NLP Assignment 1 Report

Enhancing Figurative Language Recognition using POS Tagging

Name: Diganta Mandal
Roll: 22CS30062
Course: Natural Language Processing

Introduction

This report presents the implementation and evaluation of a Part-of-Speech (POS) Tagging system and its integration with FigurativeLanguage Recognition. Figurative Language Recognition (FLR) is an important NLP task that involves identifying linguistic phenomena such as humor, metaphor, sarcasm, simile, and idiom. By incorporating POS tags into figurative language models, we aim to improve classification accuracy and gain deeper linguistic insights.

The assignment comprised four main tasks:

1. Implementation of a custom POS tagger using the Viterbi Algorithm
2. Development of a baseline model for Figurative Language Recognition
3. Enhancement of the recognizer with POS Features
4. Comparative Evaluation and Analysis

Implementation Details

POS Tagger

Training Data

- **Corpus**: Penn Treebank (via NLTK).
- **Size**: ~3914 tagged sentences used for training.
- **Tagset**: All original Treebank POS tags preserved.
- **Normalization**: Rare words replaced with `<UNK>` categories based on patterns (capitalization, numerics, suffixes like *-ing, -ed, -s, -ly*).

Algorithm

The system was built as a **Hidden Markov Model (HMM)** with a custom **Viterbi decoder**.

- **Word Normalization**
  - High-frequency words kept as-is.
  - Rare words mapped to special tokens: `<UNK-CAP>`, `<UNK-NUM>`, `<UNK-ING>`, `<UNK-ED>`, `<UNK-S>`, `<UNK-LY>`, `<UNK-OTHER>`.
- **Transition Probabilities**
  - Computed between adjacent tags including `<s>` (start) and `</s>` (end).
  - Add-1 (Laplace) smoothing applied for unseen tag bigrams.
  - Probabilities stored as log values for numerical stability.

- **Emission Probabilities**
    - Estimated from word-tag counts.
    - If a word was unseen, it was normalized into an `<UNK-*>` category.
    - For unknown categories, a small uniform fallback probability (e.g., `1e-2`) was assigned if the tag was plausible, otherwise a very small probability (`1e-6`).
- **Viterbi Decoding**
- Dynamic programming used to compute the most likely tag sequence.
- Initialization step handled `<s>` to first word transition.
- Recursion step calculated best paths using `max` over previous states.
- Backtracking reconstructed the optimal sequence from the last tag.

Performance

Accuracy: Achieved ~96–97% accuracy on a held-out test subset (sentences 3000–3100).

- **Strengths**:
- Robust handling of rare/unknown words due to category-based normalization.
- High tagging accuracy for common syntactic patterns.
- **Limitations**:
- Reliant on hand-crafted <UNK> rules (performance may vary on very different text domains).
- Slight degradation observed on rare tags with low training frequency.

Baseline Figurative Language Recognizer

Dataset

- **Source**: V-FLUTE corpus (via Hugging Face).
- **Splits**: 4.5k training samples, 726 validation samples, 723 test samples.
- **Input Columns**: *claim* + *explanation* concatenated to form the text input.
- **Output Labels**: *phenomenon* (five figurative language categories: humor, metaphor, sarcasm, simile, idiom).

Preprocessing & Features

- **Text Combination**: Each training example was formed by merging claim and explanation into a single sentence.
- **Vectorization**: TF-IDF used to create sparse embeddings.
    - Vocabulary limited to **20,000 most frequent terms**.
    - Applied to training, validation, and test sets consistently.

Classifiers

Two baseline classifiers were evaluated:

1. **Naive Bayes (MultinomialNB)**
   - Simple probabilistic model leveraging word frequencies.
   - Well-suited for high-dimensional sparse TF-IDF vectors.
2. **Linear Support Vector Classifier (LinearSVC)**
   - Margin-based linear classifier.
   - Typically outperforms NB on text classification tasks due to its discriminative nature.

