.00	1.00	1.00	200
weighted avg	1.00	1.00	1.00	200

- Lexicon-Based Results

On the dataset, VADER also demonstrated good accuracy, correctly classifying the sentiment of most tweets. But sometimes, VADER misclassified tweets that contained subtle or context-dependent sentiment, such intensifiers or sarcasm.

	precision	recall	f1-score	support
negative	1.00	1.00	1.00	422
neutral	1.00	0.69	0.82	246
positive	0.81	1.00	0.90	332

accuracy			0.92	1000
macro avg	0.94	0.90	0.90	1000
weighted avg	0.94	0.92	0.92	1000

- Comparison

Both methods worked well, but the ML model's ability to recognize small patterns in the data led it to outperform the simpler approach in cases of more complex sentiment. While VADER rule-based method is faster and more easy to use, it may not be very flexible like ML models of data, especially at a specified domain. Due to its rule based approach, VADER is a more developed and faster to implement solution compared to ML models, although if you have domain-specific data, the latter would probably be more flexible.

CONCLUSION AND FUTURE WORK

This project shows that sentiment analysis on balanced datasets may be accomplished with both lexicon-based and machine learning approaches. Although VADER is a quick and easy method, it has trouble recognizing sentiment that depends on context, including sarcasm and intensifiers. The ML models, especially Logistic Regression, performed better at identifying small sentiment patterns, which makes them more appropriate for complex text analysis. VADER is still a good option, nevertheless, for fast sentiment categorization in situations when domain-specific modification is not necessary.

Future research could look into:

1. ML models are trained on bigger, real-world datasets to evaluate performance and scalability.
2. Creating a hybrid sentiment analysis approach that combines the flexibility of machine learning models with the effectiveness of lexicon-based techniques.
3. Adding context-aware features to lexicon-based techniques, such as domain-specific sentiment dictionaries or sarcasm detection.

REFERENCES

- Prabu Palanisamy, Vineet Yadav, and Harsha Elchuri, "Serendio: Simple and Practical Lexicon-Based Approach to Sentiment Analysis," SemEval 2013.
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.
- Scikit-learn Documentation: Machine Learning in Python: https://scikit-learn.org/
- Python Faker and TextBlob Libraries: https://faker.readthedocs.io/ and https://textblob.readthedocs.io/
- Dataset can be accessed at: https://www.kaggle.com/datasets/jocelyndumlao/twitter-sentiment-analysis-using-roberta-and-vader/data