BNU

BEACONHOUSE NATIONAL UNIVERSITY

Sentiment Analysis of Roman Urdu Reviews on Daraz using Transformer

PROJECT PROPOSAL

INTERNAL SUPERVISORS

Mr. Usman Nazir

GROUP MEMBERS

Muhammad Ali	F2020-591
M. Shehryar Sohail	F2020-364
Asif Ahmad Chaudhry	F2020-533

SCHOOL OF COMPUTER & IT

April 2024

Introduction:

In the contemporary digital landscape, e-commerce platforms have emerged as vital channels for businesses to engage with consumers to drive sales. Daraz, as a leading e-commerce platform in Pakistan, serves as a hub for a diverse array of products and services, catering to the needs and preferences of millions of users. The volume of user-generated content, particularly in the form of reviews, presents a goldmine of information for businesses setting to understand customer sentiments, preferences, and pain points.

However, the linguistic diversity of the user base, coupled with the prevalence of Roman Urdu as a common medium of communication, prosses unique challenges for sentiment analysis on platforms like Daraz. Roman Urdu, characterized by its blend of Urdu script with Roman script, requires specialized approaches for accurate sentiment interpretation. Furthermore, the informal and colloquial nature of user reviews adds another layer of complexity to sentiment analysis tasks.

This research endeavors to bridge the gap between user-generated content and actionable insights for businesses operating on the Daraz platform through sentiment analysis of Roman Urdu reviews. By harnessing the power of transfomers, we aim to unlock the latent sentiment embedded within user reviews, empowering businesses to make data-driven decisions and enhance customer satisfaction.

Literature Review:

While sentiment analysis has garnered considerable attention in the field of natural language processing, its application to Roman Urdu reviews on e-commerce platforms remains underexplored. Existing literature primarily focuses on sentiment analysis in widely spoken languages, with limited research addressing the nuances of Roman Urdu text.

Studies on sentiment analysis often rely on lexicon-based approaches, machine learning algorithms, or deep learning models trained on large-scale datasets in languages like English. However, the effectiveness of these approaches diminishes when applied to Roman Urdu due to the scarcity of labeled data and linguistic peculiarities.

While transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), have demonstrated remarkable performance in various NLP tasks, their adaptation and optimization for Roman Urdu sentiment analysis remain relatively unexplored. By leveraging the capabilities of these cutting-edge models, this research seeks to bridge the gap between existing methodologies and the unique challenges posed by Roman Urdu text.

Despite these advancements, challenges persist in adapting machine learning models to the unique characteristics of Roman Urdu, especially in the context of e-commerce reviews on platforms like Daraz. Addressing these challenges requires innovative methodologies that integrate linguistic expertise with machine learning techniques to achieve accurate and reliable sentiment analysis.

Open Problems:

1. Roman Urdu, as a fusion of Urdu and English script, encapsulates a rich tapestry of linguistic nuances, cultural references, and contextual subtleties. Sentiment analysis in Roman Urdu demands a sophisticated understanding of these intricacies to accurately capture and interpret user sentiments.
2. The efficacy of sentiment analysis models cannot be confined by a single product category. To ensure broad applicability and relevance the proposed framework must demonstrate scalability and robustness across diverse product domains, ranging from electronics to fashion and beyond.
3. Beyond sentiment classification, the ultimate goal is to derive actionable insights that empower businesses to make informed decisions in real-time. This necessitates not only the accurate analysis of Roman Urdu reviews but also the seamless integration of the sentiment analysis framework into existing e-commerce infrastructure.

Researcher’s Expertise:

1. **Muhammad Ali:** With a background in Python programming and experience in developing neural networks for classification and regression tasks, Muhammad brings a wealth of expertise in machine learning and deep learning methodologies. His proficiency in Python and neural networks will be instrumental in implementing and fine-tuning the sentiment analysis model proposed in this research.
2. **Shehryar Sohail:** As an expert in data analytics, Shehryar brings invaluable insights into the preprocessing and analysis of large datasets. His expertise in data wrangling, feature engineering, and exploratory data analysis will be crucial in preparing the Roman Urdu Daraz review dataset for sentiment analysis and identifying relevant patterns and trends.

