Course: Information Retrieval

Major: Computer Science and Technology

**Declaration**

I declare that all the research and developed software presented in this report is my own, unless stated otherwise in the text.

**Text Classification for Song Lyrics and News Data**

**Abstract**: News articles are widely used samples for text classification owing to their clear structures, specific topics, and formal language use. In contrast, song lyrics are more complicated in structure, expressions, emotions, and are used less for classification. In this project, with the aim of providing basis for related recommendation systems, two data sets, the Song Lyrics Data Set and the News Data set are collected on the Internet via web crawler. Some mainstream preprocessing and feature engineering methods in NLP are utilized to improve the quality of data and extract features from the two data sets. Five traditional machine learning models and two deep learning models are used to perform text classification. Overall, the performance of models on the News Data Set has overwhelmingly exceeds the performance of models on the Song Lyrics Data Set, with the best accuracy reaching 88%. The results have shown that deep learning models do not show any advantage over traditional machine learning models in the Song Lyrics Data Set.

**Keywords**: text classification; natural language processing; machine learning; news articles; song lyrics

Contents

[1. Introduction 5](#_Toc111233222)

[2. Datasets 5](#_Toc111233223)

[2.1 Data Collection 5](#_Toc111233224)

[2.2 Data Preprocessing 6](#_Toc111233225)

[2.2.1 Data Structuring 6](#_Toc111233226)

[2.2.2 Duplicates Removal 7](#_Toc111233227)

[2.2.3 Extension Removal 7](#_Toc111233228)

[2.2.4 Abbreviations Extension & Tokenization 8](#_Toc111233229)

[2.3 Exploratory Data Analysis and Visualization 8](#_Toc111233230)

[3. Feature Engineering 10](#_Toc111233231)

[3.1 Word Frequency (TFIDF) 10](#_Toc111233232)

[3.2 Simple Sentimental Analysis 12](#_Toc111233233)

[3.3 Part of Speech 14](#_Toc111233234)

[3.4 Number of Words 14](#_Toc111233235)

[3.5 Repeated Words 15](#_Toc111233236)

[3.6 Features Used in Classification 16](#_Toc111233237)

[4. Methods 18](#_Toc111233238)

[4.1 Logistic Regression (LR) 18](#_Toc111233239)

[4.2 Support Vector Machine (SVM) 19](#_Toc111233240)

[4.3 Random Forest (RF) 19](#_Toc111233241)

[4.4 AdaBoost 19](#_Toc111233242)

[4.5 Voting (LR, SVM, RF) 20](#_Toc111233243)

[4.6 TextCNN (Text Convolution Neural Network) 20](#_Toc111233244)

[4.6.1 Words to Vectors 20](#_Toc111233245)

[4.6.2 Neural Network Structure 20](#_Toc111233246)

[4.7 Transformer 21](#_Toc111233247)

[4.7.1 Words to Vectors 21](#_Toc111233248)

[4.7.2 Neural Network Structure 21](#_Toc111233249)

[5. Results 22](#_Toc111233250)

[5.1 Traditional Machine Learning Models 22](#_Toc111233251)

[5.2 Deep Learning Models 24](#_Toc111233252)

[5.2.1 TextCNN 24](#_Toc111233253)

[5.2.2 Transformer 25](#_Toc111233254)

[6. Discussion 26](#_Toc111233255)

[6.1 Data 26](#_Toc111233256)

[6.2 Features 26](#_Toc111233257)

[6.3 Models 26](#_Toc111233258)

[6.4 Potential Improvements 27](#_Toc111233259)

[References 28](#_Toc111233260)

1. Introduction

News articles with different categories normally adopt a clear structure, specific topics, standardized words and show a strong emotional tendency. Hence, news data are widely used in text classification. In contrast, song lyrics tends to use colloquial expressions with complex emotions, and there is very little research on the use of lyrics data for text classification of songs by artists. However, the song lyrics written by the same artist sometimes share traceable styles, and song lyrics classification based on artists may be able to provide reference basis for artist recommendations.

In this project, a song lyrics dataset and a news dataset are collected by web crawler implemented with Python. Several mainstream methods in Natural Language Processing (NLP) are utilized to perform data preprocessing and feature engineering. After that, 5 traditional machine learning models implemented by *scikit-learn*are used for classification and the performance of models on two data sets are analyzed and compared. Deep learning models TextCNN and Transformer implemented by *pytorch* are used to perform classification on the Song Lyrics Data Set.