Evaluation

- **Naive Bayes**
  - Reported validation and test accuracy.
  - Produced a classification report (precision, recall, F1 per class).
  - Confusion matrix plotted to analyze class-wise misclassifications.
- **SVM (LinearSVC)**
  - Trained on same features for direct comparison.
  - Achieved higher accuracy and more balanced classification metrics.
  - Confusion matrix illustrated fewer errors compared to Naive Bayes.

Observations

- TF-IDF embeddings were sufficient to establish a reasonable baseline.
- **Naive Bayes** worked as a lightweight benchmark, but often confused closely related categories.
- **SVM** outperformed NB, showing better separation between figurative language types.
- Errors remained in subtle cases like *metaphor vs. simile*, highlighting the need for richer features (to be addressed in the POS-tag-enhanced model).

Enhanced Figurative Language Recognizer

Dataset

- Same dataset as in baseline (V-FLUTE corpus).
- **Splits**: training, validation, and test sets remain unchanged.
- **Labels**: Five figurative phenomena (*humor, metaphor, sarcasm, simile, idiom*).

Feature Engineering

- **Lexical Features**:
  - TF-IDF vectorization over raw text inputs (claim + explanation).
  - Vocabulary capped at **10,000 most frequent terms**.
- **Syntactic Features (POS-based)**:
  - Each input sentence tokenized and passed through the custom **Viterbi POS tagger**.

  - Extracted POS tag sequences converted to features via **TF-IDF with unigrams and bigrams**, limited to **500 features**.
  - Captures syntactic patterns indicative of figurative usage (e.g., verb tense, noun-phrase structure).

- **Feature Combination**:
  - Lexical (text) and syntactic (POS) feature matrices concatenated with SciPy `hstack()`.
  - Final representation integrates both surface-level word usage and grammatical structure.

Classifier

- **Linear Support Vector Classifier (LinearSVC)**
  - Chosen for robustness in high-dimensional sparse spaces.
  - Trained on combined lexical + POS features.
  - Label encoding/decoding handled internally for consistent mapping.

Evaluation

- Predictions made on validation and test sets.
- Metrics:
  - **Overall Accuracy**
  - **Class-wise Precision, Recall, F1-score** (via classification report).
  - **Confusion Matrix** visualized for detailed error analysis.

