NLP Project

Sentiment Analysis

Abstract

This project looks at ways to find out what people express in Arabic tweets. We tried two different methods: Machine Learning, and knowledge-based. We cleaned the tweets, turned them into a form the model can read, and then used these methods to classify them. The results help us understand which method works better and how hard it is to analyze feelings in Arabic tweets.

Introduction

Social media is very popular, and people often share their feelings using sites like Twitter. Understanding these feelings automatically, especially in Arabic, can be hard because Arabic is a complex language. This project tries to understand these feelings by seeing if tweets are positive, negative, or neutral.

Background

Before, most research on understanding feelings from text focused on English. Arabic is less studied because it is a complex language. Earlier research used both machine learning (where the computer learns from examples) and rule-based methods (using set rules to decide). We build on this by testing these methods on Arabic tweets.

Approach

Data Preprocessing

Tweets are preprocessed to remove noise and irrelevant information, including URLs, usernames, hashtags, and special characters.

Feature Extraction

We use Count Vectorizer and TF-IDF Vectorizer to transform text data into a format suitable for model input. We experiment with n-gram ranges to capture more contextual information, crucial for understanding the sentiment in languages with rich morphology like Arabic.

Sentiment Classification

Machine Learning Model

We utilized the Multinomial Naive Bayes. This classifier is particularly effective for text classification due to its assumption of independence between features, making it scalable and robust when dealing with high-dimensional data.

Knowledge-Based Approach

The knowledge-based method involves using a lexicon where words are pre-assigned sentiment scores. Sentiment analysis is performed by calculating the sum of the sentiment scores of the words present in a tweet. The final sentiment of the tweet is determined based on the cumulative score: positive if the sum is above zero, negative if below zero, and neutral if it equals zero.

Rule-Based Approach

This approach involves defining a set of predetermined rules to classify sentiments. We constructed rules based on the presence of keywords that are commonly associated with positive or negative sentiments. The classification is determined by scanning the tweet for these keywords and assigning a sentiment based on their occurrence.

Experiments

Dataset

ASTD: Arabic Sentiment Tweets Dataset

This dataset contains over 10k Arabic sentiment tweets classified into four classes subjective positive, subjective negative, subjective mixed, and objective.

Configuration and Metrics

Models were trained using default configurations with adjustments in the vectorization phase (n-grams and max features). The primary evaluation metrics used were accuracy, precision, recall, and F1-score.

Results

Machine Learning-Based Classification Report (Unigrams)

Label	Precision	Recall	F1-score	Support
NEG	0.51	0.19	0.28	333
NEUTRAL	0.35	0.04	0.06	170
OBJ	0.69	0.96	0.80	1276
POS	0.27	0.02	0.04	160
Overall				1939

Label	Precision	Recall	F1-score	Support
Accuracy			0.67	

Machine Learning-Based Classification Report (N-grams)

Label	Precision	Recall	F1-score	Support
NEG	0.63	0.11	0.19	333
NEUTRAL	0.33	0.02	0.03	170
OBJ	0.67	0.98	0.80	1276
POS	0.38	0.02	0.04	160
Overall				1939
Accuracy			0.67	

Machine Learning-Based Classification Report (TF-IDF and Trigrams)

Label	Precision	Recall	F1-score	Support
NEG	0.46	0.25	0.32	333
NEUTRAL	0.32	0.05	0.09	170
OBJ	0.70	0.94	0.80	1276
POS	0.43	0.06	0.11	160
Overall				1939
Accuracy			0.67	

- **Knowledge-Based Sentiment Analysis Accuracy:** 75.02%
- **Rule-Based Accuracy:** 68.90%

Comparison

All ML methods resulted in almost the same accuracy with a little difference in other metrics, while the knowledge-based achieved 75% because a lot of the tweets are marked as OBJ which also means NEUTRAL.

Conclusion

Working with Arabic data can be challenging, it takes hard work to make an efficient methodology, but from the models we build and using knowledge-based we noticed the it’s not impossible.

Team Contribution

Muhannad Faden 2140899

Abdulwahab Almutlak 2141261

Saleh Alqurashi 2140977

Hisham Molawi 2141063