Tweets Sentiment Analysis and Classification

Project Proposal

By

Maryam AlBahri
albahri.m@northeastern.edu

Shan Lin
lin.shan1@northeastern.edu

Min Yao
yao.min1@northeastern.edu

May 25th , 2025

Project Description

Traditional decision-making models in areas like finance, health, and consumer behavior often overlook public sentiment, which can rapidly shift outcomes. Social media platforms like X (formerly Twitter) have become key channels where mass opinion influences real-world events often unpredictably. Viral sentiment, as seen in cases like the GameStop short squeeze in 2021, can override data-driven logic and disrupt markets or public trust. Research shows that social media sentiment acts as a real-time indicator of public mood, yet many predictive systems still fail to incorporate this emotional layer, making them vulnerable to unanticipated shifts.

Integrating sentiment analysis into predictive models enhances their robustness and adaptability across domains. Neural NLP techniques are particularly well-suited for capturing nuanced sentiment in short, informal social media text, offering deeper contextual understanding and detecting subtle emotional cues that traditional methods often overlook.

This project aims to build a neural NLP-based sentiment classifier for tweets, using models like LSTM and BERT. While finance is used as an example, the approach is broadly applicable to any domain sensitive to real-time public sentiment. Commercial sentiment tools are often costly and opaque, and third-party datasets may introduce bias. Developing a custom model ensures control, transparency, and adaptability making it a reliable tool for sentiment-driven analysis in high-stakes environments.

Problem Statement

This project aims to identify the most effective modeling approach for short, informal text like tweets. Using the large-scale Sentiment140 dataset, we will evaluate neural architectures such as LSTM and BERT to determine which best captures sentiment in social media language. The chosen model will be fine-tuned to build a general-purpose sentiment classifier, with potential applications in financial sentiment analysis, social media monitoring, brand reputation tracking, and public feedback systems.

In pursuit of this goal, the project seeks to answer the following key questions:

- Which model performs best on tweet sentiment classification?
- What trade-offs exist between model complexity and performance?
- How well do different models generalize to unseen or real-world tweets?
- Can the selected model be effectively applied in real-world domains such as finance, brand monitoring, or public feedback systems?

Dataset Selection

The dataset used in this project is Sentiment140, sourced from Kaggle and originally developed by researchers at Stanford University.

➤ Source URL: Sentiment140 Dataset on Kaggle

➤ Reference: Go, A., Bhayani, R., & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Stanford University.

➤ Description: The Sentiment140 dataset contains 1.6 million tweets annotated for sentiment using distant supervision based on emoticons. Tweets are labeled with one of three sentiment classes:

- **0 = negative**
- **2 = neutral**
- **4 = positive**

➤ Dataset Language: All tweets are in English, having been filtered by language during collection by the authors.

➤ File Structure

The dataset includes the following fields:

Column	Description
target	Sentiment label (0 = negative, 2 = neutral, 4 = positive)
id	Unique tweet ID
date	Date and time the tweet was posted
flag	Query status always set to "NO_QUERY" (legacy field, can be ignored)
user	Username of the tweet author
text	Raw tweet text

Assumptions

1. **Label Integrity:**
   We assume that the sentiment labels provided in the Sentiment140 dataset represent ground truth. Our model is trained to replicate or generalize these labels, not to reverse-engineer or infer the labeling logic used by the original annotation system.
2. **Independence from Source Bias:**
   We assume that any potential bias embedded in the original labeling process is not being explicitly learned or mimicked by our model. The objective of this project is to learn sentiment patterns from labeled examples, without attempting to decode or replicate the decision-making heuristics of the dataset creators.

3. **Supervised Learning Framework:**
   The model is developed under a supervised learning framework where we treat the labels as fixed targets. We do not attempt to validate or question their correctness during training, and we assume no access to the original labeling algorithm or its internal logic.
4. **Generalization Focus:**
   The goal of this study is to identify and train the most effective model architecture for classifying sentiment in short texts like tweets, rather than to critique or deconstruct the annotation methodology of the dataset.

Approach & Methodology

Our approach begins with a thorough analysis of the dataset and follows a structured pipeline to develop, compare, and deploy the most effective sentiment classification model for short text data such as tweets. The methodology involves following main stages:

1. Exploratory Data Analysis (EDA)
We'll start with an exploratory data analysis to examine class distribution, tweet length, vocabulary richness, and noise (e.g., URLs, hashtags, emoticons), helping us identify imbalances and patterns that could affect model performance.

2. Data Preprocessing
Preprocessing will involve tools such as NLTK and custom scripts to clean the data by applying lowercasing, punctuation removal, tokenization, stop word removal, lemmatization, and eliminating user mentions, hashtags, and URLs. This prepares clean, consistent input for classical and neural models.

3. Data Splitting and Validation Approach
We'll apply a stratified 80/20 train-test split to maintain label balance and use methods like 10-fold cross-validation during training to promote robust generalization and minimize the effects of random data variance.