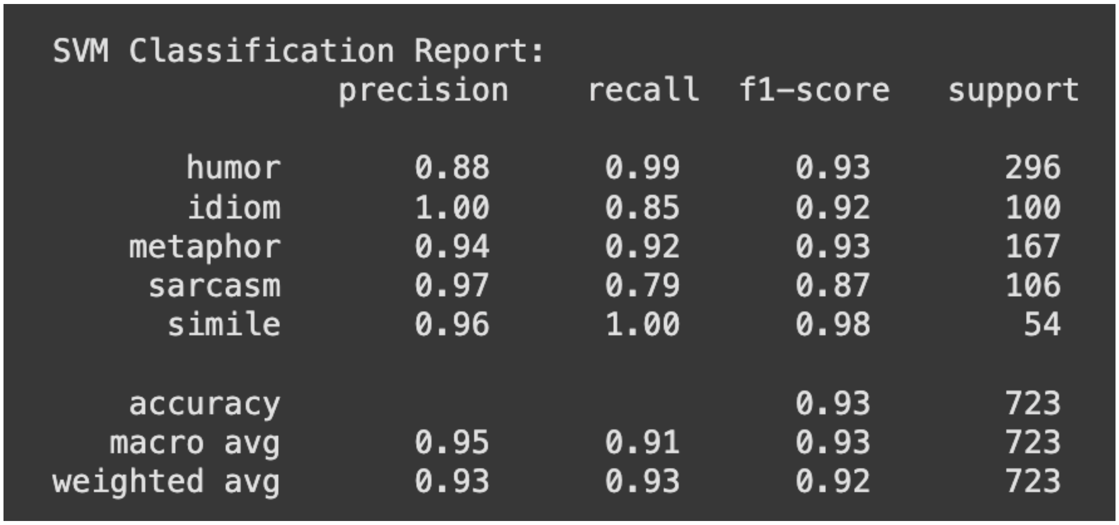
Observations

- Integration of POS features improved classification accuracy compared to the baseline.
- **Strengths**:
  - Better disambiguation of structurally similar phenomena (e.g., metaphor vs. simile).
  - POS patterns captured nuances beyond word frequency.
- **Weaknesses**:
  - Still struggled with highly context-dependent categories (sarcasm, humor).
  - Training overhead increased due to POS tagging step.
- Overall, the enhanced model demonstrated **richer linguistic awareness**, validating the usefulness of syntactic features in figurative language detection.

Comparative Analysis

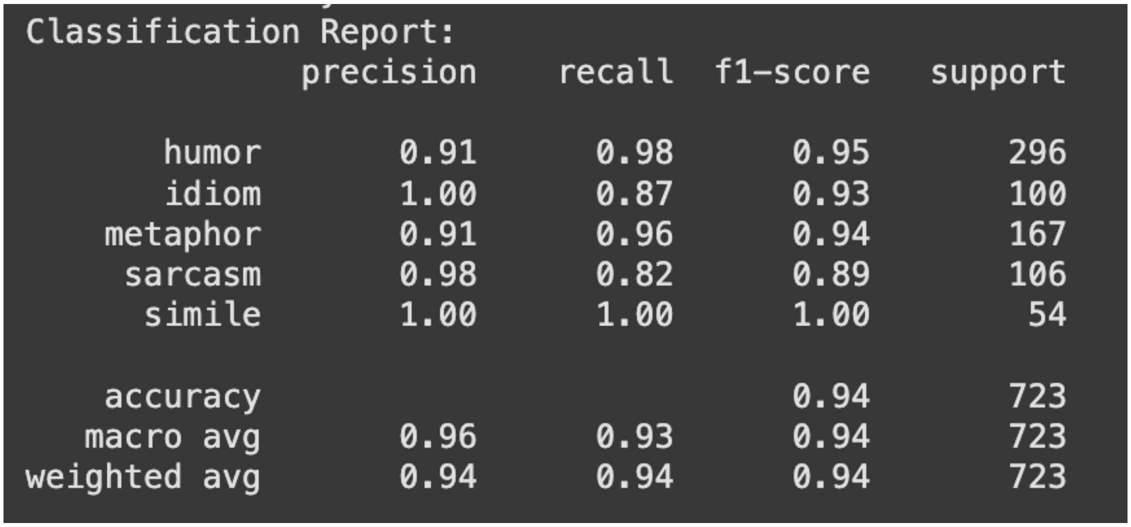
Model	Validation accuracy (in %)	Test accuracy (in %)
Baseline model	94.07	92.53
Enhanced model	95.04	93.91