3. **Asif Ahmad Chaudhry:** With a strong background in data science and expertise in data visualizations techniques, Asif adds a unique perspective to the research team. His proficiency in visualizing complex data structures and communicating insights effectively will enhance the presentation and interpretation of results obtained from the sentiment anaylsis model.

Problem Statement:

This research endeavors to address the following pivotal challenges:

1. **Fine-grained Sentiment Analysis:** Develop an advanced sentiment analysis framework capable of discerning nuanced sentiments within Roman Urdu Daraz reviews. This involves not only identifying positive, negative, and neutral sentiments but also capturing subtle variations in tone, emotion, and subjective interpretation.
2. **Cross-domain Generalization:** Assess the generalization capabilities of the sentiment analysis model across diverse product categories on Daraz. A model that performs well in one category may not necessarily generalize to others. Therefore, ensuring robust performance and adaptability across different domains is crucial for real-world applications.
3. **Practical Implementation and Deployment:** Devise strategies for the seamless integration of the sentiment analysis model into Daraz's operational framework. This includes considerations such as real-time monitoring, scalability, resource efficiency, and user privacy, ultimately enabling actionable insights generation and decision-making.

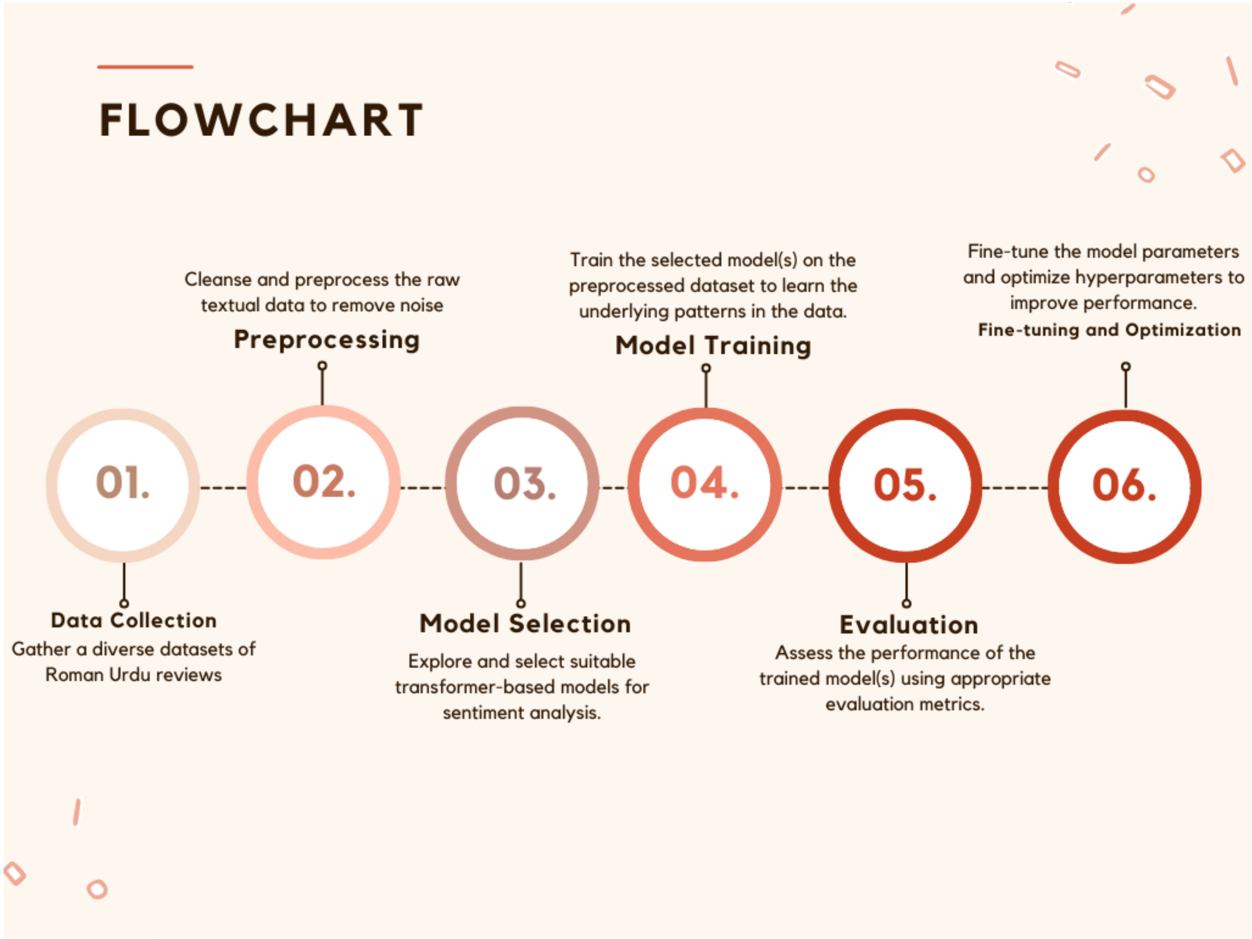
Methodology:

The methodology encompasses a comprehensive approach that integrates linguistic analysis and machine learning techniques:

1. **Comprehensive Data Acquisition**
   - Curate a diverse and extensive dataset comprising Roman Urdu reviews across various product categories on Daraz. The dataset should encompass a wide range of products, user demographics, and review sentiments to ensure representativeness and relevance.
2. **Advanced Preprocessing Techniques**
   - Employ state-of-the-art preprocessing methods to cleanse and refine the textual data, mitigating noise, handling spelling variations, and enhancing the model's ability to extract meaningful sentiment features. This may involve techniques such as tokenization, stemming, lemmatization, and character normalization.

3. **Transformer-based Model Exploration**
    - Investigate the efficacy of transformer-based architectures, such as BERT, RoBERTa, and XLNet, in capturing nuanced sentiments within Roman Urdu text. Experiment with different model variants, fine-tuning strategies, and pretraining objectives to optimize performance for the target task.
4. **Fine-tuning and Optimization Strategies**
    - Employ advanced fine-tuning and hyperparameter optimization techniques to tailor the selected transformer-based model for optimal performance in Roman Urdu sentiment analysis. This involves adjusting model parameters, learning rates, batch sizes, and other hyperparameters to strike a balance between accuracy, efficiency, and generalization.
5. **Cross-domain Evaluation and Validation**
    - Conduct rigorous cross-domain validation experiments to evaluate the model's generalization capabilities across diverse product categories. This entails training and testing the models on datasets representing different product domains, assessing performance metrics, and identifying potential sources of bias or overfitting.
6. **Interpretability & Explainability**
    - Investigate methods for interpreting and explaining the model's predictions, focusing on identifying salient features and linguistic patterns that contribute to sentiment classification decisions. Enhance the transparency and trustworthiness of the model by providing insights into its decision-making process.

Figure A:



FLOWCHART

01. Data Collection

Gather a diverse datasets of Roman Urdu reviews

02. Preprocessing

Cleanse and preprocess the raw textual data to remove noise

03. Model Selection

Explore and select suitable transformer-based models for sentiment analysis.

04. Model Training

Train the selected model(s) on the preprocessed dataset to learn the underlying patterns in the data.

05. Evaluation

Assess the performance of the trained model(s) using appropriate evaluation metrics.

06. Fine-tuning and Optimization

Fine-tune the model parameters and optimize hyperparameters to improve performance.

Figure A shows the representation of major methodology in a flow chart.

Research Plan & Timeline:

1. **Literature Review & Data Collection (Week 1)**
   - Muhammad Ali will focus on identifying relevant literature on sentiment analysis methodologies, particularly in the context of transformer-based models. Shehryar Sohail will lead the effort in collecting a diverse dataset of Roman Urdu reviews from Daraz, ensuring comprehensive coverage across different product categories
2. **Preprocessing & Model Selection (Week 2)**
   - Asif Ahmad will oversee the preprocessing phase, implementing advanced techniques to clean and prepare the raw textual data for analysis. Muhammad Ali and Shehryar Sohail will collaborate on exploring and selecting suitable transformer-based models,

considering factors such as architecture, pretraining objectives, and computational efficiency.

