Daniel Adam, I.D. 34247563
B.Sc. in Computer Science with specialization in Data Science
Topics in Natural Language Processing
Ben Gurion University of the Negev

Stylometry Approach Detecting Writing Style Changes in Poetry Text

Abstract

This study investigates the evolution of Maya Angelou's writing styles throughout her career using stylometric analysis and word embeddings with BERT. Examining her poetry from 1971 to 1983 and comparing it with her works from 1990 to 2014, feature extraction techniques and unsupervised clustering (K-means machine learning) were employed to identify distinctive writing styles. While thematic shifts are evident, the analysis suggests that the core stylistic elements of Angelou's poetry remained consistent.

Introduction

Maya Angelou is a revered figure in American literature, with her poetry rooted deeply in the African American experience, exploring themes of adversity, resilience, identity, and empowerment. Her works, including "I Know Why the Caged Bird Sings" and "Still I Rise," reflect the struggles and triumphs of African Americans. This study analyses the stylistic developments in Angelou's poetry, examining whether her writing style remained consistent or underwent changes over decades.

Thesis Statement

Preliminary analysis indicates that, despite thematic transformations reflecting Maya Angelou's personal growth and experience due to natural aging, the fundamental stylistic elements of her writing maintained consist of.

Corpus

The analysis utilized selected poems by Maya Angelou, categorized into a training corpus (poems from 1971-1983) and a testing corpus (poems from 1990-2014). The training corpus includes poems from her earlier, while the testing corpus features later more single works. The division facilitates a comparative analysis of Angelou's stylistic choices across her career. The corpus contains overall 173 poems, where 134 poems belong to the periods of 1971-1983 of the train corpus aka initial state and 39 poems belong to the period 1990-2014 of the test corpus aka final state. Thus, the learning is imbalanced in size.

Methodology

<u>Data Preprocessing:</u>
Dividing Maya Angelou's poems into training and testing corpora, with tokenization and removal of stop words via the Natural Language Toolkit (NLTK).

Feature Extraction:
Extracted 11 stylometry features from each poem creating an unlabelled feature matrix for train and test corpus.

1. Basic Text Features:
   Document length, Mean Sentence Length, Mean Word Length, Readability
2.
3. Lexical Usage:
   Lexical richness, Function word frequencies, Content word frequencies

4. Semantic Analysis:
   Semantic repetition

5. Punctuation Usage:

6. Sentiment Analysis:
   Sentiment polarity (negative, neutral, and positive), Sentiment strength

Analysis Tools:
Using various Python libraries for processing and modelling:

- NLTK: for foundational text processing tasks.
- BERT: for semantic repetition, capable of identifying meaning repetition beyond word-level analysis like ngrams.
  Also used for sentiment analysis and sentiment strength.
- Scikit-learn: for K-Means clustering and Principal Component Analysis (PCA)
- Gensim: for topic modelling, concentrating on parts of speech analysis.

Clustering and PCA:
Unsupervised learning was applied to a feature matrix representing poems, with PCA reducing dimensions and K-Means clustering identifying stylistic clusters.
Elbow Method was used to choose the optimal k.

Visualization:
Visualization techniques were used to compare feature distributions across corpora, with each feature individually plotted using Matplotlib.

Results

Feature Distributions Results

The analysis of the poetic works of Maya Angelou from two distinct periods revealed several trends in the textual features:

1. Document Length: (Fig.1)
   The distribution of document lengths showed a notable overlap between the train and test sets, with a slight trend toward shorter documents in the later works.

2. Mean Sentence Length: (Fig.2)
   The mean sentence length exhibited a concentrated range with a similar distribution for both sets. However, the later works (test set) displayed a minor shift towards brevity in sentence construction.

3. Mean Word Length: (Fig.3)
   The mean word length distributions were closely aligned, suggesting a consistent preference for word choice across the two periods.

4. Readability: (Fig.4)
   The readability scores across the train and test sets displayed a high degree of similarity, with both distributions peaking in the same range. This suggests that the complexity level of Angelou's poetry, in terms of reader accessibility, remained consistent.

5. Lexical Richness: (Fig.5)
   A comparison of lexical richness revealed a modest increase in diversity in Angelou's later poetry, as indicated by a rightward shift in the distribution for the test set.

6. Semantic Repetition: (Fig.6)
   Analysis of semantic repetition indicates a decrease in the test set, suggesting a diversification in the use of semantic content in Angelou's later work.

7. Function Words Frequency: (Fig. 8)
   The function words frequency analysis presents a similar shape for both the train and test sets, suggesting a consistent use of function words throughout Angelou's poetry. This aspect of her writing style appears to be a hallmark of her work, largely unchanged over the decades analysed.

8. Content Words Frequency: (Fig.9)
   The frequency distribution of content words between the train and test sets shows that Angelou's use of content-rich words is relatively stable over time, with a slight shift towards more frequent use in her later works. The distributions overlap considerably, indicating that while the subject matter may have evolved, the density of content words within her poems remained similar.

9. Punctuation Usage: (Fig.7)
   The analysis showed a similar pattern in the use of punctuation between the train and test sets, with a minor increase in punctuation variety in the later works.

10. Sentiment Indicator: (Fig.10)
    The sentiment analysis demonstrated a significant presence of neutral sentiment in

both sets. Notably, the later works (test set) show a slightly higher occurrence of positive sentiments.

11. Sentiment Strength: (Fig.11)
    In terms of sentiment strength, the later works tend to exhibit a reduced intensity, with the distribution skewed towards milder sentiment strength.

Clustering Results

Elbow Method for optimal k: (Fig.12)
The plot shows a notable elbow at k=3, suggesting that this is the optimal number of clusters for our dataset, providing a balance between the total variance explained and the number of clusters used.

Train and Test Data Clustering: (Fig.13)
Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space for visualizing the clustering results. The PCA scatter plots for both the train and test data indicate distinct groupings of poems, with the train set demonstrating tighter clustering, possibly signifying a more pronounced distinction between styles within this period. The test set, while also showing clusters, has them more dispersed, suggesting a potential broadening of style in Angelou's later work.

Most Common Non-stop Words Results

Train Corpus: (Fig.14)

Year	Collection	Two most common words
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part One - Where Love is a Scream of Anguish	love, home
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two - Just Before the World Ends	pow, like
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part One	em, pickin
1975	"Oh Pray My Wings Are Gonna Fit Me Well — Part Two	nobody, alone
1975	"Oh Pray My Wings Are Gonna Fit Me Well — Part Three	lain, thus
1975	"Oh Pray My Wings Are Gonna Fit Me Well — Part Four	bad, honky

1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Five	see, children
1978	And Still I Rise — Part One - Touch Me, Life, Not Softly	men, woman
1978	And Still I Rise — Part Two - Traveling	let, old
1978	And Still I Rise — Part Three -And Still I Rise	ai, like
1983	Shaker, Why Don't You Sing?	life, night

Test Corpus: (Fig.15)

Year	Collection	Two most common words
1990	I Shall Not Be Moved	man, men
1993	On the Pulse of Morning	river, rock
1995	A Brave and Startling Truth	come, people
2005	Amazing Peace	peace, us
2006	Mother: A Cradle to Hold Me	left, returned
2009	We Had Him	know, us
2014	His Day Is Done	done, day

<u>Most Common POS Results</u>

Train Corpus: (Fig.16)

Year	Collection	Two most common POS
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two	Noun singular, Adjective
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part One	Noun singular, Adjective
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Two	Noun singular, Personal pronoun
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Three	Noun singular, Adjective
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Four	Noun singular, Adjective
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Five	Noun singular, Preposition
1978	And Still I Rise — Part One - Touch Me, Life, Not Softly	Noun singular, Preposition
1978	And Still I Rise — Part Two - Traveling	Noun singular, Adjective
1978	And Still I Rise — Part Three - And Still I Rise	Noun singular, Preposition
1983	Shaker, Why Don't You Sing?	Noun singular, Preposition
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two	Noun singular, Adjective

Test Corpus: (Fig.17)

1990	I Shall Not Be Moved	Noun singular, Preposition
1993	On the Pulse of Morning	Noun singular, Other
1995	A Brave and Startling Truth	Noun singular, Preposition
2005	Amazing Peace	Noun singular, Preposition
2006	Mother: A Cradle to Hold Me	Noun singular, Personal pronoun
2009	We Had Him	Noun singular, Preposition
2014	His Day Is Done	Noun singular, Preposition
1990	I Shall Not Be Moved	Noun singular, Preposition

<u>Most Common Text Structure Results</u>

Train Corpus: (Fig.18)

Year	Collection	Two most common Text Structure
971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part One	Question
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two	Question
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part One	Exclamation
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Two	Starting Conjunction
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Three	Question
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Four	Question
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Five	Question
1978	And Still I Rise — Part One - Touch Me, Life, Not Softly	Starting Conjunction
1978	And Still I Rise — Part Two - Traveling	Question
1978	And Still I Rise — Part Three - And Still I Rise	Question
1983	Shaker, Why Don’t You Sing?	Starting Conjunction

Test Corpus: (Fig.19)

1990	I Shall Not Be Moved	Question
1993	On the Pulse of Morning	Starting Conjunction
1995	A Brave and Startling Truth	Question
2005	Amazing Peace	Question
2006	Mother: A Cradle to Hold Me	Question
2009	We Had Him	Question
2014	His Day Is Done	Question

1990	I Shall Not Be Moved	Question

Most Common Topics Results

Train Corpus: (Fig.20)

Year	Collection	Two most common Topics
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part One	love, home
1971	Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two	life, night
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part One	em, pickin
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Two	nobody, alone
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Three	lain, thus
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Four	naught, black
1975	Oh Pray My Wings Are Gonna Fit Me Well — Part Five	see, children
1978	And Still I Rise — Part One - Touch Me, Life, Not Softly	men, woman
1978	And Still I Rise — Part Two - Traveling	let, round
1978	And Still I Rise — Part Three - And Still I Rise	ai, black
1983	Shaker, Why Don't You Sing?	life, bird

Test Corpus: (Fig.21)

Year	Collection	Two most common Topics
1990	I Shall Not Be Moved	man, men
1993	On the Pulse of Morning	river, rock
1995	A Brave and Startling Truth	come, people
2005	Amazing Peace	peace, look
2006	Mother: A Cradle to Hold Me	left, thank
2009	We Had Him	know, nothing
2014	His Day Is Done	done, Mandela

Conclusions

Following conclusions upon the results from feature extractions, clustering, and most common topics can be drawn.