# Datasets

There are two datasets collected by web crawler for the experiments, one is the lyrics dataset, and the other is the news dataset.

## Data Collection

The song lyrics data set is collected from the official website of NetEase Cloud Music (<https://music.163.com/>). 41 most popular artists over the recent three decades are selected, and song lyrics from all of their albums are collected. The archives are classified according to the artist, and the lyrics of each song are written in a plain text file. The raw data are 3711 plain text files distributed over 41 folders according to the artists.

The news data set is collected from multiple major news website including CNN (<https://edition.cnn.com>), BBC ([http://www.bbc.com](http://www.bbc.comc)), Business Insider ([https://www.businessinsider.com](https://www.businessinsider.com/)), Huff Post ([https://www.huffpost.com](https://www.huffpost.com/)), NYTimes ([https://www.nytimes.com](https://www.nytimes.com/)), Washington Post ([https://www.washingtonpost.com](https://www.washingtonpost.com/)), and China Daily ([http://www.chinadaily.com.cn](http://www.chinadaily.com.cn/)), from 2018 to 2022. The data set contains 36 different types of news, including “POLITICS”, “BUSINESS”, “ENTERTAINMENT”, etc. The raw data set contains 120818 pieces of news and is collected in a 120818-rows 3-columns table. The 3 columns are the types, title, and main body of the news data, respectively.

## 2.2 Data Preprocessing

Data preprocessing is performed to discard the invalid data and improve the quality of data for better analysis.

### 2.2.1 Data Structuring

To optimize data structure, making further processing steps more convenient, the data in each data set are merged into a *Pandas* data frame. Meanwhile, the invalid data including the empty and incomplete data are discarded.

* *Song Lyrics Data Set*

Considering the presence of no-lyric pure music, all empty plain text files are discarded, and a 35693 data frame is obtained. The following table shows the attributes and some examples of this data set.

**Table 2-1:** The Data Frame Structure of the Lyric Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | **artist** | **song** | **text** |
| **Description** | Name of the artist (label) | Name of the song | Lyrics from the song |
| **Example** | Adele | All I Ask - Adele | [00:14.230] I will leave my heart at the door\n... |

* *News Data Set*

The news data set is already structured in a *csv* table, and *read\_csv()* is used to convert it into a data frame. For reducing the computing costs and improving the quality of the data, articles that have more than 800 words and the articles with empty body are discarded. The resulting data frame is of size 1088163. The following table shows the attributes and some examples of this data set.

**Table 2-2:** The Data Frame Structure of the News Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | **class** | **head** | **body** |
| **Description** | Type of the news  (label) | The topic of the news | The text body of the news |
| **Example** | WORLD NEWS | Palestinians vow to stay on West Bank land des... | ['Jerusalem (CNN) The leader of a Palestinian v... |

### 2.2.2 Duplicates Removal

For the Song Lyrics Dataset, since one specific song can be covered by different singers and there might be multiple versions of a song by the same singer, it is possible to have duplicated lyrics in the data set. For News Dataset, a news article may belong to two or more categories, or there may be duplicates of articles from the breaking news at different dates.

To reduce the ambiguous classification, the above duplicates are discarded. The following table shows the number of the duplicates in the two data sets.

**Table 2-3:** The Number of the Duplicates in the Two Data Sets

|  |  |  |  |
| --- | --- | --- | --- |
|  | **No. of Duplicates** | **No. of Duplicates in Different Class** | **No. of Duplicates in the Same Class** |
| **Song Lyrics Data Set** | 18 | 10 | 8 |
| **News Data Set** | 330 | 223 | 107 |

### 2.2.3 Extension Removal

For both datasets, the contents inside the brackets are usually composed of the additional information, which is not helpful for classification. Except the content inside the square brackets in the News Dataset, which contains informative texts for classification, all the contents including the brackets themselves are removed. The following table shows the typical contents in the brackets from the two datasets respectively.

**Table 2-4:** The Contents inside the Brackets from Two Data Sets

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Round Brackets** | **Square Brackets** | **Curly Brackets** |
| **Typical Content from**  **Lyric Dataset** | Chorus, harmonic lyrics | Timeline | Non-lyrics lines and notes |
| **Typical Content from**  **News Dataset** | Source of the news | Informative contents, news body | Non-informative contents |

Apart from the content in the brackets, since the line break sign “\n” will be read as a word while it should not be, they are also removed.

