**1.Introduction and motivation**

**1.1 Introduction**

Poetry is a form of artistic expression, capturing the essence of human emotions and experiences. Despite its increasing popularity, this field remains largely uncharted by computational analysis. While existing classification methods have shown us efficiency in processing text but there are some unique challenges presented by poetry for effective analysis. This project explains the realm of poem classification, utilizing advanced data analysis techniques such as preprocessing text, word embedding, Support Vector Classification (SVC), logistic regression, TF-IDF, word2vec, matplotlib, pie chart to analysis the data.

**1.2 Motivation**

The motivation behind this project is to infuse technology into the observation of traditional poetic forms and capabilities offered by modern computational methodologies. As technology advances the chance to unlock the patterns and meaning within poetry becomes promising. As other studies have shown us the power of computational analysis this project mainly focuses on exclusively based on genre-based classification. This focus on genre-based classification helps us understand each poem and its categorization.

**2. Data set Description**

The dataset for this analysis is sourced from the Kaggle and its owners are poetry foundation, an organization established in 2003 with the goal “to discover and celebrate the best poetry and to place it before the largest audience”. The foundation’s mission is deeply rooted in promoting a diverse range of poetry and its extensive website serves as a repository of thousands of poems. The corpus we deal with is mainly focused on three genres containing love, mythology and folklore, nature. This dataset has 509 unique values, each one of them with its author and its subjects. Each poem in our dataset will be tokenized and used for different kinds of tasks.

A graph showing different colored squares

Description automatically generated

Fig 1

**3. Research Questions**

* How does the size of the dataset impact the performance of different text representation techniques and classifier algorithms in poetry genre classification?
* What role does class imbalance play in the accuracy of genre classification, especially for underrepresented genres like "Mythology & Folklore"?
* In what ways can the performance of Word2Vec embeddings be improved for poetry genre classification, considering the unique challenges posed by the context-dependent nature of poetic language?
* Can the incorporation of more advanced semantic analysis techniques or pre-trained language models enhance the classification accuracy, particularly for genres with complex thematic elements in poetry?
* What impact does the choice of genre categories (Love, Mythology & Folklore, Nature) have on the overall performance of the classification models, and how generalizable are these findings to other poetic genres?
* To what extent does the performance of the TF-IDF technique in poetry genre classification align with its strengths in capturing thematic features and document-level analysis?
* How does parameters in word2vec model effect accuracy of genre classification in poetry?

**4. Literature Review**

The literature review in the context of poetry genre classification and computational analysis reveals a burgeoning interest in merging traditional literary studies with advanced technological methodologies. Despite the increasing popularity of computational approaches in text analysis, poetry remains a relatively unexplored domain. Existing studies have primarily focused on prose, leaving a gap in the understanding of the unique challenges posed by poetic language.

Prior research in text classification has demonstrated the efficiency of methods such as Support Vector Classification (SVC), logistic regression, TF-IDF, and word embedding techniques like Word2Vec and GloVe. However, their application to poetry presents distinctive challenges due to the nuanced and context-dependent nature of poetic expression. This project addresses these challenges by investigating how different text representation techniques and classifier algorithms perform in the context of poetry genre classification.

The choice of dataset from the Poetry Foundation, a prominent organization dedicated to promoting diverse poetry, adds credibility to the study. The dataset, comprising genres such as love, mythology and folklore, and nature, is used to explore the impact of dataset size on classification performance and the role of class imbalance, especially for underrepresented genres.

The research questions posed in the project reflect a comprehensive inquiry into the intricacies of poetry genre classification, considering factors like the effectiveness of Word2Vec embeddings, the influence of genre categories on model performance, and the potential enhancement of accuracy through advanced semantic analysis techniques or pre-trained language models.

The project's methodology involves thorough text preprocessing, data cleaning, tokenization, and various text representation techniques, including TF-IDF and word embedding methods. Two major classifiers, logistic regression and SVC, are employed for evaluation. The literature review sets the stage for the methods used, emphasizing the need for careful consideration of the unique characteristics of poetic language.

Overall, the literature review contextualizes the project within the evolving landscape of computational analysis in poetry, showcasing the existing gaps, methods employed in related studies, and the project's contribution to advancing the understanding of genre classification in this distinct literary form.

**5.Methods and Implementation**

**5.1 Text preprocessing**

Text preprocessing is a fundamental step in preparing the dataset for analysis. In the context of our project on poem genre classification, the poem foundation dataset was subjected to a series of preprocessing steps such as stop word removal to enhance the quality of data.

**5.2 Data cleaning**

The raw dataset underwent a cleaning process to get rid of irrelevant characters, symbols and artifacts. This step is aimed to ensure that the analysis is focused on the essential linguistic content of the poem.