Feature analysis

- Various metrics like document length, sentence length, and word length indicates Maya Angelou's consistent stylistic approach, regardless of the larger train corpus size.
- Readability scores emphasize her sustained approach to accessibility in poetry.
- Lexical richness shows a slight increase in diversity over time.
- Semantic repetition decreases, hinting at evolving language use rather than a shift in style.
- Punctuation patterns remain similar, with minor developments in later works.
- Distribution of function and content words maintains its shape, reflecting core characteristics of her writing style.
- Neutral sentiments predominate in both corpora, with later poems showing a rise in positive sentiment and milder sentiment intensity, possibly mirroring changes in thematic content rather than style.

Clustering analysis

- The clustering of poems in the train corpus is dense and varied, suggesting diverse styles or themes, while the test corpus shows a more concentrated clustering, implying a more consistent style in later works.
- Both periods show positive distributions along the primary axis, indicating a foundational style or thematic element that Maya Angelou consistently maintained throughout her career.
- The variation along the second axis, especially in the earlier works, points to a range of explorations within her poetry, but without significant oppositional shifts between the two periods.

Topic analysis

- Early works touch on themes like love and home, nobody and alone, see and children, men and women, life and bird, let and round
  → hence we could state that the initial period expresses self and social identities, and the complexity of human connections.
- Later works touch on themes like come and people, peace and look, left and thank, done and Mandela, know and nothing
  → which could express desire for unity as self-reflection, which is reasonable and suits a stage of older years of Maya Angelou.

Overall, Maya Angelou's writing style demonstrates remarkable stability over time, with subtle thematic and emotional evolutions rather than significant stylistic changes.

Sources and Figures

Resources

1. Maya Angelou
   Link: https://en.wikipedia.org/wiki/Maya_Angelou
   Link: https://encyclopediaofarkansas.net/entries/maya-angelou-1085/

2. Poetry of Maya Angelou
   Link: https://en.wikipedia.org/wiki/Poetry_of_Maya_Angelou

3. PDF: THE COMPLETE COLLECTED POEMS OF MAYA ANGELOU – Random House
   Link:
   chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://antilogicalism.com/wp-content/uploads/2020/07/maya_angelou.pdf

4. Sources for poems not in the pdf:
   - *A Brave and Startling Truth:*
   Link: https://www.poetry-chaikhana.com/Poets/A/AngelouMaya/ABraveand/index.html-
   Amazing peace:
   Link: https://www.oprah.com/oprahshow/maya-angelous-amazing-peace
   - MOTHER, A CRADLE TO HOLD ME:
   Link: HTTPS://ADOPTION-BEYOND.ORG/MOTHER-A-CRADLE-TO-HOLD-ME-BY-MAYA-ANGELOU/

   - We Had Him:
   Link: https://allpoetry.com/poem/14326539-We-Had-Him-by-Maya-Angelou

   - His Day Is Done:
   Link: https://www.nelsonmandela.org/news/entry/dr.-maya-angelou-his-day-is-done-a-tribute-poem-for-nelson-mandela