### 2.2.4 Abbreviations Extension & Tokenization

To perform the tokenization, *RegexpTokenizer* in the *nltk* library is utilized. During tokenization, the punctuations will be considered as signs to divide sentences into words, and the abbreviations such as “won’t” will be divided into two words such as “won” and “t”, which will cause confusion in classification. Thus, the abbreviations are extended back to their standard forms by the following table. Note that this step is only applied to the Song Lyrics Data Set since the format of words in news articles is standard and this step is not required to the News Data Set.

**Table 2-5:** Abbreviations Converting Table

|  |  |
| --- | --- |
| **Abbreviation** | **Standard Form** |
| won’t | will not |
| can’t | cannot |
| couldn’t | could not |
| I’m | I am |
| ain’t | is not |
| ‘ll | will |
| ‘t | not |
| ‘ve | have |
| ‘s | is |
| ‘re | are |
| ‘d | would |

After abbreviation extension, the texts are tokenized as the separated words, and stemming is performed to convert all words into their common basic forms.

## 2.3 Exploratory Data Analysis and Visualization

To better understand the data, we perform a simple exploratory data analysis and visualization on the number of texts per class and the number of words per text.

* *Song Lyrics Data Set*

The average number of songs per artist is 87. There are artists with less songs, which may affect the performance of classification models since most models are designed for balanced data. The distribution of number of words per song is nearly symmetric, but there are some outliers with more than 1000 words per song. Figure 2-1 and Figure 2-2 show the number of songs per artist and the number of words per song, respectively.

图表, 直方图

描述已自动生成

**Figure 2-1:** The Number of Songs per Artist

图表, 直方图

描述已自动生成

**Figure 2-2:** The Number of Words per Song

* *News Data Set*

Most categories of news have less than 4000 articles. The data from this data set is also imbalanced and the distribution of number of words per news articled is skewed. Most articles have less than 100 words, and very few of them has more than 200 words. Figure 2-3 and Figure 2-4 show the number of news articles per category and the number of words per news article, respectively.

日程表, Teams

描述已自动生成

**Figure 2-3:** The Number of News Articles per Category

图形用户界面, 应用程序, Teams

描述已自动生成

**Figure 2-4:** The Number of Words per News Article

# **Feature Engineering**

Feature engineering is an essential step since the quality of features will largely influence the performance of the traditional machine learning models. In this project, mainstream feature engineering methods in NLP are utilized on both data sets, and the extracted features are analyzed with randomly selected classes.

## **Word Frequency (TFIDF)**

In traditional NLP, Term Frequency – Inverse Document Frequency (TFIDF) is one of the most important features in word frequency analysis. Term frequency measures how frequently a word appears in a specific text document, and inverse document frequency is the inverse frequency of the same word in all documents. The weights are calculated as the following formula:

where is the weight of term within document , is the frequency of in , is the total number of documents and is the number of documents containing . The words that appear frequently are assigned lower weights, and the words that are unusual are assigned higher weights. *TfidfTransformer* in *scikit-learn* is utilized to compute the TFIDF vectors.

For each document, the TFIDF vector is calculated. Each element of the vector represents one word, and its value is calculated by multiplying its frequency by the corresponding weight. The value will then be normalized by the total number of words in a document. The TFIDF score is the sum of the elements in the TFIDF vector. A higher TFIDF score indicates a more unusual pattern of the word’s appearing in a document.

* *Song Lyrics Data Set*

Figure 3-1 shows the cosine distance of TFIDF vectors between each pair of the selected artists and Figure 3-2 shows the boxplot of the TFIDF scores for each song of the selected artists.

图表

描述已自动生成

**Figure 3-1:** Cosine Distance Between TFIDF Vectors of Each Pair of Artists

图表

描述已自动生成

**Figure 3-2:** The TFIDF Score of Songs of 25 Selected Artists

Overall, the similarity of TFIDF for different artists is very high. The lyrics of Harry Styles and Olivia Rodrigo are most distinguishable compared with all other selected artists. However, the distributions of TFIDF scores of songs of different artists vary, and particularly, Drake has a higher median and variance, which means that he tends to use more unusual words in his lyrics.

* *News Data Set*

Figure 3-3 shows the cosine distance of TFIDF vectors between each pair of the selected classes from the News Data Set and Figure 3-4 shows the boxplot of the TFIDF scores for each article from the selected classes.