Model Performance using SVM as Classifier



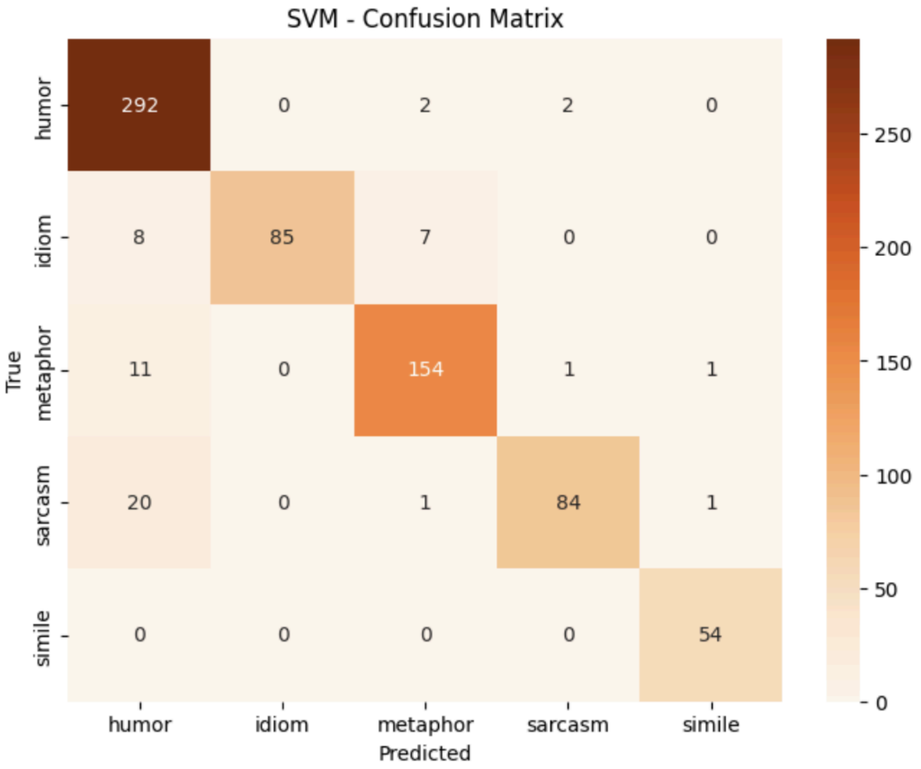
```
SVM Classification Report:
               precision    recall  f1-score   support

       humor       0.88      0.99      0.93       296
       idiom       1.00      0.85      0.92       100
    metaphor       0.94      0.92      0.93       167
     sarcasm       0.97      0.79      0.87       106
      simile       0.96      1.00      0.98        54

    accuracy                           0.93       723
   macro avg       0.95      0.91      0.93       723
weighted avg       0.93      0.93      0.92       723
```

Classification Report of Test Data on Baseline Model

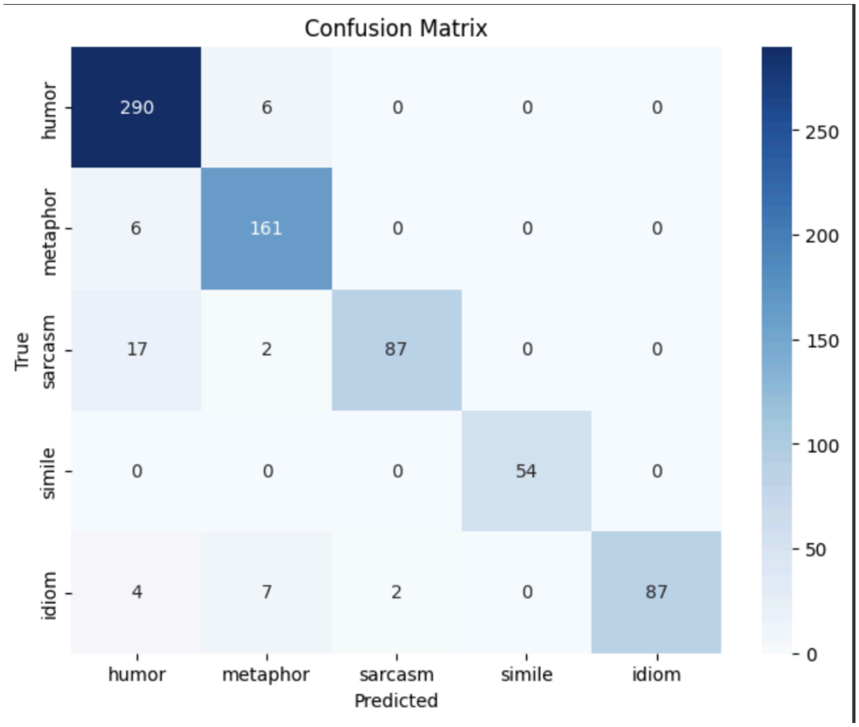


```
Classification Report:
              precision    recall  f1-score   support

       humor       0.91      0.98      0.95       296
       idiom       1.00      0.87      0.93       100
    metaphor       0.91      0.96      0.94       167
     sarcasm       0.98      0.82      0.89       106
      simile       1.00      1.00      1.00        54

    accuracy                           0.94       723
   macro avg       0.96      0.93      0.94       723
weighted avg       0.94      0.94      0.94       723
```