Figures

Fig. 1

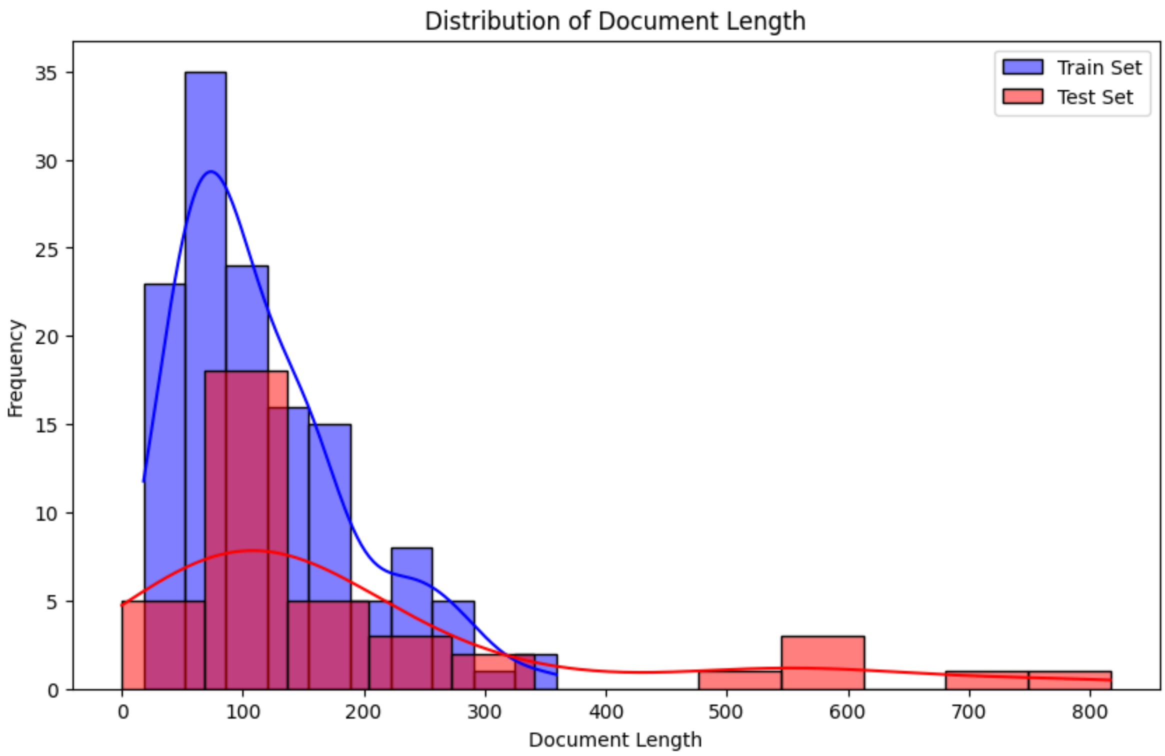


Fig. 2

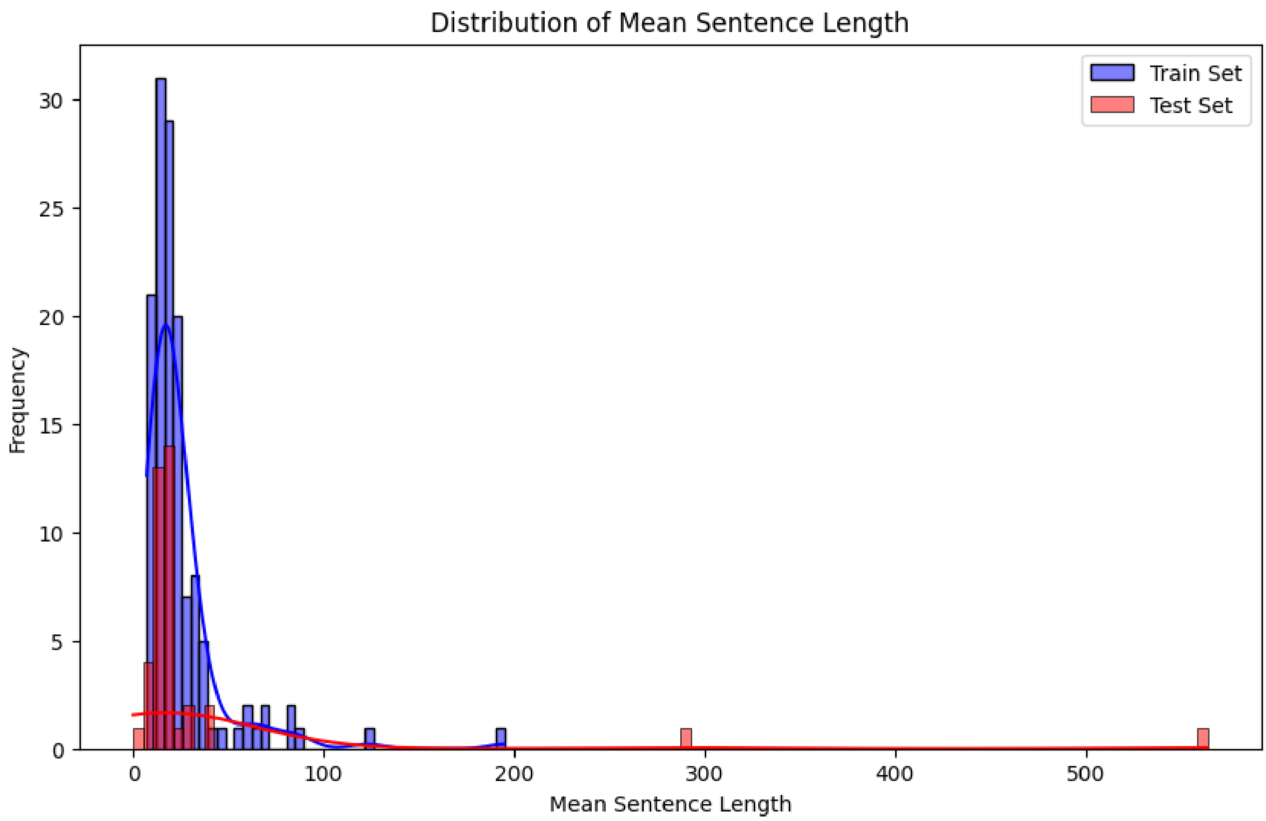


Fig. 3

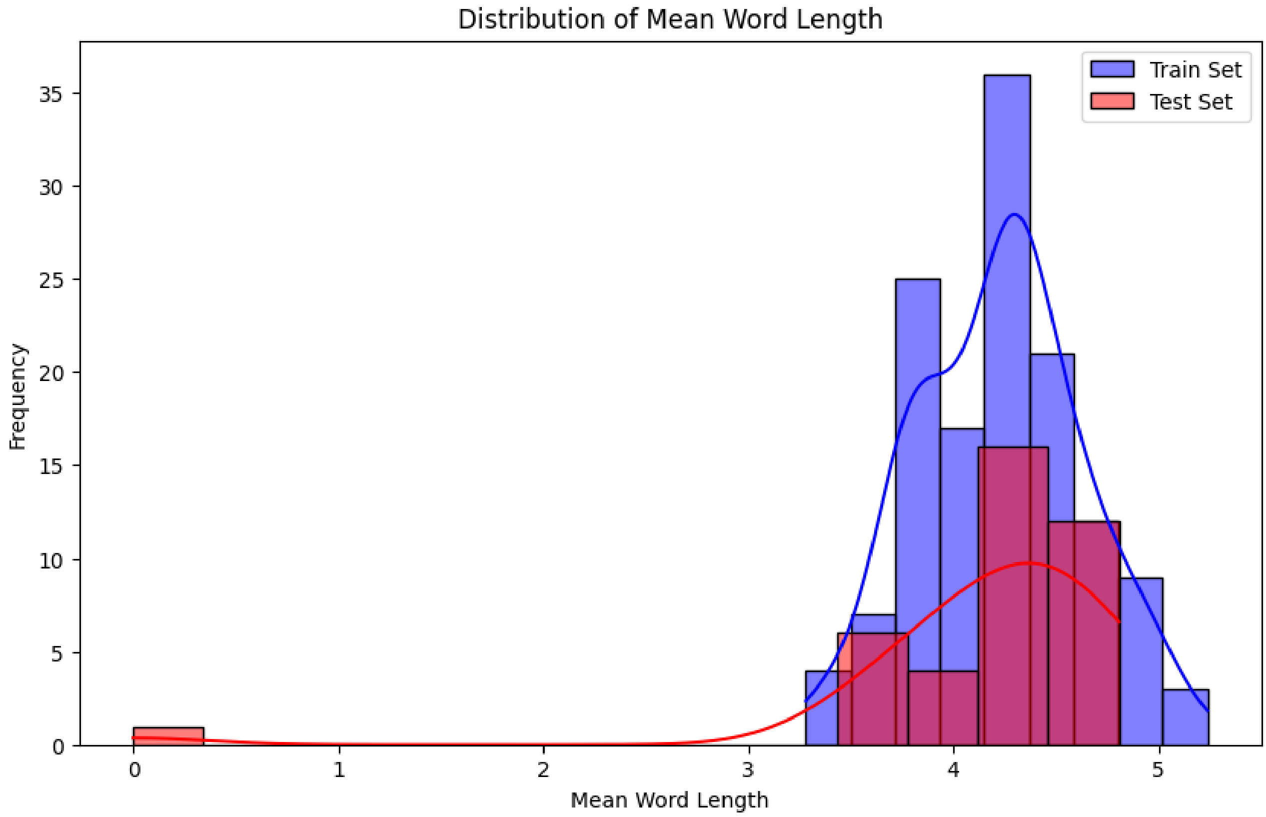

Distribution of Mean Word Length
Train Set
Test Set
Frequency
Mean Word Length
0
5
10
15
20
25
30
35
0
1
2
3
4
5

Fig. 4

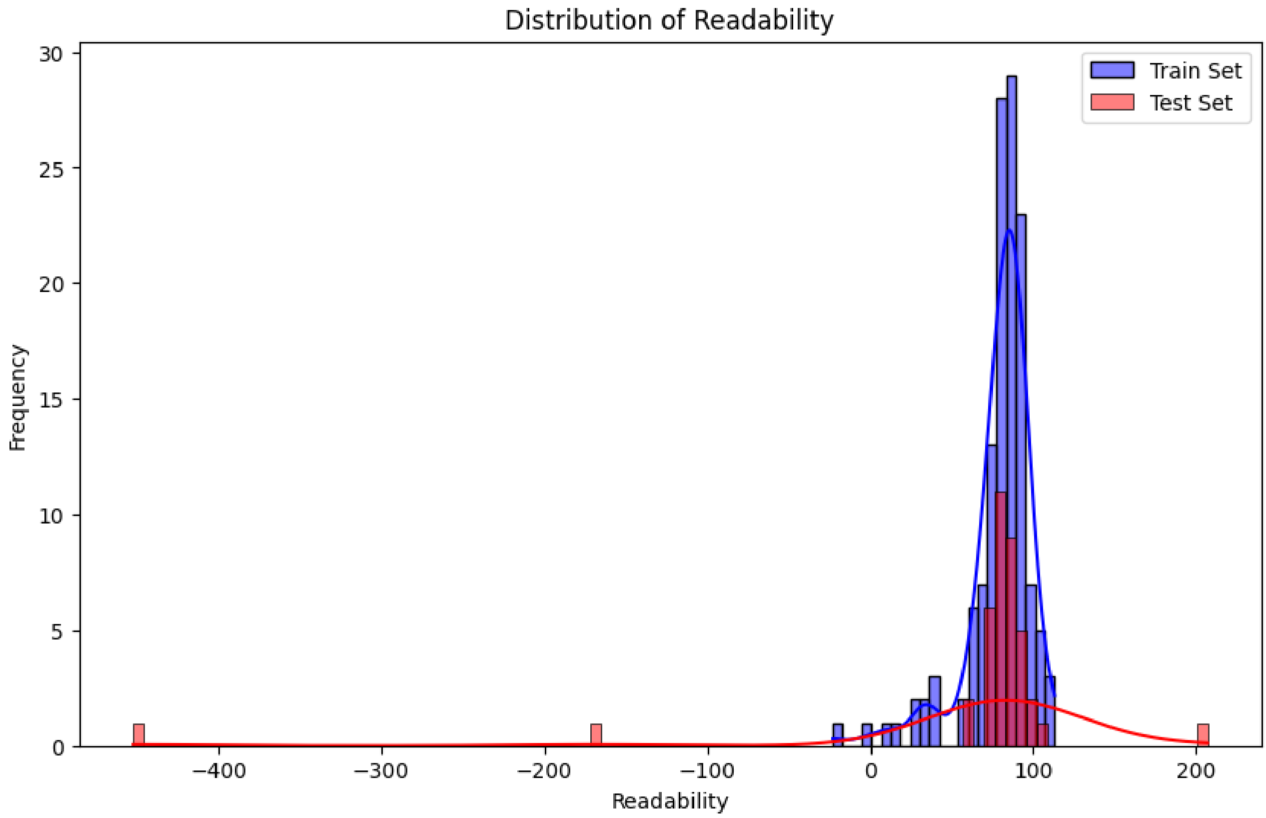

Distribution of Readability
Train Set
Test Set
Frequency
Readability
0
5
10
15
20
25
30
−400
−300
−200
−100
0
100
200

Fig. 5

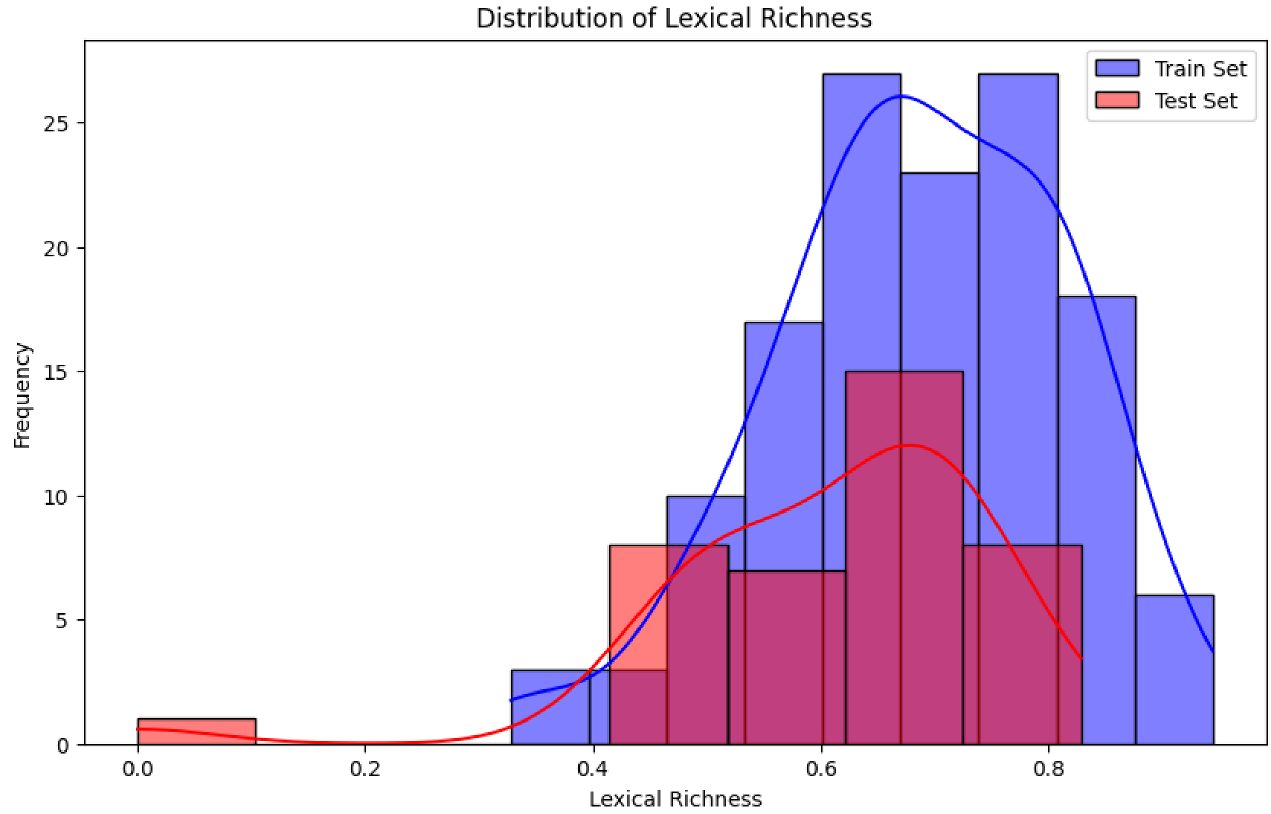

Distribution of Lexical Richness
Train Set
Test Set
Frequency
Lexical Richness
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8