图片包含 背景图案

描述已自动生成

**Figure 3-3:** Cosine Distance Between TFIDF Vectors of Each Pair of Classes

图表, 箱线图

描述已自动生成

**Figure 3-4:** The TFIDF Score of News Articles from 10 Selected Classes

Compared to the Song Lyrics Data Set, the similarity between different classes is very low, except that the similarity of articles from “FIFTY” and “PARENTS” is higher than others. One potential reason is that readers that are in their fifties are usually parents, so there is a strong relationship between the two types of news articles.

## Simple Sentimental Analysis

In recent years, sentimental analysis has been one of the most popular fields in NLP. One key aspect of sentimental analysis is to “understand” the emotions, contents, and opinions hidden in the texts, which is typically measured by what is called polarity. Subjectivity and emotional tendencies are the essential features for song lyrics and news article. Different from polarity, subjectivity is a value that measures the degree and strength of the author’s sentiment and feelings in the text. In this project, the *TextBlob* library in *Python* is utilized to obtain the values of polarity and subjectivity with respect to a segment of text.

* *Song Lyrics Data Set*

Figure 3-5 shows the polarity and subjectivity of different songs, where each point represents a song and different colors of points represent different artists.

图表, 散点图

描述已自动生成**Figure 3-5:** Polarity and Subjectivity of Songs

It can be observed that there is not a specific area that is dominated by one artist, so there is no obvious difference in the polarity and subjectivity of songs from different artists. This is probably because the most popular artists tend to try songs of different styles, emotions, and levels of subjectivity.

* *News Data Set*

Figure 3-6 shows the polarity and subjectivity of different news articles, where each point represents an article and different colors of points represent different categories of articles.

图表, 散点图

描述已自动生成

**Figure 3-6:** Polarity and Subjectivity of News Articles

Compared to the Song Lyrics Data Set, it is noticeable that in the News Dataset, there are certain colors of points, for example, the points for news articles from “FIFTY”, do appear close to each other. However, for many other categories of articles, the points from the same class does not appear in cluster, since a specific category of news articles may be different in polarity and subjectivity based on the concrete events being discussed.

## **Part of Speech**

The part of speech structure is a feature that reflects the style of each artist and each type of news articles. In this project, the *nltk* library is used to check the part of speech for each word, and the ratio of nouns, adjectives, verbs, and adverbs of each document is calculated. Figure 3-7 and Figure 3-8 shows the ratio of nouns from the randomly selected classes from the Song Lyrics Dataset and the News Dataset as the example, respectively. Furthermore, the differentials, standard deviations, and mean of the ratios are also considered as features for classification.

图表, 箱线图

描述已自动生成

**Figure 3-7:** The Ratio of Nouns in the Song Lyrics Data Set

图表, 箱线图

描述已自动生成

**Figure 3-8:** The Ratio of Nouns in the News Data Set

## Number of Words

Number of words is a feature that reflects some part of writing style of an artist or the typical length of a type of news article. Figure 3-9 and Figure 3-10 show the distribution of the number of words for each class in the two data sets.

图表, 箱线图

描述已自动生成

**Figure 3-9:** Number of Words in Song Lyrics Data Set

图表, 箱线图

描述已自动生成

**Figure 3-10:** Number of Words in News Data Set

It is noticeable that there are some artists, such as Drake, tend to use more words in a song. However, the number of words does not vary greatly as the lyric dataset for the News Dataset.

## Repeated Words

Repeated words may be a way to express strong emotions or to put the emphasis on certain stories or events, so the ratio of unique words can also be considered as a feature for classification. The ratio ranges from 0 to 1, and the larger the ratio, the less repeated words there are in a song.

Figure 3-11 is a boxplot that shows the ratio of unique words from randomly selected artists. The ratio is typically between 0.2 and 0.4, but there are some artists, such as Coldplay, Oasis, and Queen, who use less repeated words in their songs.

图表, 箱线图

描述已自动生成

**Figure 3-11:** Ratio of Unique Words in Song Lyrics Data Set

Figure 3-12 shows the ratio of unique words from different categories of news articles. Normally, news articles tend to use less repeated words, since most news articles are short and avoid the use of repeated expressions.

图表, 箱线图

描述已自动生成

**Figure 3-12:** Ratio of Unique Words in News Data Set

## Features Used in Classification

The features used in classification for traditional machine learning models are basically composed of stems, TFIDF, polarity, part of speech and class related features. Normally the means, standard deviation, differential and the differential of the standard deviation in the same class are considered.