**5.3 Tokenization:**

Tokenization involved breaking down the cleaned poems into individual words or tokens. The NLTK library was utilized to implement this process, creating a structures representation of the text data for further analysis.

**5.4 Text representation:**

**5.4.1 TF-IDF (term frequency – inverse document frequency)**

The tokenized poems were transformed into numerical vectors using TF-IDF presentation. This technique gives weights to each term based on its frequency in a particular poem relative to its occurrence across all poems. The resulting TF-IDF matrices provided a a numerical foundation for understanding the importance of words within the dataset.

**5.4.2 Word embedding techniques:**

**5.4.2.1Word2vec:**

Word2vec is an advanced word embedding technique that was employed to understand semantic relationship between words. This technique converts word into vectors in high dimensional space which helps the model to understand the meaning of the words. The genism library helps us to use this technique for our dataset.

**5.4.2.1 GloVe(global vectors for word representation )**

GloVe, another powerful word embedding, was utilized to generate vector representation for words based on global statical information. The pre trained GloVe embedding was leveraged to capture semantic connections and contextual connections within the poetry.

**5.5 Training and testing**

The prepared datasets are represented in various forms such as TF IDF matrices, word2vec embedding and GloVe embeddings, and were split into training and testing sets. These partitions allowed for the evaluation of model performance on unseen data.

**5.6 Classifier algorithms**

Two major classifiers were used in this project such as logistic regression and support vector classifier (SVC).

**5.6.1 Logistic regression**logistic regression, a linear classification algorithm, was applies to predict poems and their genres based on the transformed textual features. The Sckit-learn library was utilized to implementation of logistic regression models.

**5.6.2 SVC (support vector classifier)**

Support vector classifier is a power non liner classification algorithm used to capture intricate patterns within textual data. The scikit – learn library was utilized to implement and train svc models.

**5.7 Evaluation Metrics:**

The performances of each combination of text representation and classifiers algorithm was evaluated using accuracy scores. The accuracy scores provided a quantitative measure of the model’s ability to correctly predict poem genre on the testing data.

**6. Result and analysis:**

The result, expressed in terms of accuracy scores, were analyzed to identify the most effective combination of text representation and classifier algorithm for poem genre classification. The finding offers the limitation of different methods, guiding the computational approaches in the domain of poem analysis. Below I’m going to show multiple combinations of text representation and classifier from the highest accuracy to the lowest one.

**(TF-IDF + logistic regression)**

Accuracy: 0.77

Classification Report:

precision recall f1-score support

Love 0.80 0.93 0.86 76

Mythology & Folklore 0.67 0.15 0.25 13

Nature 0.65 0.58 0.61 26

accuracy 0.77 115

macro avg 0.71 0.55 0.57 115

weighted avg 0.75 0.77 0.74 115

**TF-IDF + SVC**

Accuracy: 0.76

Classification Report:

precision recall f1-score support

Love 0.82 0.89 0.86 76

Mythology & Folklore 0.67 0.15 0.25 13

Nature 0.59 0.65 0.62 26

accuracy 0.76 115

macro avg 0.69 0.57 0.57 115

weighted avg 0.75 0.76 0.73 115

**Word2Vec + logistic regression**

Accuracy: 0.76

precision recall f1-score support

Love 0.82 0.89 0.86 76

Mythology & Folklore 0.50 0.08 0.13 13

Nature 0.60 0.69 0.64 26

accuracy 0.76 115

macro avg 0.64 0.55 0.54 115

weighted avg 0.73 0.76 0.73 115

**Word2Vec + SVC**

Accuracy: 0.75

precision recall f1-score support

Love 0.81 0.88 0.84 76

Mythology & Folklore 0.67 0.15 0.25 13

Nature 0.59 0.65 0.62 26

accuracy 0.75 115

macro avg 0.69 0.56 0.57 115

weighted avg 0.74 0.75 0.72 115

**Trained model with poem dataset Word2Vec+ SVC**

Accuracy: 0.70

precision recall f1-score support

Love 0.70 0.84 0.76 62

Mythology & Folklore 0.00 0.00 0.00 8

Nature 0.68 0.62 0.65 45

accuracy 0.70 115

macro avg 0.46 0.49 0.47 115

weighted avg 0.65 0.70 0.67 115

**Trained model with poem dataset word2vec+ logistic regression**

Accuracy: 0.70

precision recall f1-score support

Love 0.81 0.80 0.81 76

Mythology & Folklore 0.00 0.00 0.00 13

Nature 0.42 0.65 0.52 26

accuracy 0.68 115

macro avg 0.41 0.49 0.44 115

weighted avg 0.63 0.68 0.65 115

**Analysis:**

The best combination of text representation and classifier for our project is TF-IDF + logistic regression which achieved an accuracy of 77% in which the model excelled in classifying poems under the ‘love’ gene demonstrating high precision. Recall, and F1 scores.