Fig. 6

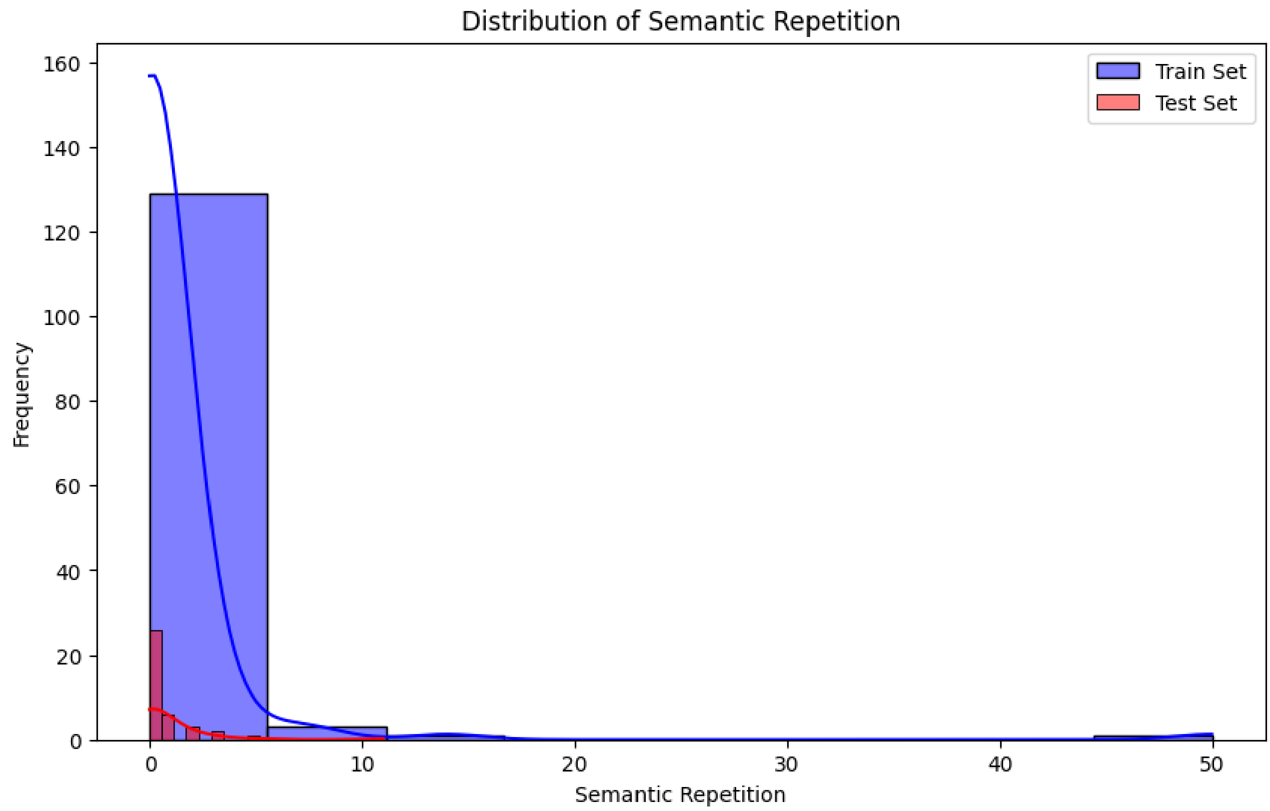

Distribution of Semantic Repetition
Train Set
Test Set
Frequency
Semantic Repetition
0
20
40
60
80
100
120
140
160
0
10
20
30
40
50

Fig. 7

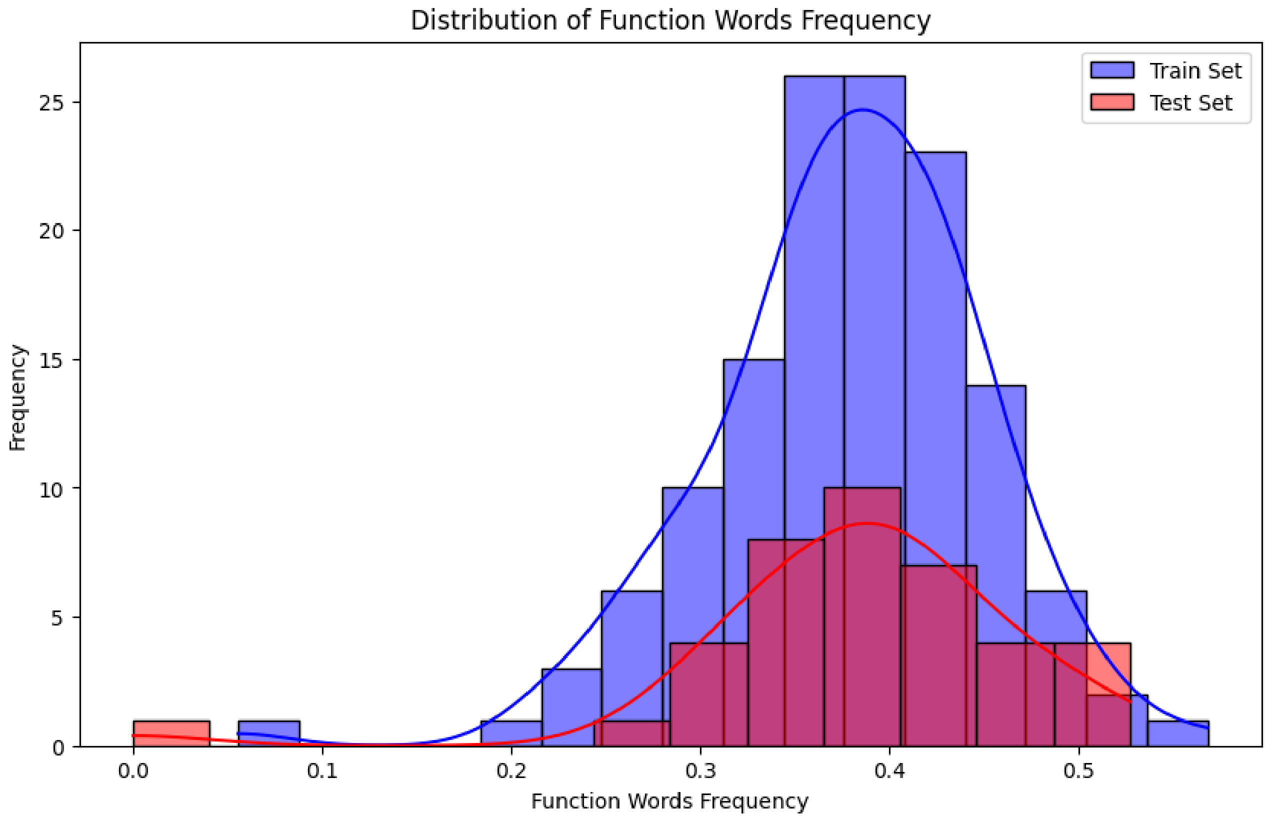

Distribution of Function Words Frequency
Train Set
Test Set
Frequency
Function Words Frequency
0
5
10
15
20
25
0.0
0.1
0.2
0.3
0.4
0.5

Fig. 8

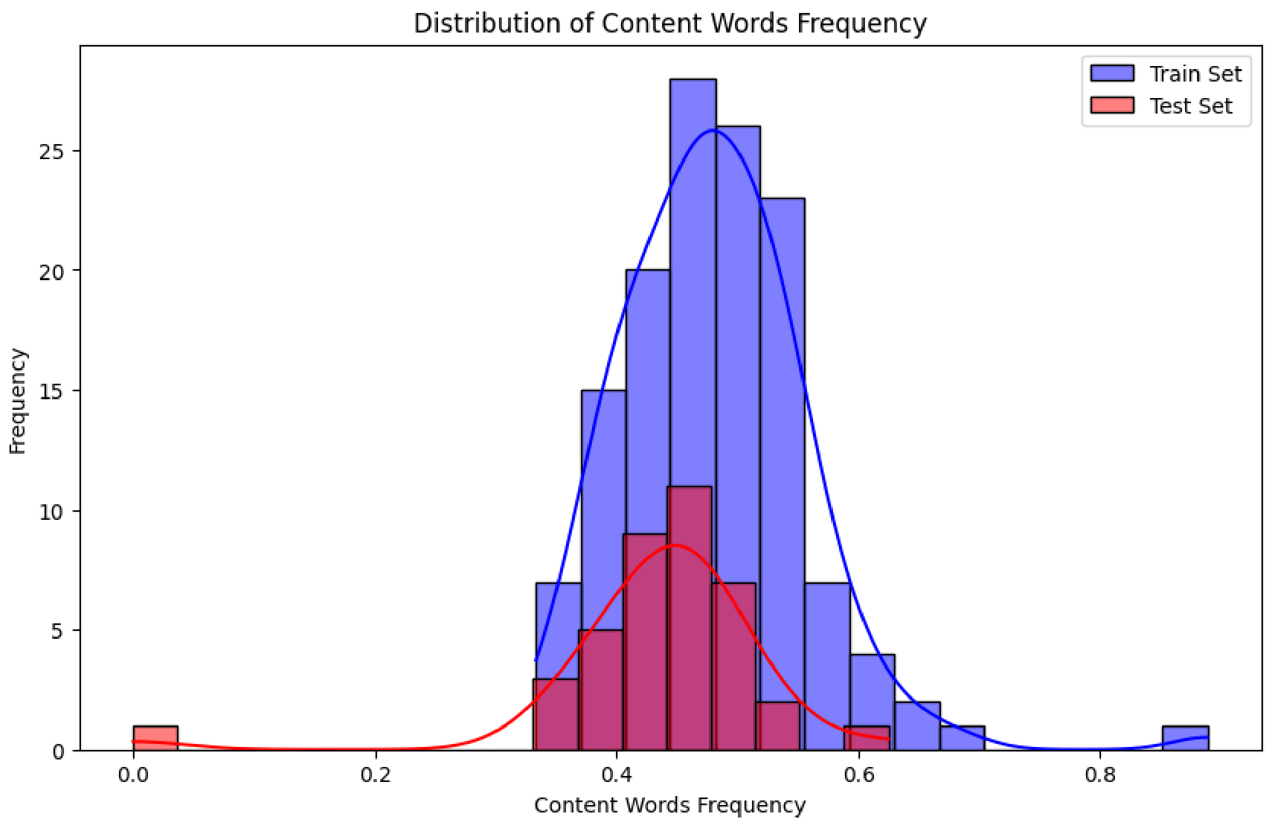

Distribution of Content Words Frequency
Train Set
Test Set
Frequency
Content Words Frequency
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8

