PA1: Linear Models & Text Classification

Behrad Rezaie
COMP 550, McGill University, 261102374

***Abstract*—This project explores the classification performance of linear models (Logistic Regression, LinearSVC) in semantic and morphology differentiation.**

***Index Terms*—Linear Models, Sentiment, Morphology**

I. Introduction & Problem Setup

Text classification based on semantics and word morphology is an important task in natural language processing. This project evaluates the applicability of linear models such as Logistic Regression and Linear SVC for the task of differentiating text by semantics and by text morphology. Specifically, the two following experiments were performed: the classification of sentences as positive/negative sentiments, and the detection of plurals in sentences.

For the sentiment classification, the hypothesis was that the model would perform slightly better than a random guessing, limited by its inability of forming complex connections with features to deeply understand underlying semantic meanings. For plurals detection in a sentence, the hypothesis is that performance will be much higher, since it is a much easier task given the presence of certain suffixes strongly indicates the presence of plurals.

II. Dataset Generation

LLM tools were used to generate 200 sentences for each category of Negative Comment, Positive Comment, Sentence without plurals, and Sentence with at least one plural. Each experiment's categories were combined, randomized, and split into training and test sets via a 90/10 split, as the train set would be used as the validation set during cross-validation.

III. Preprocessing & Feature Extraction

A number of preprocessing steps were experimented with, including the lowercasing of text, removal of punctuation, removal of common english stop words, stemming, and lemmatization. Results were recorded with and without the application of each of these methods, individually and collectively.

To extract features, the **TF-IDF-Vectorizer** tool from the scikit-learn library was used. This identifies and extracts relevant n-grams from a given corpus based off their computed TF-IDF score.

For sentiment detection, the n-gram extractor was configured to extract word-level unigrams and bigrams. Additionally, a manually added feature representing the count of "negative" words such as "bad, no, not, n't" in each sentence was also provided to the model.

For plural detection, the vectorizer was configured to work with character-level unigrams and bigrams, which extracted meaningful combinations of 1-2 character groupings within the corpus.

IV. Model Training & Parameter Exploration

Logistic Regression and LinearSVC models were trained on the dataset, with a grid search across their different regularization methods and strengths, through a 5-fold cross validation process.

V. Results and Conclusions

Across all experiments, both logistic regression and linearSVC performed similarly, with negligible differences in performance.

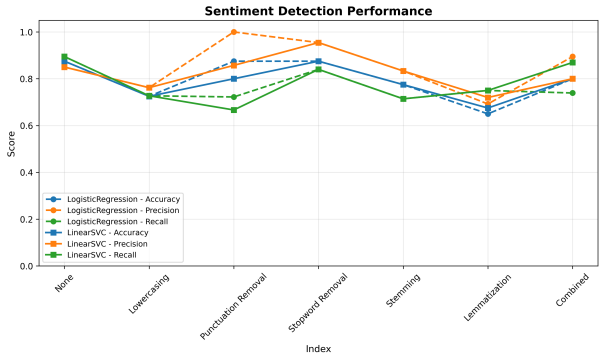

Fig. 1. Sentiment Classification Metrics vs Preprocessing Techniques

As seen in figures 1 and 2, performances differed significantly depending on the preprocessing steps applied, with little variance between the two types of models. Performance was measured with no preprocessing methods applied, one method applied in isolation at a time, and all methods applied simultaneously. Across the 7 experiments, sentiment detection accuracy hovered

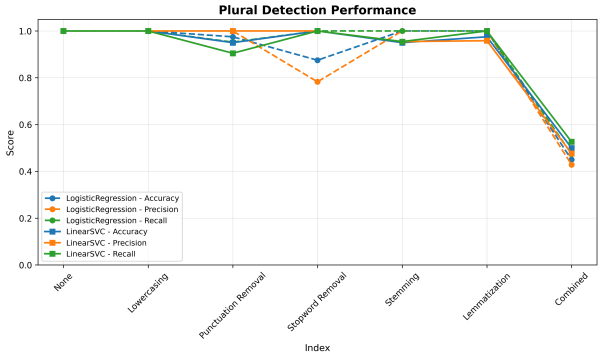

Fig. 2. Plural Detection Metrics vs Preprocessing Techniques

around 75-80%, reaching a maximum of 87.5% when stopwords were removed, and a minimum of 65% with lemmatization applied. This indicates the model performed best when words without much semantic content were removed from the inputs, but suffered significantly when words were lemmatized into their canonical form, indicating these words lost too much information.

For plural detection, performance was at 90-100% for most experiments, dropping significantly down to 50% only when all preprocessing methods were applied simultaneously. This indicates the loss of too much information, likely due to the removal of key parts of words due to stemming and lemmatization, resulting in loss of information that would hint at the presence of plurals.

The weights of the features contributing the most to each classification was explored as well. In the experiment with the best sentiment detection performance, terms such as *"unpleas","neg experi","poorli"* contributed the most to negative classification, whereas *"pleasant","exceed","posit experi"* contributed most to positive classifications. For plural detection, character n-grams such as *"re","s ","ese"* were the strongest indicators of plurals. This can be tied to the model learning the plural forms of words, as well as verb conjugation, such as *"are"* indicating the subject to be in plural form.

VI. Limitations of Experiment

This experiment relied entirely on LLM-generated data, which introduces bias in sentence structure and formation that might not be representative of real-world data. Additionally, the scope of this experiment is limited to the exploration of only two linear models for relatively simple classification tasks. Their applicability on tasks requiring deeper language semantic understanding would be significantly more limited, and more complex architectures would have to be explored. Finally, more powerful text processing techniques and features, such as Part-Of-Speech tagging, were not explored. Their addition into these experiments could yield interest results and enable deeper understanding of text, even for simple linear models.