Final Report

“Automated Romantic Quote Classification Using a Sentence-scoring Mechanism”

Introduction & Motivation

This project focuses on the extraction and classification of romantic quotes using classic natural language processing (NLP) techniques. The system is designed to identify romantic quotes from a given dataset and assign a romantic score based on 4 key metrics. This classification system is intended to aid in curating romantic content for applications such as literature analysis.

The primary motivation behind this project stems from the need to effectively flag sentences from text that can serve as potential romantic quotes as a fun way of textual analysis. With the vast amount of textual data available on the internet, manual classification becomes impractical. Automating this process not only saves time but also ensures consistency and accuracy in identifying romantic quotes. This project aims to develop an algorithm that can distinguish romantic content from non-romantic content with high precision, recall, and F1 score, making it useful for various real-world applications.

Related Work

The paper “*DirectQuote: A Dataset for Direct Quotation Extraction and Attribution in News Articles*” [1] has a similar goal in which they extract and attribute quotes from news articles to identify credible experts for fact-checking. Utilizes a BERT-based Question Answering model. model to attribute quotes using semantic role labeling and other NLP techniques.

Another example is *“Journalism AI – Quotes extraction for modular journalism”* [2] which makes use of a trained SpaCy model using techniques such as NER for quote extraction. Other methodologies include the use of sentiment lexicons, such as the Opinion Lexicon, and keyword-based extraction for text categorization. Studies like “*Quote Attribution in Literary Works”* [3] focus on attributing quotes to characters in literature. These models often combine syntactic parsing with semantic analysis to determine the speaker and context.However, little to no work was found on romantic text classification as quotes.

System Description

The system for extracting and classifying romantic quotes is designed with multiple interrelated components. Here is a detailed description of each part of the system.

1. **Data Splitting and Preprocessing**:

The data is first split into a train set with 43188 quotes (romantic only) and a test set with 12957 quotes(all genres). The test and train text data is first converted to lowercase to help reduce analysis complexity by treating words like "Love" and "love" as the same. Sentences are tokenized using the `word_tokenize` functions and stopwords (such as "and", "the", "is") are removed. Words are reduced to their base or root form using the `WordNetLemmatizer` from NLTK.

2. **Keyword Extraction**:

The system extracts keywords from the training dataset by counting the frequency of each word after preprocessing. Words that appear more frequently in romantic quotes are considered significant keywords. A `Counter` object from the `collections` module is used to tally the occurrences of each word. Words that are identified as romantic keywords are those with higher frequencies in the training set.

3. **Scoring Mechanism**:

Equation: 0.5K + 0.3S + 0.1R + 0.1D

Each sentence in the test set is scored based on the combined scores of the 4 metrics. These 4 metrics were selected because their unique combination came together to deliver the best results. Weights have been assigned to the metrics as shown above;

The `score_sentence_by_keywords` function calculates the score by summing the frequencies of individual words, bigrams, and trigrams from the keyword frequency dictionary. Sentiment analysis is performed using the Opinion Lexicon from NLTK. Additional positive and negative words specific to romantic contexts are included. Each word in the sentence is checked against these lexicons, and the sentiment score is calculated by summing positive word counts and subtracting negative word counts. The readability of a sentence is evaluated using the Flesch-Kincaid readability formula, which considers the number of words, sentences, and syllables. A higher readability score indicates easier readability.

Lexical diversity score is calculated by dividing the number of unique words in a sentence by the total numbThe sentiment_score function calculates the sentiment of a sentence by analyzing the presence of positive and negative words. It uses pre-defined sets from the opinion_lexicon and extends them with additional romantic and negative words. The function tokenizes the sentence and scores it by incrementing for positive words and decrementing for negative ones.er of words. Higher lexical diversity indicates a richer vocabulary.

Each sentence is assigned a combined score using a weighted sum of the keyword frequency score, sentiment score, readability score, and lexical diversity score. The weights for these components are predetermined based on their importance (e.g., keyword weight: 0.5, sentiment weight: 0.3, readability weight: 0.1, lexical diversity weight: 0.1). The threshold for classification is determined from the training set's score distribution. Sentences with combined scores above this threshold are classified as romantic, while those below are classified as non-romantic.

4. **Evaluation**:

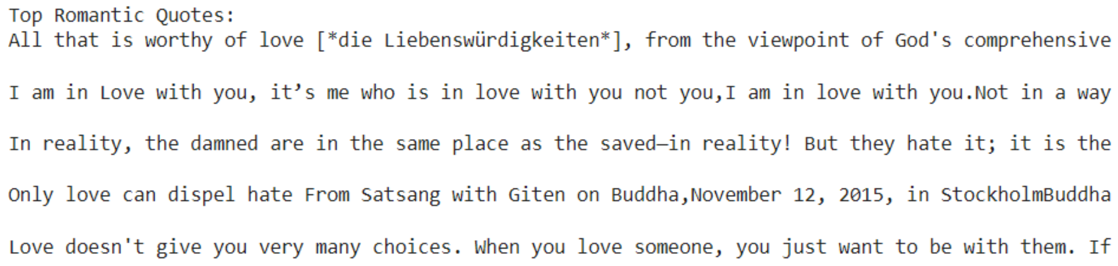
The system's performance is evaluated using the test set. Actual labels (romantic or non-romantic) are compared against predicted labels.