As we know word2vec is a better text representation compared to TF-IDF but the result above shows something else the performance of TF-IDF over Word2Vec can lead to various factors, each influencing the effectiveness of the model.

**Spare Representation**

TF-IDF generates a sparse representation of the text, giving the importance of unique words in each document. In poetry, specific words carry significant thematic weight, TF-IDF could capture these distinctive features effectively.

**Document-level Analysis**

Poetry often relies on individual words to convey nuanced meaning. TF-IDF works by focusing on terms frequency within each document. This might give TF-IDF better performing text representation in this kind of dataset.

**Meaning in Poetry**

Word2vec captures semantic relationships between words, it may be challenging in poetry where words often have context-dependent meanings. TF-IDF focus on individual words might be more suitable for the unique characteristics of poetic language.

As we can see we used two different classifiers logistic regression and support vector classifier (SVC) the performance of logistic regression was better in most of the combinations compared to SVC. These can be the factors that excelled logistic regression over SVC.

**Dataset size**

The dataset we used is relatively small compared to other datasets. Logistic regression usually tends to perform well on smaller datasets. SVC on other hand is a complex algorithm and may require large dataset to learn patterns.

It's crucial to note that the performance of different models depends on the specific characteristics of your dataset, the nature of the problem, and the intricacies of the genre classification task in poetry. Experimenting with hyperparameter tuning, exploring different algorithms, or considering ensemble methods could further refine your model's performance.

**Trained model with poem dataset word2vec+ SVC**

Accuracy: 0.70

precision recall f1-score support

Love 0.70 0.84 0.76 62

Mythology & Folklore 0.00 0.00 0.00 8

Nature 0.68 0.62 0.65 45

accuracy 0.70 115

macro avg 0.46 0.49 0.47 115

weighted avg 0.65 0.70 0.67 115

In this combination we can see we have mythology and folklore has achieved precision, recall, F1score as 0 this means the model did not classify any instances belonging to this class. This usually occurs when there are no true positives (TP) for a specific class.

**Potential reason for this scenario:**

**Insufficient data**

The class "Mythology & Folklore" may be underrepresented in your dataset. With only 8 instances, the model might not have had enough examples to learn and generalize the distinguishing features of this genre effectively as you can observe from figure 1.

**Class imbalance**

Class imbalance occurs when one class has significantly fewer instances compared to others. In your case, if "Mythology & Folklore" has fewer examples, the model might struggle to learn the characteristics specific to this class, leading to poor performance.

**Feature representation**

Word2Vec embeddings are learned based on the context of words in the dataset. If the words associated with "Mythology & Folklore" are not well-represented in the embedding space, the model may struggle to identify them during classification.

This can be improved by collecting more data for “mythology and Folklore” class to provide the model with a more diverse set of examples for training. We can use more techniques to improve class imbalance. By addressing these solutions, the model accuracy can be improved drastically.

**7.Conclusion**

In the pursuit of automating the classification of poetry genres, this project explored various text representation techniques and classifier algorithms. We also faced different challenges in the interaction of computational analysis and literary expression. The analysis, utilizing TF-IDF, Word2Vec, Logistic regression, and Support Vector Classifier (SVC), provided valuable insights into the strengths and limitations of different model configurations:

**Key Findings:**

1. TF-IDF, paired with Logistic Regression, emerged as the most successful combination, achieving the highest accuracy of 77%. The sparse representation and document-level focus of TF-IDF appeared particularly effective in capturing genre-specific features in the poetry dataset.
2. The project faced challenges in accurately classifying genres with limited representation in the dataset, notably "Mythology & Folklore." Insufficient data for this class contributed to a lack of successful predictions, underscoring the importance of addressing class imbalances for robust model performance.
3. Word2Vec, while a powerful technique, exhibited variable performance, especially in classifying less-represented genres. The intricacies of poetic language and the context-dependent nature of word meanings presented challenges for embedding models.

**Improvements for future works:**

1. Improving the dataset with additional examples, especially for underrepresented genres, could enhance the model's ability to improve performance.
2. Further fine-tuning of model hyperparameters, especially for the SVC, could potentially lead to improved classification results. Exploring a broader range of parameters and kernel functions may reveal optimal configurations.
3. Exploring advanced semantic analysis techniques or pre-trained language models could capture more subtle nuances in poetic language, potentially improving the classification of genres with complex thematic elements.

This project shows the relationship between computational algorithms and the analysis of poetic expressions. The challenges of classifying genre in poetry, coupled with the need for diverse datasets, highlight the nature of literature analysis. As technology and methodology continues to advance, the fusion of computational linguistic and literary studies presents exciting possibilities for complexities of poetic expression and genre classification. This project serves as advancement in exploration of future endeavors in this literature type of computational analysis.