Project Design: Twitter Sentiment Classification

Jeffrey Clancy
Esteban Felix Tapia
Stepan Ochodek
Alen Pavlovic

Master of Science in Applied Data Science
University of Chicago

Course: Linear and Nonlinear Models for Business Application
Professor: Dr. Utku Pamuksuz
TA: Irem Pamuksuz

August 3, 2024

1 Executive Summary

This project aims to develop a robust sentiment analysis tool using the Sentiment140 dataset, which includes 1.6 million tweets with sentiment labels. The goal is to classify sentiment (positive or negative) in Twitter messages, employing both linear and non-linear models. This tool will help organizations understand sentiment in contexts like product feedback, film recommendations, electoral analysis and many others.

We will utilize machine learning algorithms such as Logistic Regression, SVM, and advanced models like LSTM and BERT (*if time permits*). Additionally, we will implement Explainable AI (XAI) techniques like LIME and SHAP to provide transparency in model decision-making.

Our methodology includes data preprocessing, exploratory data analysis, model training and validation and XAI implementation. We will use cross-validation and independent test sets to evaluate model performance, including the simulated production environment with the "live tweets".

Expected outcomes include a sentiment classification tool. Potential challenges include handling the diversity and noise in Twitter data and ensuring model explainability. A risk management plan will address these issues to ensure model robustness and reliability.

2 Introduction

2.1 Background and Context

Social media platforms like Twitter have become rich sources of real-time information, capturing the diversity of human expression across various contexts. This project leverages the Sentiment140 dataset, a large collection of tweets annotated with sentiment labels, to develop a sentiment analysis tool.

2.2 Problem Statement

The primary challenge is to accurately classify the sentiment of tweets (positive or negative) using a combination of linear and non-linear machine learning models. This classification must be effective even for texts that do not contain explicit emoticons.

2.3 Objectives

The main objectives of this project are:

1. **Develop a Sentiment Analysis Tool**: Create a tool that can classify sentiments in tweets using both traditional and advanced machine learning models.

2. **XAI Techniques**: Use XAI methods like LIME and SHAP to make the model's decision-making process transparent and interpretable.

3. **Evaluate Model Performance**: Use cross-validation and independent test sets to assess the performance of the models.

4. **Causal Analysis**: Apply causal inference techniques to uncover and understand causal relationships within the data.

3 Literature Review

3.1 Related Work

Sentiment analysis has been a important area of research, particularly for applications in product reviews, movie reviews, and electrocal analysis.

The Sentiment140 dataset, introduced by Go, Bhayani, and Huang (2009), leveraged distant supervision by using emoticons as noisy labels for sentiment classification (":)" positive vs ":(" negative). Their study showed that machine learning algorithms could achieve over 80% accuracy in classifying sentiments of tweets, laying the groundwork for further exploration of Twitter data for sentiment analysis and our interest in the topic.

3.2 Data Source Description

The Sentiment140 dataset consists of 1.6 million tweets, each annotated with a sentiment label (positive or negative) based on the presence of emoticons. The dataset was collected using the Twitter API between April 6, 2009, and June 25, 2009. Each tweet in the dataset includes the tweet text, the sentiment label, and other metadata such as the date of the tweet, the query used to collect the tweet, and the user ID.

Key Features of the Dataset:

- **Size**: The dataset contains 1.6 million rows.
- **Variance**: Tweets in the dataset contain a variety of words and expressions, reflecting the diverse nature of Twitter content.
- **Real-World Applicability**: The data was collected directly from Twitter, making it applicable to real-world scenarios.

4 Methodology

4.1 Data

The Sentiment140 dataset will be used as the primary data source for this project. The following steps outline the data preprocessing procedures:

1. **Data Cleaning**: Remove unnecessary characters, duplicate tweets.
2. **Tokenization**: Split the text into individual words or tokens.
3. **Stop Words Removal**: Remove common words that do not contribute to sentiment (e.g., "the", "is").
4. **Feature Extraction**: Convert text data into numerical features using techniques TF-IDF or word embeddings.