Precision: 86.7 %	**Recall**: 82.8 %
F1 score: 84.7	**True Positives (TP):** 789; **False Positives (FP):** 1370 **False Negatives (FN):** 1854; **True Negatives (TN):** 8944

Table 1 illustrates the evaluation metrics.

User Interaction

For easier evaluation the current system gets a test set as input to classify quotes as romantic or non romantic. However the user can also input a piece of text or a novel from which they can get scored sentences based on romantic appeal. They can then select the top quotes from the list of the highest scoring sentences.



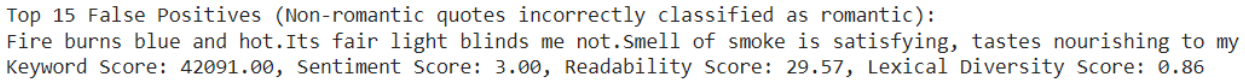
```
Top Romantic Quotes:
All that is worthy of love [*die Liebenswürdigkeiten*], from the viewpoint of God's comprehensive

I am in Love with you, it’s me who is in love with you not you,I am in love with you.Not in a way

In reality, the damned are in the same place as the saved—in reality! But they hate it; it is the

Only love can dispel hate From Satsang with Giten on Buddha,November 12, 2015, in StockholmBuddha

Love doesn't give you very many choices. When you love someone, you just want to be with them. If
```

Figure 1 shows example output.

The user interface allows for easy input of new datasets and retrieval of classification results. There is a url variable where the user can put the link to any book from the GutenBerg Corpora. Additionally there is a text variable as well which can take any kind of english text as input and score the sentences, outputting the highest scoring sentences in the end.

Discussion

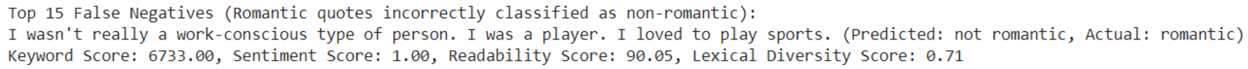
As observed from Table 1, the system demonstrates high precision and recall in identifying romantic quotes. The F1 score is also decent. However, there are instances of false positives and false negatives, which highlight the complexity of natural language and the subjective nature of romance. Analyzing the individual key metric scores of the edge cases reveals that in the case of FPs the model's reliance on keyword frequency and sentiment analysis leads to misclassification when keywords appear in non-romantic contexts.



```
Top 15 False Positives (Non-romantic quotes incorrectly classified as romantic):
Fire burns blue and hot.Its fair light blinds me not.Smell of smoke is satisfying, tastes nourishing to my
Keyword Score: 42091.00, Sentiment Score: 3.00, Readability Score: 29.57, Lexical Diversity Score: 0.86
```

Figure 2 illustrates how individual scores for FPs are printed.

FNs demonstrate high readability and lexical diversity scores showing they are well-constructed sentences but might not have enough romantic keywords or lexical diversity to get high enough scores. In both these scenarios refining the sentiment analysis or incorporating more contextual understanding could help improve results significantly.



```
Top 15 False Negatives (Romantic quotes incorrectly classified as non-romantic):
I wasn't really a work-conscious type of person. I was a player. I loved to play sports. (Predicted: not romantic, Actual: romantic)
Keyword Score: 6733.00, Sentiment Score: 1.00, Readability Score: 90.05, Lexical Diversity Score: 0.71
```

Figure 3 illustrates how individual scores for FNs are printed.

There are also some interesting outputs observed. For example, inputting pride and Prejudice from the GutenBerg Corpora extracts these quotes from the novel among others.

- *"I have been used to consider poetry as the food of love,"*
- *"A lady's imagination is very rapid; it jumps from admiration to love, from love to matrimony, in a moment."*
- *"Such a change in a man of so much pride excited not only astonishment but gratitude--for to love, ardent love, it must be attributed; and, as such, its impression on her was of a sort to be encouraged, as by no means unpleasing, though it could not be exactly defined."*

Each of these quotes are quite famous, indicating the system has a satisfactory approach so far. One area of interest is how closely the quotes that are famous worldwide from novels compare to those that the system outputs from the same text. Hidden patterns can be uncovered by analyzing why other notable quotes from Pride and Prejudice received lower scores compared to these.

Conclusion

This project presents a robust framework for extracting and classifying romantic quotes using NLP techniques. The system achieves a balanced combination of precision, recall, and F1 score, making it a valuable tool for applications requiring romantic text analysis. Future work involves incorporating more advanced language models (e.g., BERT, GPT), expanding the dataset, contextual analysis and exploring other additional features to enhance classification accuracy. This system is relevant to the course as it integrates key concepts from NLP and text analysis. It demonstrates practical application of theoretical knowledge in developing a functional system that addresses a real-world problem using a classic rule based approach rather than making use of Machine Learning algorithms. Thus the project aligns with the course objectives of understanding and applying traditional NLP algorithms for text processing and classification.

Author: Mahnoor Shafiq

References

[1] https://arxiv.org/abs/2110.07827
[2] https://github.com/JournalismAI-2021-Quotes/quote-extraction
[3] https://aclanthology.org/2022.lrec-1.628.pdf