University Of Engineering and Technology Lahore

Title:

Sentiment Analysis Report

Subject:

Natural Language Processing

Submitted to:

Dr. Quratulain

Submitted by:

Mahnoor Imran (2022-CS-726) - A

Introduction

Sentiment analysis is an exciting field within Natural Language Processing (NLP) that helps computers understand emotions or opinions expressed in text. In this project, our goal was to analyze the sentiments (positive or negative) in Urdu sentences taken from YouTube transcripts. Sentiment analysis can be extremely useful for businesses, social media platforms, and content creators to gauge public opinion or feedback, and we aimed to apply this to Urdu text.

Methodology

Here's how we approached the project:

1. **DataCollection:**
   We collected transcripts from a variety of Urdu-language YouTube videos to ensure diversity in sentence structure and sentiment. The goal was to gather an equal number of **positive** and **negative** sentences for a balanced dataset.
2. **Translation:**
   We used **Google Translate** to convert English into Urdu. This helped standardize the dataset for Urdu-based sentiment analysis.
3. **Text File Creation:**
   Each transcript was saved as a separate **.txt** file, including:
   - Sentences
   - Timestamps
   - YouTube video URLs
4. **Data Preprocessing:**
   - Cleaned the text by removing incomplete, noisy, or irrelevant lines.
   - Created a structured **Excel file** using pandas, with columns: Sentence, Timestamp, Video URL.
   - The dataset was labeled manually as either **Positive** or **Negative**.
5. **Sentiment Prediction Using Transformers:**
   Instead of traditional ML models, we used **pre-trained transformer-based models** for sentiment analysis.
6. **Manual Review and Correction:**
   After prediction, we **manually checked** the results for incorrectly labeled sentences.
7. **Evaluation:**
   Using the final labeled dataset (Ground Truth vs. Model Prediction), we calculated:
   - **Accuracy**
   - **Precision**
   - **Recall**

- **F1-Score**
  These metrics were computed using sklearn to evaluate how well the transformer model performed on sentiment classification.

Dataset

- **Source**: YouTube Urdu-language video transcripts.
- **Size**: 40200 sentences, split positive and negative sentences.
- **Preprocessing**:
  - We made sure all sentences were either in Urdu or translated into Urdu.
  - The sentences were then cleaned and tokenized, making them ready for the machine learning models.

The dataset was split into 80% for training the models and 20% for testing them. Each sentence was labeled as either "Positive" or "Negative" based on its sentiment.

Result

Here's how the different models performed:

Model	Accuracy	Precision (N)	Recall (N)	F1-Score (N)	Precision (P)	Recall (P)	F1-Score (P)
Logistic Regression	0.93	0.93	0.94	0.94	0.92	0.91	0.92
Naive Bayes	0.89	0.88	0.94	0.91	0.91	0.84	0.87
SVM	0.94	0.95	0.94	0.95	0.92	0.94	0.93
XGBoost	0.90	0.90	0.93	0.91	0.90	0.90	0.90
LightGBM	0.90	0.90	0.93	0.91	0.91	0.90	0.92

Error Analysis

- Here are examples of misclassified sentences from traditional models:

Sentence	Ground Truth	Predicted
وہ ہم پر بلاوجہ الزام لگا رہے ہیں۔	Negative	Positive
میرے بیٹے کی کوئی بری عادت نہیں ہے	Positive	Negative
انکل، آرکیٹیکٹس نے نرما کے خوابوں کے گھر کو ڈیزائن کرنے کے لیے کام شروع کر دیا ہے۔	Positive	Negative
بہت سارے تحائف۔ -	Positive	Negative
نہیں، نہیں. یہ اچھی بات ہے۔	Positive	Negative

میں آپ کو جنم دینے سے مایوس ہوں۔	Negative	Positive
مجھے مایوس مت کرو۔	Positive	Negative
تم مجھے اس دنیا میں سب سے زیادہ عزیز ہو۔	Positive	Negative
میں ناخوش ہوں۔	Negative	Positive
جب آپ کا شوہر آپ کے ساتھ ہو تو آپ کو کسی سے ڈرنے کی ضرورت نہیں ہے۔	Positive	Negative

Conclusion

Overall, the sentiment analysis models did a great job classifying Urdu text into positive and negative categories, with the SVM model showing the best results. We learned that while machine learning can be very effective, there's still room for improvement especially in handling more complex sentence structures or sarcasm. Future work could involve using more advanced techniques, such as deep learning models, or expanding the dataset to make the model more robust.