Fig. 9

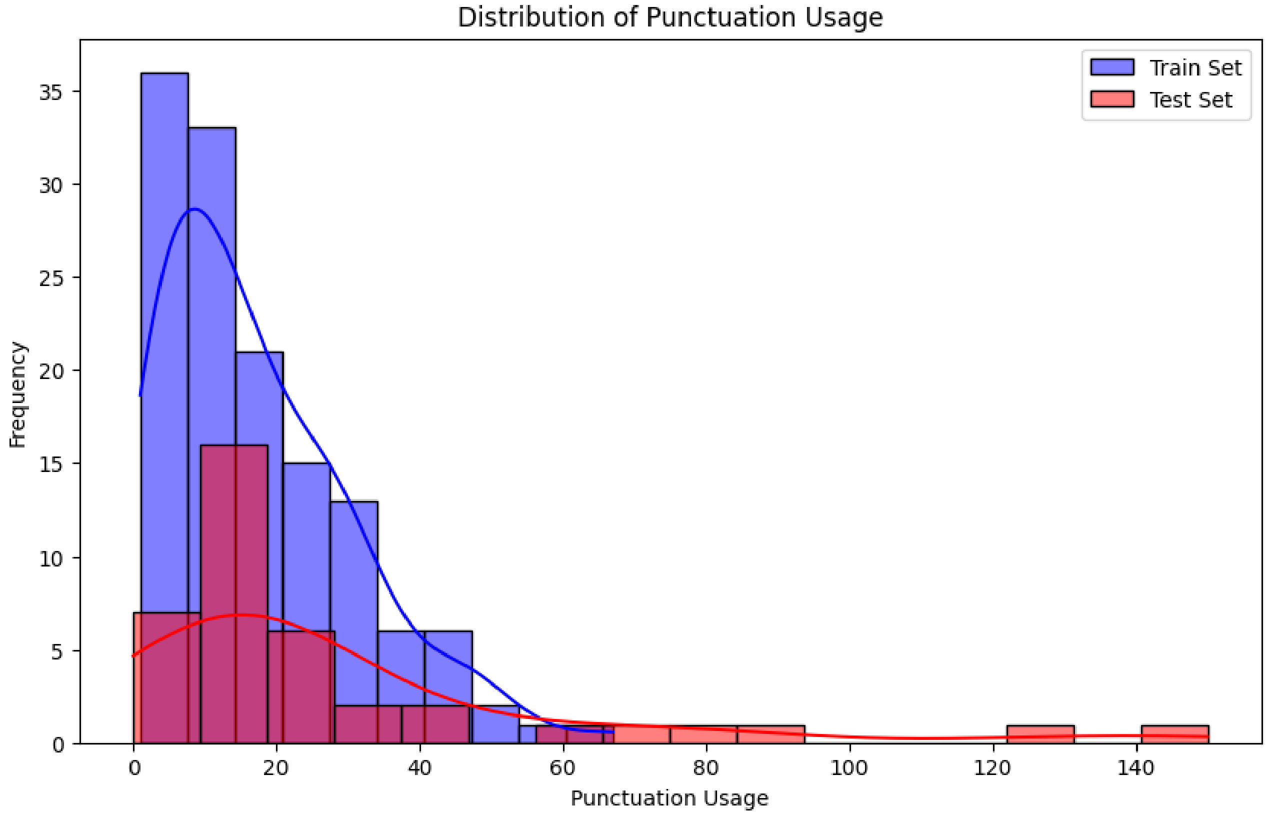

Distribution of Punctuation Usage
Train Set
Test Set
Frequency
0
5
10
15
20
25
30
35
0
20
40
60
80
100
120
140
Punctuation Usage

Fig. 10

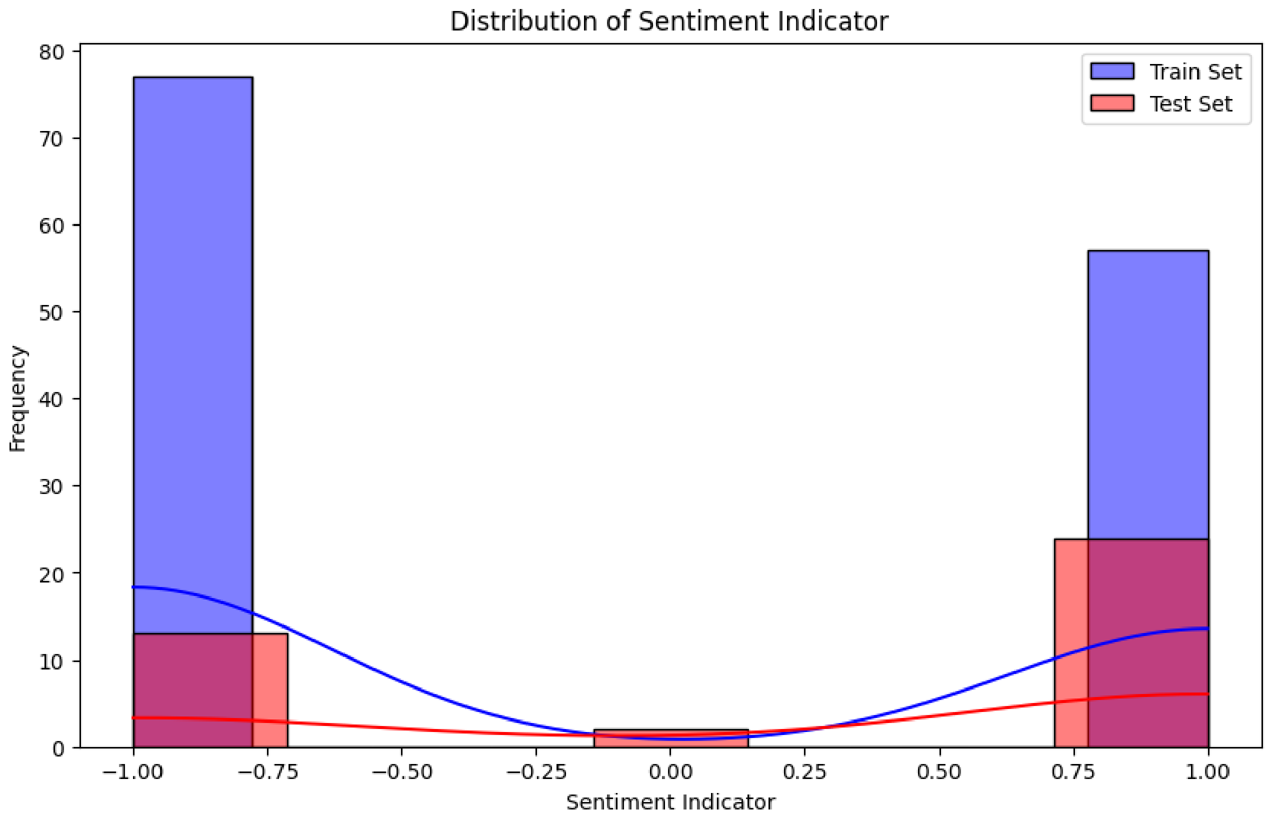

Distribution of Sentiment Indicator
Train Set
Test Set
Frequency
0
10
20
30
40
50
60
70
80
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Indicator

Fig. 11

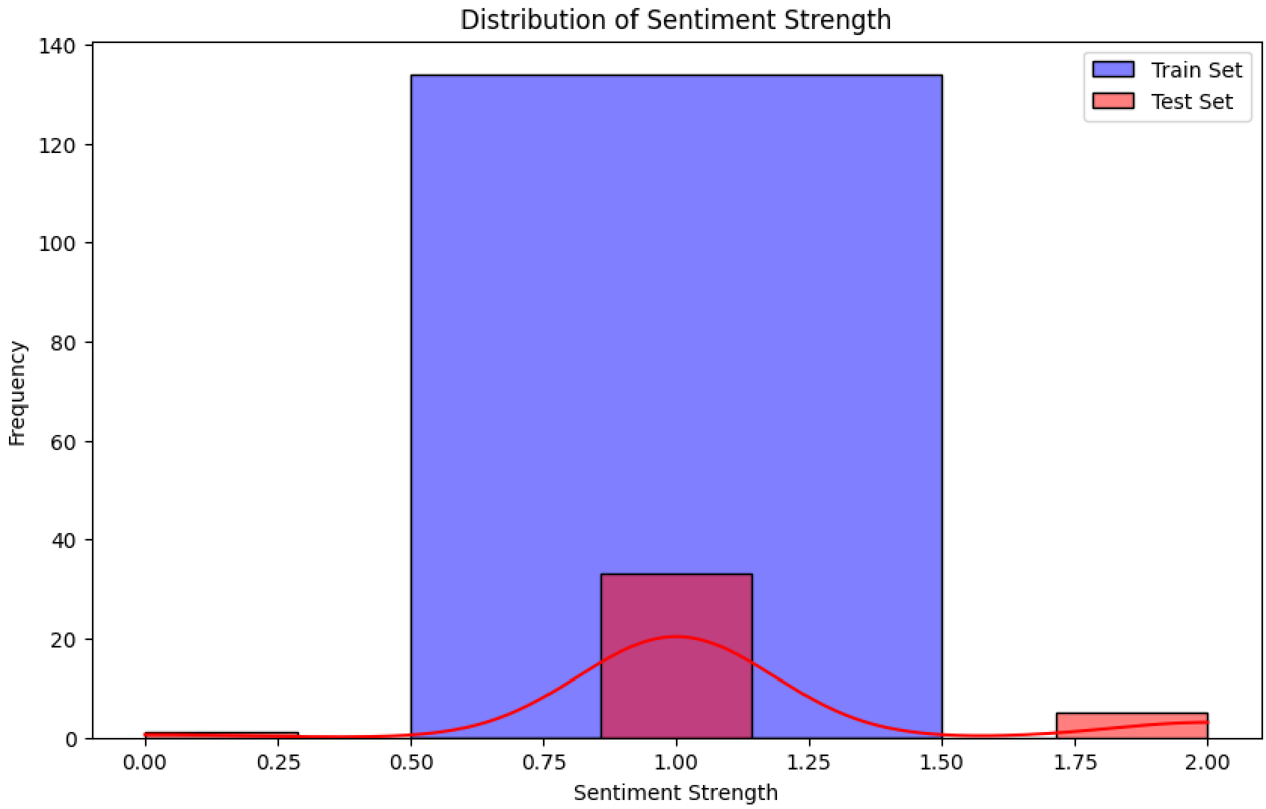

Distribution of Sentiment Strength
Train Set
Test Set
Frequency
140
120
100
80
60
40
20
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Sentiment Strength

