# STATS 507 - Final Project - Proposal

Shenyi Tang

November 2024

## 1 Overview

### 1.1 Motivation

Sentiment analysis has become a valuable tool across numerous industries, playing a critical role in social media listening, email feedback analysis, customer support ticket prioritization, product review mining, and personalized marketing. By extracting customer reviews and comments, businesses can gain insights into customer experiences, predict trends, and make data-driven decisions to improve customer satisfaction and engagement.

I previously interned at a social media listening company that served prominent businesses by delivering periodic reports on customer sentiment regarding products and campaigns. A key metric in these reports was the net sentiment rate, which calculated overall sentiment using a company's proprietary tool. While I utilized pre-existing tools during my internship, my coursework in STATS 507 has motivated me to build my own sentiment analysis model, aiming to achieve performance similar to professional tools in the industry.

### 1.2 Datasets

The dataset consists of user evaluations of 20 different apps, with each evaluation comprising two parts: a user-provided comment and a rating (ranging from 1 to 5). My goal is to develop a sentiment analysis model to classify the comments into three sentiment categories—positive, neutral, and negative—using the ratings as labels. Specifically:

- Ratings of 1 and 2 will be classified as negative.
- Ratings of 3 will be classified as neutral.
- Ratings of 4 and 5 will be classified as positive

### 1.3 Expected Insights

From this project, I aim to gain insights particularly from the model development and performance analysis. First of call, to go over an end-to-end process of data analysis from data pre-processing, model building to result interpretation. Secondly, to develop a sentiment analysis model that performs as well as possible given the data and computational constraints. Last but not least, I hope to strengthen the ability to know the limitations of current approach and propose possible areas and methods (e.g., data quality, feature engineering, model architecture) for future improvements.

## 2 Prior Work

Most literature classify sentiment analysis techniques into 3 categories: Lexicon-Based methods, machine learning methods and deep learning methods. [1]

In machine learning approaches, a SVM classifier [2] could be used to search the optimal hyperplane to separate classes, which performances good effectivity and stability in high dimensional spaces. While Naive Bayes provides a model depends on Bayes Theorem and Bag of Words feature Extraction, it use nodes to represent random and independent variables and edges to represent the relationship, that are used to find relationships among a large

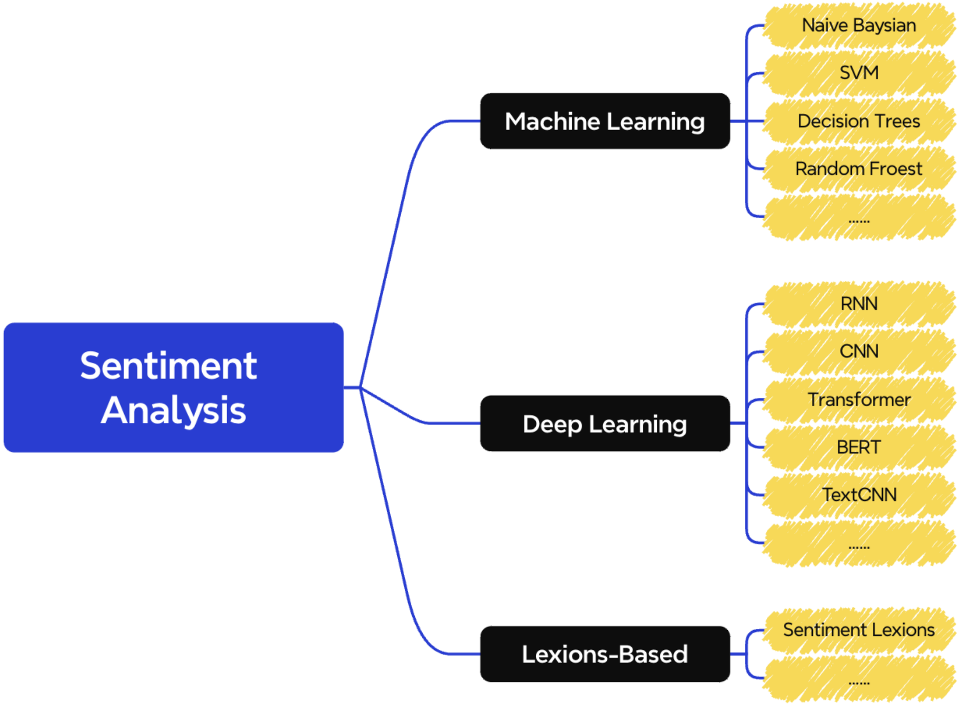

Figure 1: Sentiment Analysis Techniques

number of words. Wan and Gao [3] used Bayesian networks, SVM Decision Tree and Random Forest algorithms, and Bayesian networks performed best.

The Lexicon-Based approach, also known as the knowledge-based approach, is one of the primary methods for sentiment analysis and relies on a predefined list of words called an opinion lexicon. This approach works effectively in avoiding the problem that a word means definitely opposite in different domains. Well-known dictionaries can be used to create sentiment lexicons.

For deep learning methods, CNN can also achieve notable success in NLPs besides its original employment in computer vision.Kim [4] utilized a CNN model on top of the pre-trained word2vec model to do the sentence-level sentiment classification. Li et al. [5] introduced a bidirectional LSTM model to capture the relationship between target words and sentiment polarity words in a sentence without using a sentiment lexicon.

# 3 Preliminary Results

## 3.1 Data Understanding

The dataset consists of user reviews and corresponding ratings (1-5) for 20 different apps. Ratings are used as sentiment labels, with 1-2 mapped to ”negative,” 3 to ”neutral,” and 4-5 to ”positive.” Initial exploration indicates that most reviews fall into the ”positive” category, creating a potential class imbalance issue that could impact model performance. Additionally, user comments often include emojis and informal language, requiring preprocessing steps to handle such elements effectively.

### 3.2 Tools and Models

I plan to use PyTorch to build an LSTM-based model for sentiment analysis. LSTM is well-suited for sequential data like text, as it can capture contextual information across word sequences.

I will use PyTorch in depth, focusing on building and optimizing the LSTM model. I will expore techniques like SMOTE, class weighting in the loss function, or oversampling minority classes to handle imbalance data. Moreover, I plan to learn model optimization to propose possible methods to enhance the model performance.

## 4 Project Deliverables

A successful project will produce a sentiment analysis model capable of classifying app reviews into positive, neutral, or negative categories with reasonable accuracy and robustness. The deliverables will include: (1) a pre-processing pipeline to clean, tokenize and embed text data (2) a trained LSTM model (3) a performance evaluation to assess the model's effectiveness, with insights on strengths and weaknesses.

The sub-goals are to tune hyperparameters and propose possible methods to optimize the model's performance and avoid overfitting.

## References

[1] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[2] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," 1998.

[3] Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in *2015 IEEE international conference on data mining workshop (ICDMW)*. IEEE, 2015, pp. 1318–1325.

[4] Y. Kim, "Convolutional neural networks for sentence classification. arxiv 2014," *arXiv preprint arXiv:1408.5882*, 2019.

[5] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional lstm with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, 2020.