Table 3-1 and Table 3-2 show the features used in classification for traditional machine learning models in two data sets. Note that “std” refers to standard deviation, and “diff” refers to the differential value. There is no “stem\_per\_line” related feature in the News Data Set, since the collected data has no line break sign “\n”.

* Song Lyric Data Set

**Table 3-1:** Features for Song Lyric Dataset

|  |  |
| --- | --- |
| **Feature Category** | **Name of Features** |
| **Stems** | n\_stems\_mean, n\_stems\_std, unique\_stems\_ratio\_mean, unique\_stems\_ratio\_std, stems\_per\_line\_mean, stems\_per\_line\_std, n\_stems, unique\_stems\_ratio, stems\_per\_line, n\_stems\_diff, n\_stems\_diff\_std, unique\_stems\_ratio\_diff, unique\_stems\_ratio\_diff\_std, stems\_per\_line\_diff, stems\_per\_line\_diff\_std |
| **TFIDF** | tf\_idf\_score\_mean, tf\_idf\_score\_std, tf\_idf\_vector\_mean, tf\_idf\_vector, tf\_idf\_score, tf\_idf\_score\_diff, tf\_idf\_score\_diff\_std, vector\_similarity |
| **Polarity** | polarity\_mean, polarity\_std, polarity, polarity\_diff, polarity\_diff\_std |
| **Part of Speech** | p\_stop\_words\_mean, p\_stop\_words\_std, p\_noun\_words\_mean, p\_noun\_words\_std, p\_adj\_words\_mean, p\_adj\_words\_std, p\_verb\_words\_mean, p\_verb\_words\_mean, p\_verb\_words\_std, p\_adv\_words\_mean, p\_adv\_words\_std |
| **Artist** | artist\_artist, artist\_song, song, artist\_set\_id |

* News Data Set

**Table 3-2:** Features for News Dataset

|  |  |
| --- | --- |
| **Feature Category** | **Name of Features** |
| **Stems** | n\_stems\_mean, n\_stems\_std, unique\_stems\_ratio\_mean, unique\_stems\_ratio\_std, n\_stems, unique\_stems\_ratio, n\_stems\_diff, n\_stems\_diff\_std, unique\_stems\_ratio\_diff, unique\_stems\_ratio\_diff\_std, |
| **TFIDF** | tf\_idf\_score\_mean, tf\_idf\_score\_std, tf\_idf\_vector\_mean, tf\_idf\_vector, tf\_idf\_score, tf\_idf\_score\_diff, tf\_idf\_score\_diff\_std, vector\_similarity |
| **Polarity** | polarity\_mean, polarity\_std, polarity, polarity\_diff, polarity\_diff\_std |
| **Part of Speech** | p\_stop\_words\_mean, p\_stop\_words\_std, p\_noun\_words\_mean, p\_noun\_words\_std, p\_adj\_words\_mean, p\_adj\_words\_std, p\_verb\_words\_mean, p\_verb\_words\_mean, p\_verb\_words\_std, p\_adv\_words\_mean, p\_adv\_words\_std |
| **class** | class\_class, class\_head, head, class\_set\_id |

# **Methods**

Machine learning models are adopted to perform text classification. Both traditional machine learning models and some state-of-the-art deep learning models are included. Note that the deep learning models are only experimented on the Song Lyrics Data Set.

## 4.1 Logistic Regression (LR)

Logistic Regression (LR) is a supervised machine learning algorithm that models the discrete outcome. LR uses generalized linear model for classification. While LR is mostly used for binary classification, multinomial LR can perform multi-classification.

Multinomial LR predicts the probability of each outcome via a linear function:

where is the regression coefficient vector of outcome , and is the feature vector of the -th input. Suppose there are possible outcomes, the probability estimation of the-th outcome is:{\displaystyle f(k,i)=\beta \_{0,k}+\beta \_{1,k}x\_{1,i}+\beta \_{2,k}x\_{2,i}+\cdots +\beta \_{M,k}x\_{M,i},}

*=*

The coefficient vector is usually estimated by maximizing the posteriori probability estimation, and log-likelihood function is a common choice to maximize the probability estimation.