Fig. 12

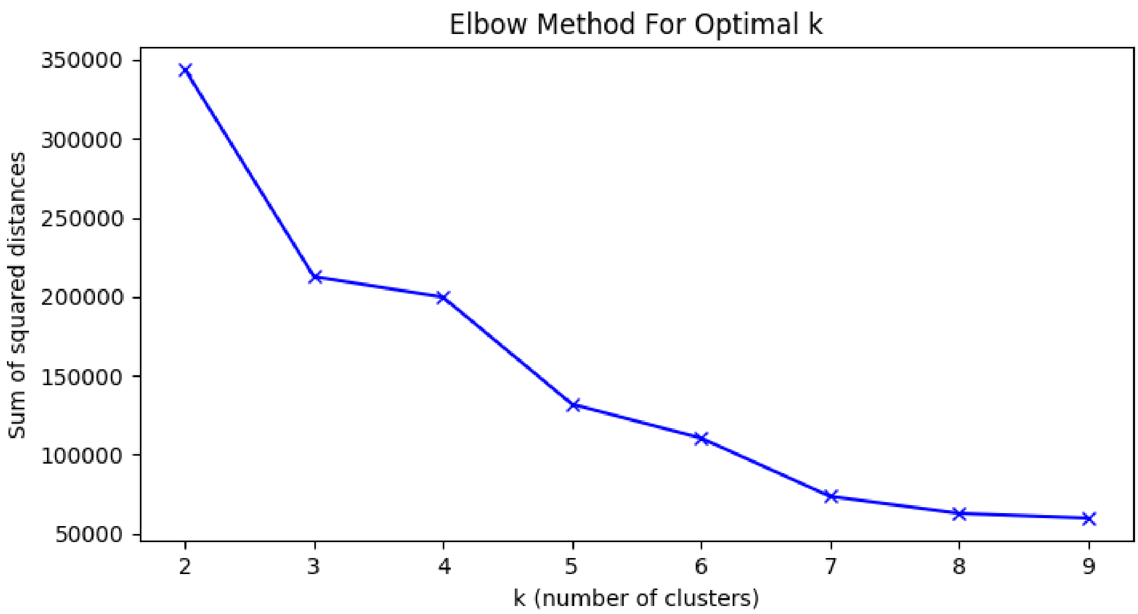

Elbow Method For Optimal k
Sum of squared distances
350000
300000
250000
200000
150000
100000
50000
2
3
4
5
6
7
8
9
k (number of clusters)

Fig. 13

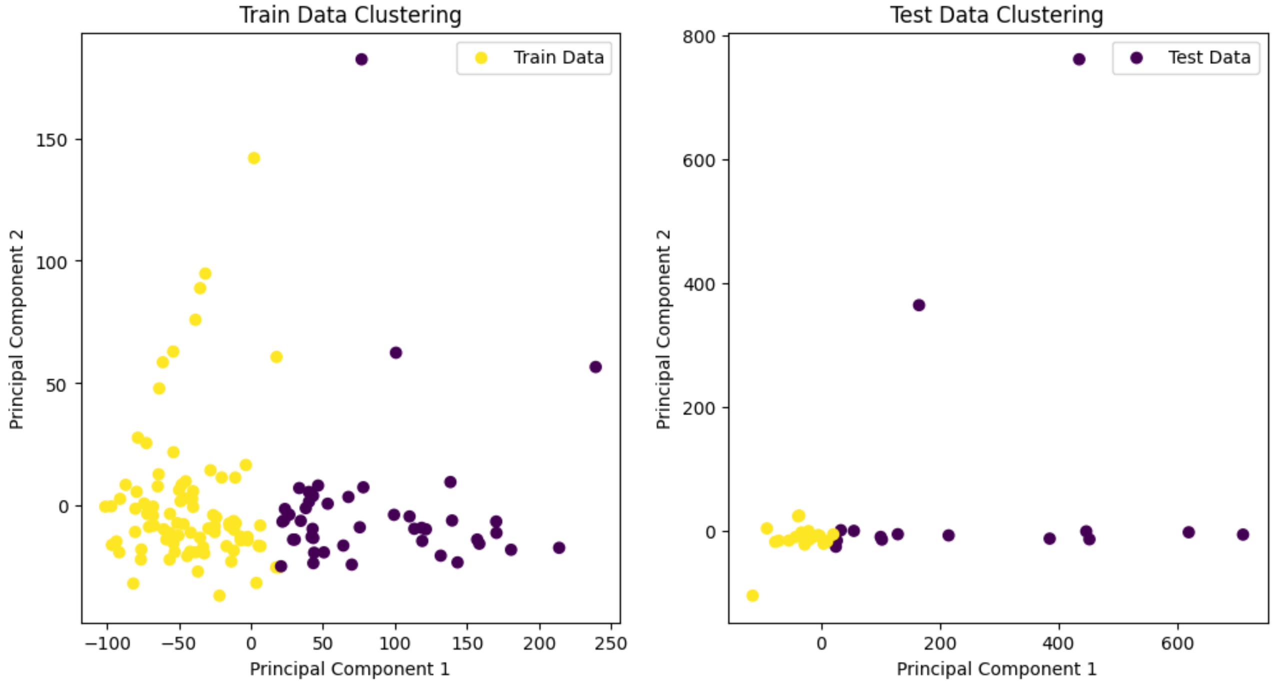

Train Data Clustering
Train Data
Principal Component 2
Principal Component 1
150
100
50
0
−100
−50
0
50
100
150
200
250
Test Data Clustering
Test Data
Principal Component 2
Principal Component 1
800
600
400
200
0
0
200
400
600

Fig. 14

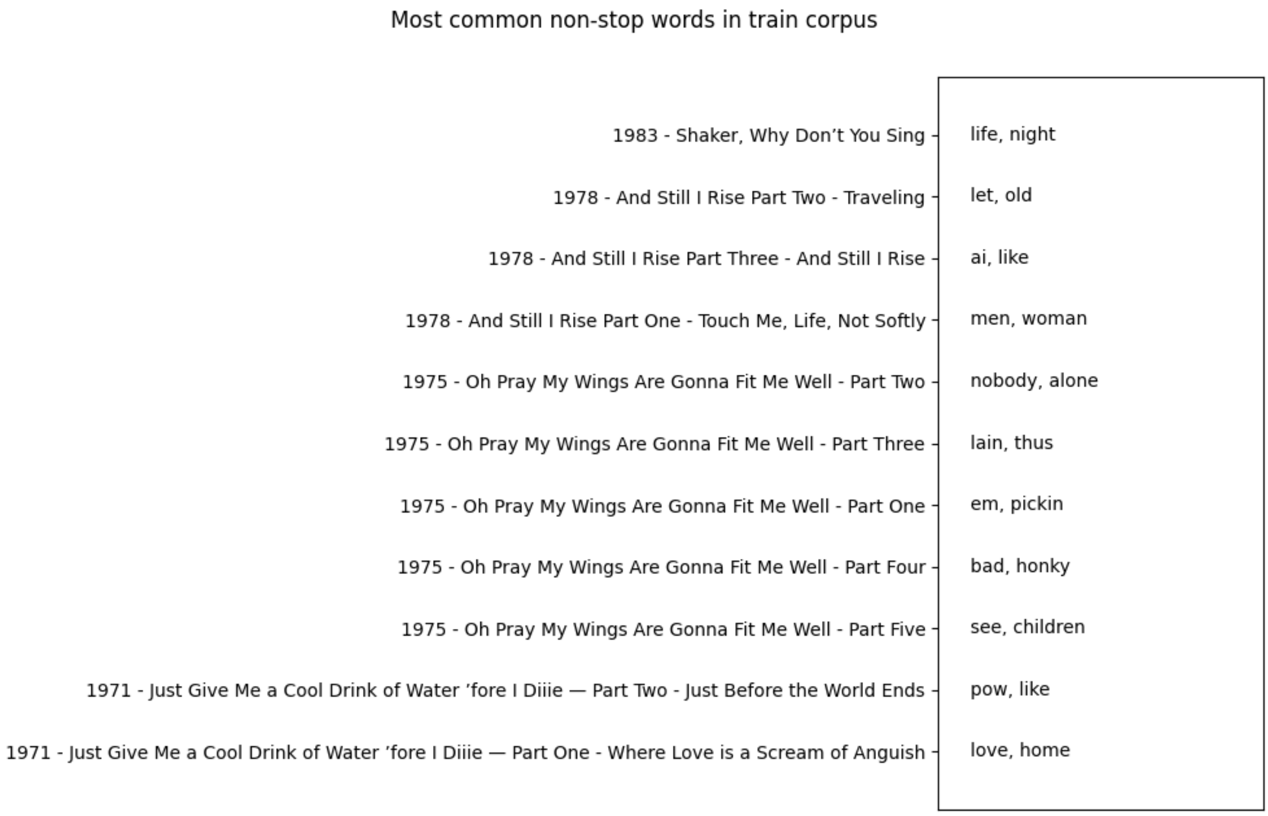

Most common non-stop words in train corpus
1983 - Shaker, Why Don’t You Sing
life, night
1978 - And Still I Rise Part Two - Traveling
let, old
1978 - And Still I Rise Part Three - And Still I Rise
ai, like
1978 - And Still I Rise Part One - Touch Me, Life, Not Softly
men, woman
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Two
nobody, alone
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Three
lain, thus
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part One
em, pickin
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Four
bad, honky
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Five
see, children
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part Two - Just Before the World Ends
pow, like
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part One - Where Love is a Scream of Anguish
love, home