4.2 Exploratory Data Analysis

Perform EDA to understand the dataset's structure, including:

1. **Sentiment Distribution**: Analyze the proportion of positive and negative tweets.
2. **Word Frequency Analysis**: Identify the most frequent words and phrases in each sentiment category.
3. **Visualization**: Use visual tools like word clouds, bar charts, and histograms to gain insights.

4.3 Model Development

Develop and train multiple machine learning models to classify tweet sentiments. The models to be implemented include:

1. **Linear Models**: Logistic Regression.
2. **Non-Linear Models**: SVM, Naive Bayes.
3. **Advanced Models**: LSTM, Bidirectional Encoder *(if time permits)*.
4. **Fine-Tuning Pre-Trained Transformer Models**: Fine-tune pre-trained transformer (e.g. BERT).

4.4 Model Training and Validation

1. **Training**: Split the dataset into training and validation sets.
2. **Validation**: Use cross-validation to evaluate model performance and prevent overfitting.
3. **Performance Metrics**: Use metrics such as accuracy, precision, recall, F1 score, and AUC-ROC to assess model performance.

4.5 Explainable AI (XAI) Techniques

Implement XAI techniques to interpret model predictions:

1. **LIME (Local Interpretable Model-Agnostic Explanations)**: Explain individual predictions.
2. **SHAP (SHapley Additive exPlanations)**: Calculate the contribution of each feature to the overall prediction.

5 Roles and Responsibilities

The description of each team member's role and responsibilities:

1. **Jeff Clancy**
   - **Role**: Project Manager/Writer and QA/Testing Specialist
   - **Responsibilities**: Ensure the project is delivered on time, coordinate tasks among team members, draft and compile all written deliverables, prepare the final presentation and report, ensure models are validated correctly, and conduct QA testing to make sure the implementation meets the project requirements.
2. **Esteban Felix Tapia**
   - **Role**: ML Engineer
   - **Responsibilities**: Focus on model development, including training and fine-tuning machine learning models, implementing Explainable AI techniques, handling data preprocessing, performing exploratory data analysis, and assisting in feature engineering and model evaluation.
3. **Stepan Ochodek**
   - **Role**: ML Engineer/Literature Reviewer
   - **Responsibilities**: Focus on model development and researching previously developed sentiment analysis models, conducting comprehensive literature review and method review analysis.
4. **Alen Pavlovic**
   - **Role**: ML Engineer
   - **Responsibilities**: Focus on model development, including conducting exploratory data analysis and feature engineering, assisting in model training and validation, applying causal inference techniques to the dataset, and analyzing and interpreting causal relationships within the data.

6 Expected Outcomes

6.1 Deliverables

The project will result in several key deliverables:

1. **Sentiment Classification Tool**: A robust model capable of classifying sentiments (positive, negative) in tweets and similar short text.

2. **Data, Code and Final Presentation**: A final presentation summarizing the project, methods (XAI and Causal Analysis), results, and insights, along with a final presentation.

6.2 Potential Challenges and Risk Management

1. **Data Diversity and Noise**: Social media data is diverse and noisy, which can affect model performance. To mitigate this, in-depth data cleaning and preprocessing steps will be undertaken.

2. **Model Explainability**: Ensuring the explainability of the models is crucial. XAI techniques will be implemented to provide transparency.

7 Conclusion

This project aims to advance sentiment analysis by developing a robust tool for classifying sentiments in Twitter data using the Sentiment140 dataset. We will implement a combination of machine learning models, including Logistic Regression, SVM, LSTM, and fine-tuned transformer models like BERT, while applying Explainable AI techniques (LIME, SHAP) and causal inference methods to enhance model transparency and provide deeper insights.

Our comprehensive methodology includes data preprocessing, exploratory data analysis, model development, and validation. We will address challenges such as data diversity and class imbalance through careful planning and risk management. The project will deliver a versatile and interpretable sentiment analysis tool, offering valuable insights for organizations to leverage social media data for strategic decision-making.

References

[1] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision.", Stanford (2009).

[2] TensorFlow Datasets: Sentiment140. Available at: `https://www.tensorflow.org/datasets/catalog/sentiment140`