## 4.2 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that is believed to be one of the most robust prediction methods. SVM is a non-probabilistic classifier, and it performs linear classification by maximizing the distance between the boundaries and the samples**.**

In SVM, the training samples are correctly classified with a nonzero margin, which is the distance between the boundaries and the sample. For the -th sample , there is:

The support vector is then introduced to minimize the regular penalty subject to for each sample in the set.

Apart from linear classification, SVM can also map features into higher dimensions via kernel functions, and the linear classifier is later applied to the mapped features.

## 4.3 Random Forest (RF)

Random Forest (RF) is a machine learning model constructed by ensembling a multitude of decision trees. To perform classification, the output of RF is the class that is selected by most trees. Assume the number of trees is , the algorithm of RF is shown below:

**Algorithm 4-1: Random Forest**

|  |
| --- |
| *for=1 to:*  *Draw a bootstrap sample of size from the training data*  *= number of variables*    *= 1*  *Grow a random-forest tree to the bootstrapped data by recursively repeating the following steps on each terminal node until the minimum node size is reached.*  *(i) Select variables randomly out of the variavles*  *(ii) Choose the best variable splitting point out of the variables*  *(iii) Split the node into two daughter nodes* |

## 4.4 AdaBoost

AdaBoost (Adaptive Boosting) is a ensemble learning model that improves weak classifiers by combining the outcomes of weak classifiers into a weighted sum that represents the final output of the boosted classifier. The classifiers are learned sequentially, and the training data is reweighted for the next classifier.

Assume there are weak classifiers and given the training data , the algorithm of AdaBoost is described as follow:

**Algorithm 4-2: AdaBoost**

|  |
| --- |
| *for =1 to :*  *Train weak classifier with {}*  *Calculate the error rate*  *for =1 to :*    *If is misclassified by :*  *=*  *Else:*  *=*  *Obtain the set of functions {}*  *Aggregate the functions* |

There are multiple choices to aggregate the classifiers, including uniform weight aggregation and non-uniform weight aggregation.

## 4.5 Voting (LR, SVM, RF)

The Voting Classifier is provided by scikit-learn 1.0.2. In our experiment, soft voting, which ensembles several classifiers and returns the average of predicted probabilities, is adopted.

## 4.6 TextCNN (Text Convolution Neural Network)

### 4.6.1 Words to Vectors

The *glove.6B.300d* in *pytorch.vocab* dictionary is used for converting words into vector. The processing of text can be simplified into vector operations in the vector space, and the similarity in the vector space can be calculated to represent the semantic similarity of the text. The similarity between two vectors is measured by the cosine of the angle between them. During comparison, only the direction of the vector is considered.

### 4.6.2 Neural Network Structure

The core idea of convolutional neural network is to capture local features, which are sliding windows composed of several words for text data. The advantage of convolutional neural network is that it can automatically combine and filter N-gram features to obtain semantic information at different levels of abstraction. The following table shows the structure of the neural network.

**Table 4-1:** Neural Network Structure of the TextCNN

|  |  |
| --- | --- |
| **(embed)** | Embedding(15707, 300) |
| **(convs)** | Conv2d(1, 100, kernel\_size=(3, 300), stride=(1, 1)) |
| Conv2d(1, 100, kernel\_size=(4, 300), stride=(1, 1)) |
| Conv2d(1, 100, kernel\_size=(5, 300), stride=(1, 1)) |
| **(dropout)** | Dropout(p=0.2, inplace=False) |
| **(fc)** | Linear(in\_features=300, out\_features=41, bias=True) |

The embedding layer is used to convert the words into the vectors defined in the dictionary. The above embedding layer used a dictionary with 15707 words and convert each word into a 300-dimension vector. After embedding, 3 convolution layers are used with kernel size (3, 300), (4, 300) and (5, 300) respectively to capture the local features.

## 4.7 Transformer

### 4.7.1 Words to Vectors

The *pytorch.vocab* dictionary is used for converting all words that the frequency is above 2 into vector.

### 4.7.2 Neural Network Structure

Transformer is an encoder-decoder architecture that uses the Attention as primary mechanism, which allows the neural network to pay more attention to relevant information. The structure of Transformer is shown in Figure 4-1.

图示

描述已自动生成

**Figure 4-1**: The Structure of Transformer

Transformer adopts the design of key-value Attention, using matrix (query), (Key), and (Value) for Attention, which is defined as:

where ，，.

The multi-head Attention based on Attention is defined as:

where ，，.