Fig. 15

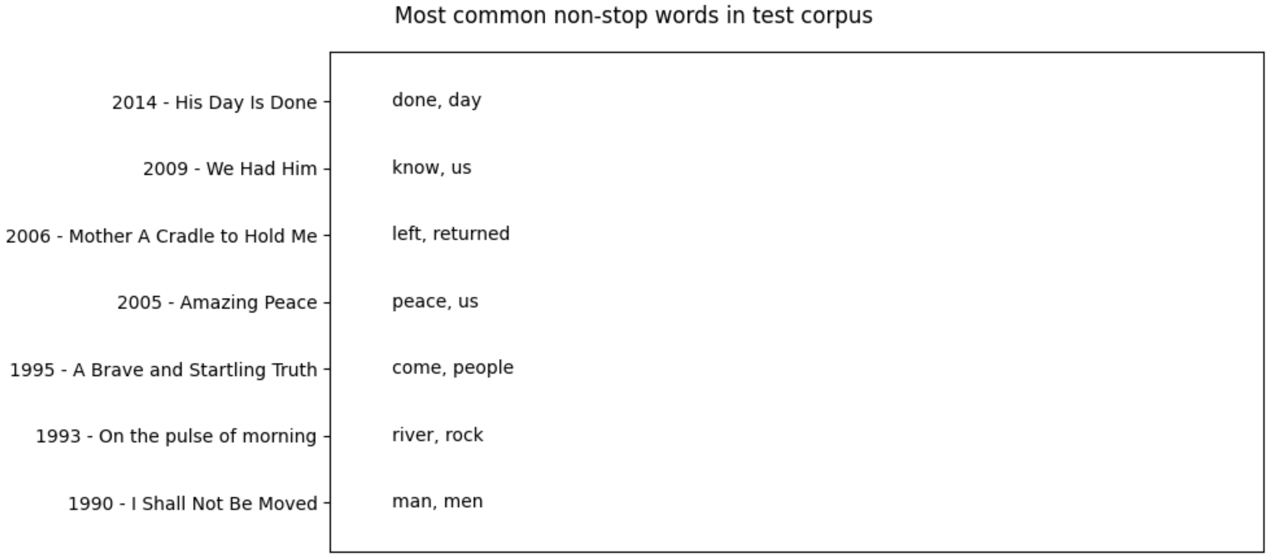

Most common non-stop words in test corpus
2014 - His Day Is Done
done, day
2009 - We Had Him
know, us
2006 - Mother A Cradle to Hold Me
left, returned
2005 - Amazing Peace
peace, us
1995 - A Brave and Startling Truth
come, people
1993 - On the pulse of morning
river, rock
1990 - I Shall Not Be Moved
man, men

Fig. 16

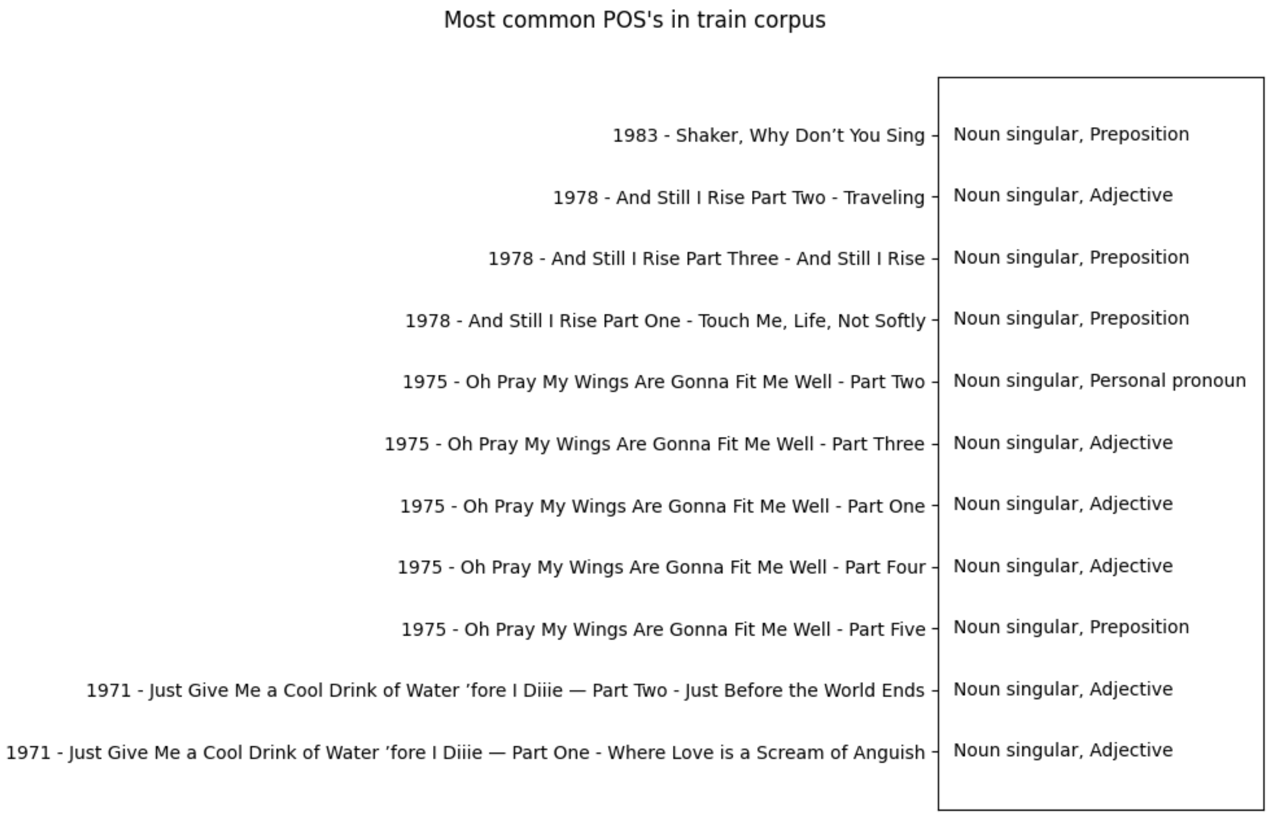

Most common POS's in train corpus
1983 - Shaker, Why Don't You Sing
Noun singular, Preposition
1978 - And Still I Rise Part Two - Traveling
Noun singular, Adjective
1978 - And Still I Rise Part Three - And Still I Rise
Noun singular, Preposition
1978 - And Still I Rise Part One - Touch Me, Life, Not Softly
Noun singular, Preposition
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Two
Noun singular, Personal pronoun
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Three
Noun singular, Adjective
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part One
Noun singular, Adjective
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Four
Noun singular, Adjective
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Five
Noun singular, Preposition
1971 - Just Give Me a Cool Drink of Water 'fore I Diiie — Part Two - Just Before the World Ends
Noun singular, Adjective
1971 - Just Give Me a Cool Drink of Water 'fore I Diiie — Part One - Where Love is a Scream of Anguish
Noun singular, Adjective

Fig. 17

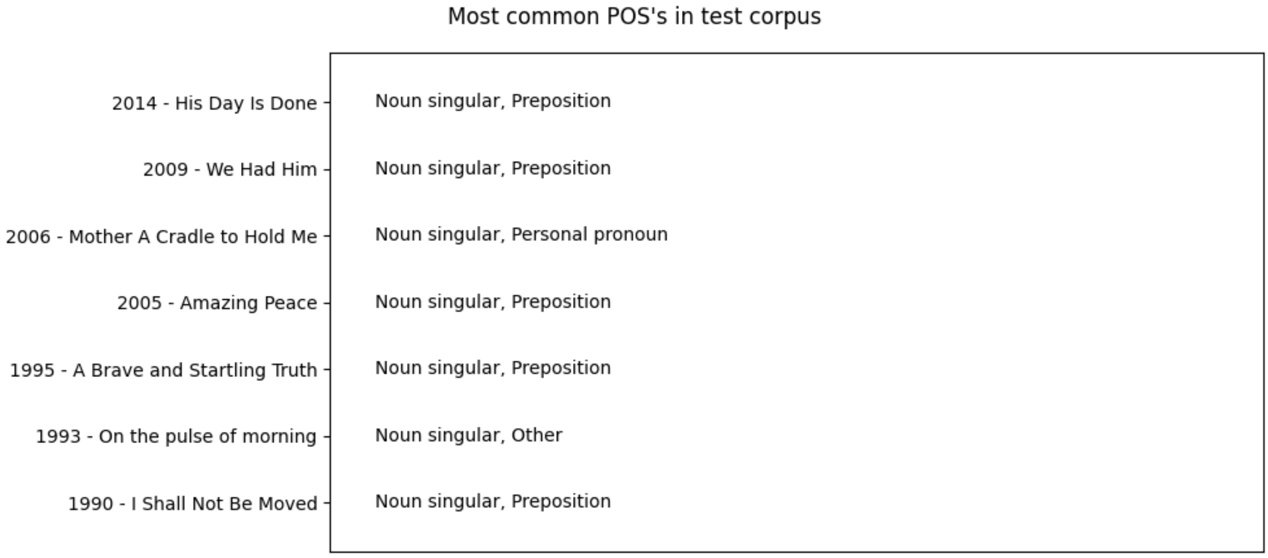

Most common POS's in test corpus
2014 - His Day Is Done
Noun singular, Preposition
2009 - We Had Him
Noun singular, Preposition
2006 - Mother A Cradle to Hold Me
Noun singular, Personal pronoun
2005 - Amazing Peace
Noun singular, Preposition
1995 - A Brave and Startling Truth
Noun singular, Preposition
1993 - On the pulse of morning
Noun singular, Other
1990 - I Shall Not Be Moved
Noun singular, Preposition

