Project Proposal

CS2.302 Computational Linguistics 2

Submission by
Naman Singhal(2024114013), Shrish Kadam(2024114015)

Project P13
Sentiment Analysis on Social Media Texts with Semantic Interpretation

GitHub Repository: `https://github.com/the-neemon/CL-2-Project`

Date: November 3, 2025

1. Project Title

Sentiment Analysis on Social Media Texts with Semantic Interpretation Using Lexicon-Based and Machine Learning Approaches

2. Research Question

How can combining lexicon-based semantic features with traditional machine learning approaches improve sentiment classification accuracy on Twitter data and what role do semantic interpretation techniques play in handling context-dependent sentiment expressions in social media text?

3. Methods

3.1. Data Preprocessing

- Text cleaning: Remove URLs, mentions, hashtags and special characters using NLTK
- Tokenization, normalization, stopword removal and lemmatization
- Emoji and emoticon handling to preserve sentiment context

3.2. Feature Engineering

Each model will first be evaluated using the full feature set, after which the best-performing model based on maximum F1 score will be tested with each feature separately to perform feature ablation

- **Contextual features**: Negation patterns and intensifier word detection
- **Semantic embeddings**: Pre-trained Word2Vec or GloVe embeddings for similarity analysis
- **Traditional features**: N-grams and POS tagging for sentiment-bearing words

3.3. Classification and Evaluation

- **Models**: Naive Bayes, Logistic Regression, Random Forest
- **Comparative analysis**: Performance evaluation with and without semantic features
- **Metrics**: Accuracy, Precision, Recall, F1-score with cross-validation
- **Error analysis**: Classification patterns and semantic limitation assessment
- **Success Analysis**: Identification of correctly classified instances and exploration of factors contributing to high model performance
- **Qualitative analysis**: Manual inspection of representative tweets to interpret semantic influences on model predictions and contextual sentiment understanding

4. Datasets

- **Primary**: Sentiment140 dataset (50,000-100,000 tweets subset) with positive, negative, neutral labels
- **Secondary**: Twitter US Airline Sentiment dataset (14,000 tweets) for cross-domain validation
- **Processing**: 80-20 train-test split with class balancing for imbalance handling

5. Key Research Paper References

1. **Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014).** "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1555-1565. [Sentiment-specific word embeddings integrating semantic information for Twitter analysis]
2. **Mohammad, S. M., Kiritchenko, S., & Zhu, X. (2013).** "NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets." *Proceedings of SemEval-2013*. [Feature engineering methodology for Twitter sentiment analysis]
3. **Pak, A., & Paroubek, P. (2010).** "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *LREC'10*. [Foundational Twitter preprocessing and analysis techniques]

Implementation and Deliverables

This project will be implemented on the GitHub repository: `https://github.com/the-neemon/CL-2-Project`. Deliverables include comparative performance analysis demonstrating semantic feature effectiveness, feature ablation studies, detailed error analysis and complete code with documentation. The project leverages established libraries (NLTK, scikit-learn etc.) ensuring feasibility while contributing insights into semantic interpretation's role in social media sentiment analysis.