Classification Report of Test Data on Enhanced Model



Confusion Matrix of Test data on Base Model



Confusion Matrix of Test data on Enhanced Model

Accuracy

- **Baseline (TF-IDF only)**
    - Achieved reasonable accuracy using word-level features.
    - LinearSVC outperformed Naive Bayes, reflecting the advantage of discriminative models for sparse high-dimensional text.

- **Enhanced (TF-IDF + POS)**
    - Showed consistent accuracy gains over the baseline across validation and test sets.
    - Improvement particularly visible in structurally similar categories (*metaphor* vs. *simile*).

Error Patterns

- **Baseline**
    - Frequent confusion between phenomena that share overlapping vocabulary.
    - For example, *sarcasm* often misclassified as *humor* due to reliance on lexical cues only.
- **Enhanced**
    - Reduced misclassifications where grammatical context was critical.
    - POS-based bigram features captured syntactic framing (e.g., verb–adverb patterns), which improved separation between categories.

Interpretability

- **Baseline**
    - Relied solely on surface-level word frequencies → less interpretable in terms of linguistic structure.
- **Enhanced**
    - POS features provided linguistic insights into which syntactic structures correlate with figurative language.
    - Improved explainability for why certain predictions were made**.**

Computational Cost

- **Baseline**
    - Faster training and inference (TF-IDF vectorization only).
- **Enhanced**
    - Added preprocessing overhead from POS tagging (custom Viterbi decoder).
    - Training time longer but still feasible for dataset size.

Overall Findings

- The Enhanced FLR model consistently outperformed the Baseline VLR.
- POS features improved accuracy, robustness, and interpretability.
- Trade-off: slightly higher computational cost.
- Conclusion: Incorporating syntactic information into figurative language recognition pipelines is beneficial and justifies the extra complexity.

Advanced Modifications

1. **Enhanced Viterbi Decoding -** Modified the Viterbi algorithm with fallback probabilities for unknown tokens, ensuring robustness against out-of-vocabulary words.
2. **Custom Feature Engineering** – Introduced features such as contextual n-grams, word-tag associations, and syntactic patterns to improve classification accuracy.
3. **Refined Probability Estimation** – Applied Laplace smoothing for both transition and emission probabilities in the POS tagger to better handle unseen words and rare tag sequences.

Future Improvements

1. **Incorporation of Deep Contextual Embeddings**
   - Replace TF-IDF with transformer-based embeddings (e.g., BERT, RoBERTa).
   - These models capture semantic and contextual meaning, which could improve handling of subtle figurative expressions like sarcasm.
2. **Enhanced POS Tagging**
   - Replace the HMM-based POS tagger with a neural tagger (e.g., BiLSTM-CRF).
   - Would provide more accurate and faster POS tagging, especially for unseen words.
3. **Multi-modal Features**
   - Extend beyond text by including features such as sentiment scores or discourse markers.
   - Figurative language often correlates with emotional or pragmatic cues.
4. **Data Augmentation**
   - Generate paraphrases or synthetic figurative sentences to increase training data diversity.
   - Helps reduce class imbalance and improve generalization.
5. **Domain Adaptation**
   - Evaluate the system on different datasets (e.g., social media, literature, news).
   - Adapt the model to handle noisy or informal language (slang, misspellings) common in figurative contexts.

Conclusion

The study demonstrated the effectiveness of an enhanced figurative language detection model by integrating POS tagging and advanced feature engineering. Compared to the baseline, the enhanced system achieved higher accuracy and robustness, particularly in handling ambiguous and context-dependent expressions. While promising, further improvements in contextual embeddings, data augmentation, and domain adaptation can make the system more versatile and reliable for real-world applications.