Fig. 18

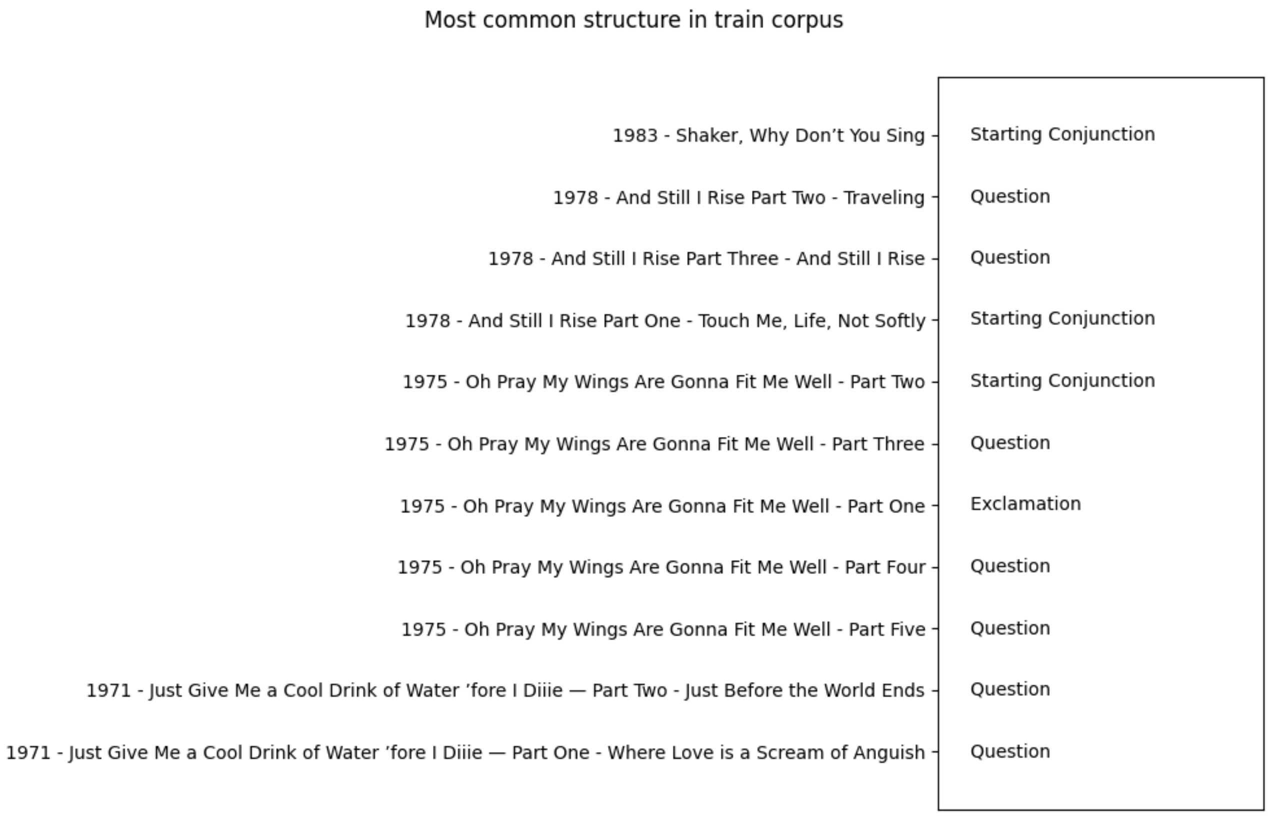

Most common structure in train corpus
1983 - Shaker, Why Don’t You Sing
Starting Conjunction
1978 - And Still I Rise Part Two - Traveling
Question
1978 - And Still I Rise Part Three - And Still I Rise
Question
1978 - And Still I Rise Part One - Touch Me, Life, Not Softly
Starting Conjunction
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Two
Starting Conjunction
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Three
Question
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part One
Exclamation
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Four
Question
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Five
Question
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part Two - Just Before the World Ends
Question
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part One - Where Love is a Scream of Anguish
Question

Fig. 19

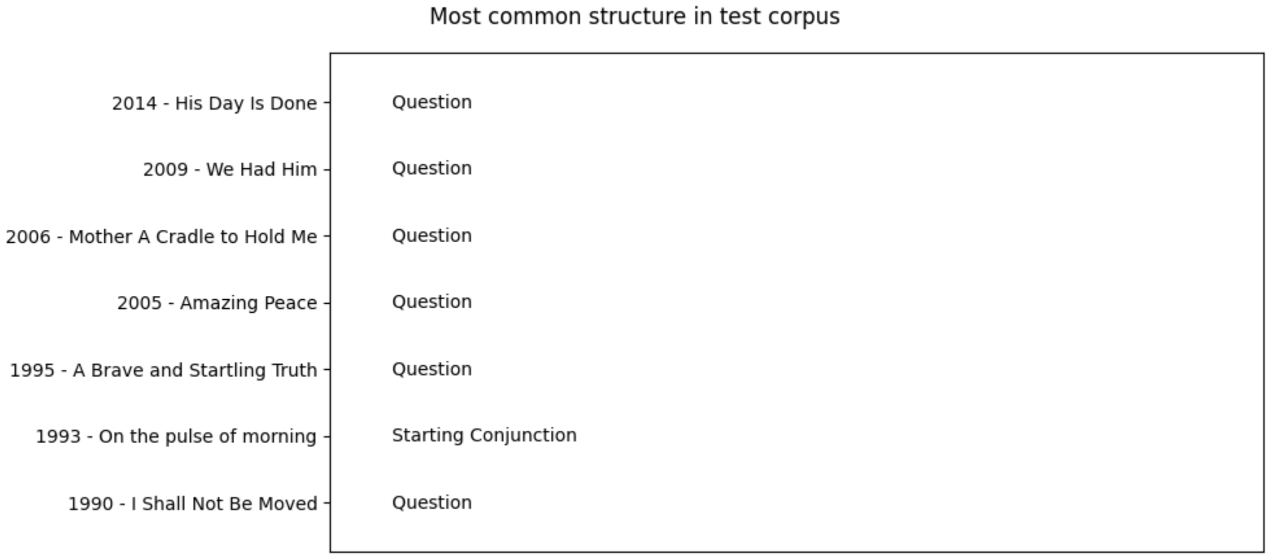
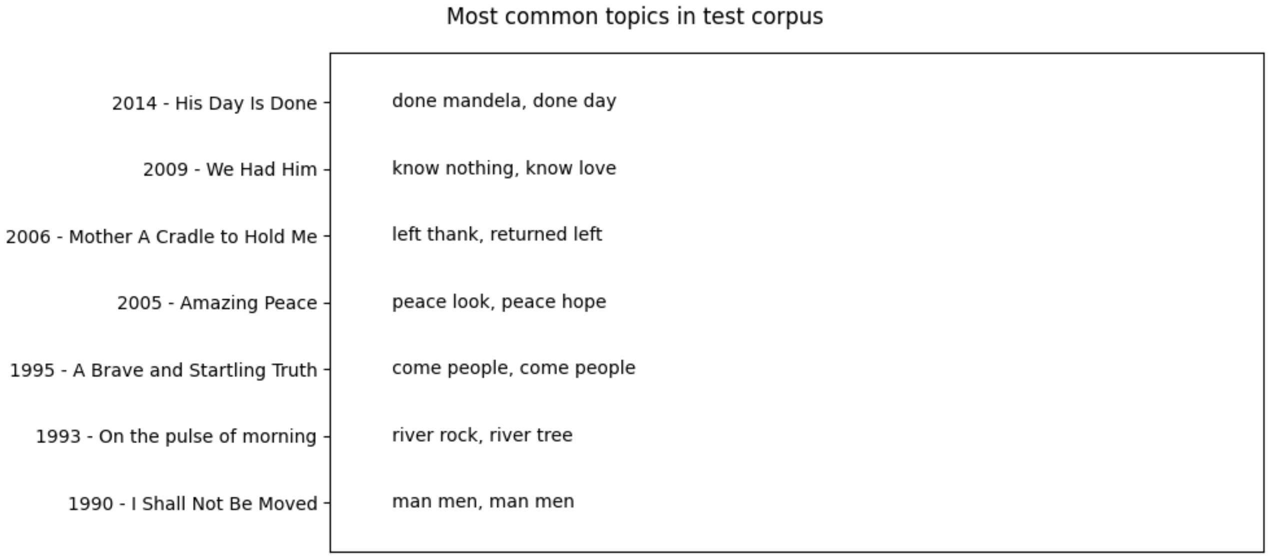

Most common structure in test corpus
2014 - His Day Is Done
Question
2009 - We Had Him
Question
2006 - Mother A Cradle to Hold Me
Question
2005 - Amazing Peace
Question
1995 - A Brave and Startling Truth
Question
1993 - On the pulse of morning
Starting Conjunction
1990 - I Shall Not Be Moved
Question

Fig. 20


Most common topics in train corpus
1983 - Shaker, Why Don’t You Sing
life night, life bird
1978 - And Still I Rise Part Two - Traveling
let round, born let
1978 - And Still I Rise Part Three - And Still I Rise
ai black, ai life
1978 - And Still I Rise Part One - Touch Me, Life, Not Softly
men woman, woman men
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Two
nobody alone, alone nobody
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Three
lain thus, lain leaves
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part One
em layin, em pickin
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Four
naught black, eyes bad
1975 - Oh Pray My Wings Are Gonna Fit Me Well - Part Five
see children, see children
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part Two - Just Before the World Ends
pow man, pow black
1971 - Just Give Me a Cool Drink of Water ’fore I Diiie — Part One - Where Love is a Scream of Anguish
love came, home love

Fig. 21


Most common topics in test corpus
2014 - His Day Is Done
2009 - We Had Him
2006 - Mother A Cradle to Hold Me
2005 - Amazing Peace
1995 - A Brave and Startling Truth
1993 - On the pulse of morning
1990 - I Shall Not Be Moved
done mandela, done day
know nothing, know love
left thank, returned left
peace look, peace hope
come people, come people
river rock, river tree
man men, man men