4. Feature Engineering and Model Training
We will train and compare three modeling approaches, progressing from traditional to advanced:

- Baseline Model: A Logistic Regression model with TF-IDF vectorization, serving as a fast, lightweight benchmark, though limited in capturing context.

- LSTM Model: A Bidirectional LSTM leveraging pre-trained embeddings (e.g., GloVe) to learn temporal and contextual sentiment in tweets.
- As an exploratory step, depending on resources, we'll explore fine-tuning a transformer model like DistilBERT to assess performance vs. resource trade-offs.

Models will be implemented using appropriate frameworks: For example, Logistic Regression using scikit-learn, LSTM in TensorFlow/Keras, and transformers (if used) via Hugging Face Transformers.

5. Model Evaluation and Recommendations

To identify the most effective model for sentiment classification, we will compare candidate models Logistic Regression, LSTM, and, if feasible, DistilBERT using validation metrics such as accuracy, precision, recall, and F1-score. Ablation analysis will be conducted to explore the impact of factors such as input length, vocabulary coverage, and misclassification patterns.

In addition to validation, we may optionally assess generalization by testing the selected model on a small external set of unseen tweets, if time and resources permit. This would provide preliminary insight into the model's robustness in handling real-world sentiment variation.

Model selection will weigh both performance and practical considerations, including computational cost, generalization capability, and deployment feasibility. Based on these results, we will recommend the most suitable approach for real-world sentiment analysis applications in domains such as finance, brand monitoring, or public feedback systems. This evaluation aims to directly address the core questions posed in the problem statement concerning model effectiveness, trade-offs in complexity, and the model's adaptability to dynamic, informal text data.

Expected Outcomes

Comparative Baseline:

- Use a traditional approach (e.g., TF-IDF with logistic regression) as a baseline for comparison.
- Benchmark performance improvements from more advanced models, including LSTM and transformers. (Prior studies suggest transformers may outperform traditional models by 10–15% on similar datasets, though actual performance will depend on tuning and domain factors.).

Model Training and Validation:

- Train the best-performing model on the Sentiment140 dataset.
- Validate against a held-out test set using accuracy and F1-score.
- Ensure consistent performance across sentiment classes.

Model Selection:

- Identify the most suitable architecture for sentiment analysis of short, informal texts.
- Compare LSTM and BERT-based models in terms of handling brevity, context, and informal language.

Functional Use Potential:

- Explore the applicability of the sentiment classifier to use cases such as financial sentiment analysis, customer feedback tracking, and social media monitoring.
- Where feasible, demonstrate basic inference or batch sentiment classification on example data to illustrate practical relevance.

Plan for Next Steps

Weeks 3–4: Feedback & Refinement

- Review instructor feedback and revise project scope if needed.
- Conduct exploratory data analysis (EDA) to understand class distribution, tweet characteristics, and potential preprocessing needs.
- Set up a GitHub repository and define team workflow.
- Begin reviewing relevant NLP methods (e.g., TF-IDF, LSTM, BERT).
- Start data preparation: clean, tokenize, and preprocess the dataset.

Weeks 5–9: Model Development

- Finalize model selection and conduct preliminary tests.
- Implement and train models using TF-IDF as a baseline and a transformer-based approach (e.g., BERT).
- Track performance using metrics such as accuracy and F1-score.
- Maintain a well-documented GitHub repository for collaboration and submission.

Week 10-12: Model Evaluation

- Evaluate final models on test data using standard metrics.
- Analyze results to identify strengths, limitations, and areas for improvement.
- Optionally, perform error analysis and suggest refinements.

Week 13-14: Findings and Final Report

- Finalize the best-performing model and supporting documentation.
- Complete the final report and submit all required deliverables.

Future Extensions & Exploration

Our custom sentiment classifier can serve as a foundational component for more complex machine learning pipelines across various application domains. Future extensions may include:

- **Predictive Modeling**: Use sentiment scores as input features in models for time series forecasting or risk analysis, e.g., predicting stock volatility or transaction volume based on public opinion trends.
- **Domain Applications**:
    - **Finance**: Enhance stock market analysis where traditional indicators fall short.
    - **Brand Monitoring & Public Health**: Improve early warning systems, consumer behavior prediction, or public response prioritization.
- **Technical Enhancements**:
    - Integrate with real-time streaming data (e.g., via Twitter API).
    - Apply domain-specific tuning to adapt the model to different contexts.

These extensions would strengthen the model's utility for real-world, decision-support systems across industries.

References

1. Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
2. Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I Hope it is not as Bad as I Fear". *Procedia - Social and Behavioral Sciences*, 26, 55–62.
3. Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
4. Cavalcante, R. C., Brasileiro, R. C., Souza, V. L. F., Nobrega, J. P., & Oliveira, A. L. I. (2016). Computational intelligence and financial markets: A survey and future directions. *Expert Systems with Applications*, 55, 194–211.
5. Go, A., Bhayani, R., & Huang, L. (2009). *Twitter Sentiment Classification using Distant Supervision*. Stanford University.
6. *GameStop short squeeze*. (n.d.). Wikipedia. https://en.wikipedia.org/wiki/GameStop_short_squeeze