# **5. Results**

## 5.1 Traditional Machine Learning Models

To evaluate the performance of the traditional machine learning models, “top-n” is used as the accuracy metric in a subset with specific number of classes, where the number of texts in the training set and in the test set are even.

* *Song Lyrics Data Set*

Table 5-1 demonstrates the accuracy of different traditional machine learning models for 16 random selected artists per set with different values for “top-n”. The differences between results from different models are not significant, but Logistic Regression has the best performance among all models

**Table 5-1:** The Accuracy of Models for the Lyric Dataset

|  |  |  |
| --- | --- | --- |
| **Machine Learning Models** | **top - n** | **Accuracy** |
| **Logistic Regression** | **1** | **0.30** |
| **2** | **0.48** |
| **4** | **0.66** |
| **8** | **0.90** |
| **Support Vector Machine** | 1 | 0.22 |
| 2 | 0.30 |
| 4 | 0.56 |
| 8 | 0.86 |
| **Random Forest**  **(Number of estimators = 10)** | 1 | 0.08 |
| 2 | 0.10 |
| 4 | 0.18 |
| 8 | 0.50 |
| **AdaBoost**  **(Number of estimators = 100)** | 1 | 0.22 |
| 2 | 0.32 |
| 4 | 0.50 |
| 8 | 0.78 |
| **Voting Classifier**  **(LR, RF, SVM)** | 1 | 0.22 |
| 2 | 0.36 |
| 4 | 0.54 |
| 8 | 0.82 |

* *News Data Set*

Table 5-2 demonstrates the accuracy of different traditional machine learning models for 32 categories per set with different values for “top-n”. It is obvious that compared to the Song Lyrics Dataset, the models have shown remarkable results in this data set, with the highest accuracy of top-n=1 reaching 88%.

**Table 5-2:** The Accuracy of Models for the News Dataset

|  |  |  |
| --- | --- | --- |
| **Machine Learning Models** | **top - n** | **Accuracy** |
| **Logistic Regression** | 1 | 0.86 |
| 2 | 0.94 |
| 4 | 0.96 |
| 8 | 0.98 |
| **Support Vector Machine** | 1 | 0.86 |
| 2 | 0.92 |
| 4 | 0.96 |
| 8 | 0.99 |
| **Random Forest**  **(number of estimators = 10)** | 1 | 0.87 |
| 2 | 0.94 |
| 4 | 0.97 |
| 8 | 0.99 |
| **AdaBoost**  **(number of estimators = 100)** | **1** | **0.88** |
| **2** | **0.96** |
| **4** | **0.96** |
| **8** | **0.98** |
| **Voting Classifier**  **(Logistic Regression + Random Forest + Support Vector Machine)** | 1 | 0.85 |
| 2 | 0.92 |
| 4 | 0.94 |
| 8 | 0.97 |

## **5.2 Deep Learning Models**

The deep learning models are only used on the Song Lyrics Data Set, where 80% of the Lyric dataset is randomly separated and the rest of the 20% is used as the validation & test set.

### 5.2.1 TextCNN

Figure 5-1 shows the train and validation accuracy.

图表, 折线图

描述已自动生成

**Figure 5-1:** Train/Validation Accuracy for Song Lyric Data Set

The train accuracy quickly converges to 100% after 20 epochs. However, the validation accuracy is about 31%, which shows significant problem of over-fitting. This may be a result from imbalanced data and the limited size of the data set.

### **5.2.2 Transformer**

To evaluate the performance of *Transformer*, 20% of the dataset is randomly split as the validation set, and the others are used as the training set. Some relevant results are shown as follows. Table 5-2 shows the result of Transformer on the Song Lyrics Data Set. Figure 5-2 and Figure 5-3 shows the curves of F1-score and loss, respectively.

**Table 5-2:** The Result of Transformer on Song Lyrics Data Set

|  |  |
| --- | --- |
| **Test Loss** | 3.3735909461975098 |
| **Test F1 Score** | 0.052619909802745626 |
| **Test Accuracy** | 0.1 |

图表, 折线图

描述已自动生成

**Figure 5-2:** F1 Score of the Train & Test Set

图表, 折线图

描述已自动生成

**Figure 5-3:** Loss of the Train & Test Set

In this case, this model fails to converge. Compared to other models, the result of Transformer is obviously degraded, with accuracy and F1 score being very low.

As we can see, for Song Lyrics Data Set, the state-of-the-art deep learning models do not show dominant advantages over traditional models, this may be a result from the limited regularity of the song lyrics data. We will provide some further analysis and discussion in the next section.