3. **Model Training & Evaluation (Week 3)**
    - Muhammad Ali will take the lead in training the selected transformer-based model(s) on the preprocessed dataset, fine-tuning hyperparameters, and monitoring convergence. Shehryar Sohail will assist in devising evaluation metrics and conducting comprehensive performance evaluations, ensuring robustness and reliability of the models.
4. **Fine-tuning & Optimization (Week 4)**
    - Muhammad Ali and Shehryar Sohail will collaborate on fine-tuning the trained models to optimize performance for Roman Urdu sentiment analysis. Asif Ahmad will provide insights into visualization techniques to aid in understanding model behavior and identifying areas for further improvement.
5. **Cross-Validation & Result Analysis (Week 5)**
    - The entire team will work collaboratively to conduct cross-domain validation experiments to assess the generalization capabilities of the sentiment analysis models across diverse product categories. Muhammad Ali, Shehryar Sohail, and Asif Ahmad will jointly analyze the results, draw actionable insights, and prepare the final research findings for dissemination.
6. **Interpretability and Explainability (Week 6)**
    - The entire team will investigate methods for interpreting and explaining model's predictions, focusing on identifying salient features and linguistic patterns that contribute to sentiment classification decisions. Enhance the transparency and trustworthiness of the model by providing insights into its decision-making process.

Figure B:

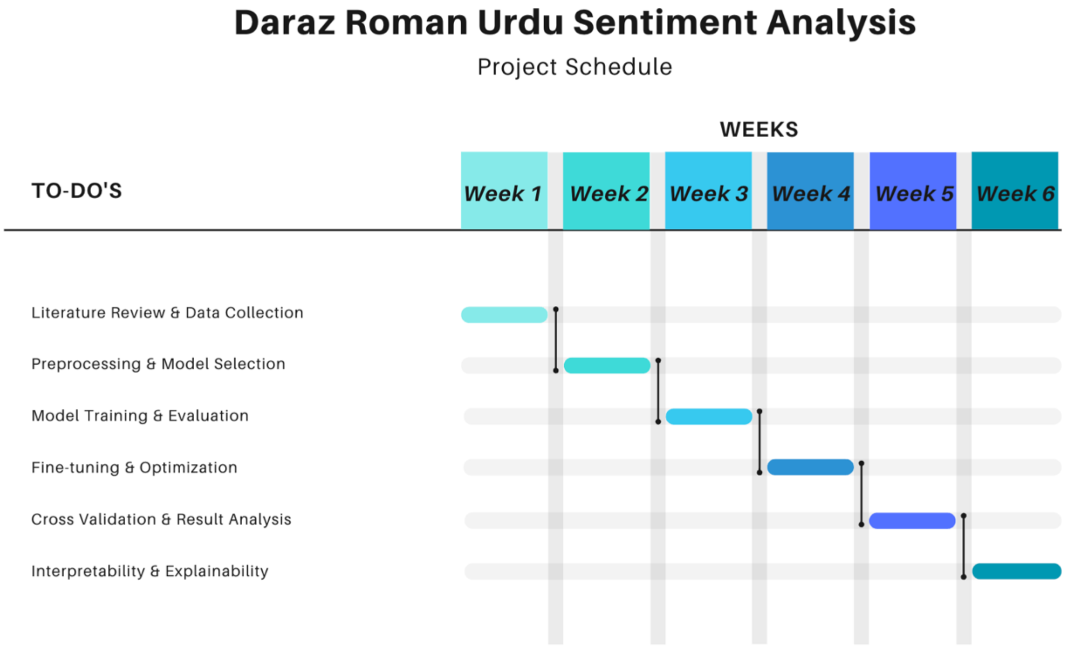


Figure B describes the project schedule in a Gantt Chart format.

References:

https://arxiv.org/pdf/1706.03762.pdf

https://link.springer.com/article/10.1007/s10462-022-10144-1

https://arxiv.org/pdf/1810.04805v2.pdf

https://arxiv.org/pdf/1907.11692v1.pdf

https://arxiv.org/pdf/1801.06146v5.pdf