# **Discussion**

In this section, we will give some discussion based on previous sections. We will compare the differences between data sets and the performance of models, as well as providing some possible interpretation.

## **Data**

We conduct experiments on two different data sets, the Song Lyrics Data Set and the News Data Set. There are some significant differences between news articles and song lyrics. On the one hand, news articles tend to use very formal languages, while songs usually prefer colloquial expressions. On the other hand, news articles in a specific category usually have clearly specified topics, with compact contents and clear emotions. Songs written by one artist may have various styles, without specific contents and stories, and with complicated emotions tendencies. Therefore, the relation between song lyrics and the artists may not be as strong.

With the differences between the two data sets and the differences in quality of features, the performance of models between data sets can also be expected. Also, in both data sets, the data among classes appears to be imbalanced, which may also limit the performance.

## Features

The performance of models, especially traditional machine learning models, depends largely on features of data. With mainstream feature engineering methods, the quality of extracted features is generally very low and do not show high variability in Song Lyric Dataset.

There are potential reasons behind this problem. The style of an artist is usually not limited, and most songs, regardless of artists, tend to use catchy and classic words. Without the rhythm of music, with only lyrics, a lot of information in the songs is lost, making it extra hard for features to be extracted.

## 6.3 Models

We adopt both traditional machine learning models and deep learning models for classification.

In the Song Lyrics Data Set, different traditional machine learning models do not show significant variance in their performances. Of all traditional machine learning models, Logistic Regression has given the best results. The deep learning model TextCNN suffers from obvious overfitting problems. However, Transformer, a deep learning model with more complex structure that is expected to discover the deeper features inside the lyrics, yields nothing but a worse result. With the data being imbalanced and insufficient, and with the regularity between song lyrics and artists being weak, the precise classification tasks seem extra difficult on this data set. Hence, the “top-n” accuracy metric is used to provide the reference basis for recommending the artists with similar styles.

In the News Data Set, the traditional machine learning models have shown remarkable performances. When top-n=1, the highest accuracy has achieved 88%. When top-n=8, the highest accuracy is 99%. Apart from the fact that this is a larger data set compared to the Song Lyrics Data Set, the length of news articles is generally longer than song lyrics, and training machine learning models on this data set is a lot more costly in computation. We did not use deep learning models on this data set because with our laptops, training deep learning models on this data set is way too time and resource consuming.

## **Potential Improvements**

Based on the methods, results and our analysis, there are several potential improvements to this project.

For both Song Lyrics Data Set and News Data Set, the data are imbalanced, which may introduce bias to the models. Improving the quality of data may also improve the performance of models. Thus, one potential improvement is to add more data to the dataset, especially for those classes that have a smaller number of documents, while for some artists they do not have enough song works for the machine learning.

Our experiment has shown that the quality of features has a big role to play in the performance of machine learning models. For Song Lyrics Data, it seems that mainstream feature engineering methods fail to extract high quality features from the data. There are currently a growing amount of research focusing on NLP and feature engineering, especially sentimental analysis, so some of the most up-to-date feature engineering methods may be able to help extract better features.

In the News Data Set, a piece of news only has one label. However, in real life, a piece of news may be associated to more than one label. If multiple labels can be introduced to some new articles, the models may have a bigger chance to correctly classified them.

# References

1) Engel, J. (1988). "Polytomous logistic regression". Statistica Neerlandica. **42** (4): 233–252. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1111/j.1467-9574.1988.tb01238.x](https://doi.org/10.1111%2Fj.1467-9574.1988.tb01238.x).

2) Bennett, Kristin P.; Campbell, Colin (2000). "Support Vector Machines: Hype or Hallelujah?". SIGKDD Explorations. 2 (2): 1–13. doi:10.1145/380995.380999. S2CID 207753020.

3) Ho, Tin Kam (1995). [Random Decision Forests](https://web.archive.org/web/20160417030218/http:/ect.bell-labs.com/who/tkh/publications/papers/odt.pdf). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

4) Freund, Yoav; Schapire, Robert E (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". Journal of Computer and System Sciences. **55**: 119–139.

5) [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

6) He, Tong, et al. "Text-attentional convolutional neural network for scene text detection." *IEEE transactions on image processing* 25.6 (2016): 2529-2541.

7) Vaswani A., et al. 2017. Attention is all you need. In NIPS, 5998–6008.