Exploiting Features for Gender Prediction of Artists Using Lyrics

|  |  |
| --- | --- |
| 000  001  002  003  004  005  006  007  008  009  010  011  012  013  014  015  016  017  018  019  020  021  022  023  024  025  026  027  028  029  030  031  032  033  034  035  036  037  038  039  040  041  042  043  044  045  046  047  048  049 | 050  051  052  053  054  055  056  057  058  059  060  061  062  063  064  065  066  067  068  069  070  071  072  073  074  075  076  077  078  079  080  081  082  083  084  085  086  087  088  089  090  091  092  093  094  095  096  097  098  099 |

|  |
| --- |
|  |
| Babafemi Oyinlola |
|  |
|  |
|  |
|  |
|  |

Abstract

This document contains the instructions for preparing a camera-ready manuscript for the proceedings of ACL-2018. The document itself conforms to its own specifications and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

# Introduction

Music is a form of art accepted across various age groups, gender, race and even, religion. It is an ancient highly valued feature of all known living creatures, pervading many aspects of daily life, playing many roles (Killin, 2017) [1]. A song contains lyrics and melody; lyrics has been used in natural language processing to perform a number of tasks such as classifying genre (Fell and Sporleder, 2014), artist recognition (Eghbal-Zadeh et al., 2015), sentiment analysis (Gomez and Caceres), information retrieval [5], annotation [6], automatic generation of song lyrics [7] however, no work has been identified which classifies the gender of artists. Hence, this work develops a model that classifies the gender of an artist from the lyrics of the artist’s song as “*female”* or *“male”* by exploiting a varied set of features such as lexical, syntactic, semantic and sentiment using text classification techniques. Text classification is used to automatically classify text documents into predefined classes. It has applications such as selective dissemination of information to consumers, spam filtering, filing patents into patent directories. The goal of this research is to identify the most important features required to successfully classify the gender of an artist.

# Related Work

In recent years, there has been an explosion of automatic text classification which is as a result of the increased availability of documents in digital form and the ensuing need to organize them [15]. This work garners from previous text classification research.

In 2014, Fell and Sporleder [2] conducted experiments to analyze lyrics and detect the genre of music the song belongs using the Weka [16] implementation of support vector machines with the default setting. With a focus on eight genres, they compared results obtained from an extended model with vocabulary, style, orientation, semantics and song structure as features to a baseline model using only vocabulary as features. Using F-score as the evaluation metric, the baseline model was seen to perform better in predicting seven of the eight genres than the extended model, however, a model with a combination of all features outperformed the other two models.

In this work, the focus is the categorization of a music artist’s gender based on lyrics which is similar to the aim of Sboev et al. [17]. In their research, they developed models for classifying text according to the author’s gender although, done with a corpus of Russian-language texts. They identified a group of features such as syntactic, morphological and emotions which they used to train various machine learning algorithms while employing deep learning. They concluded that the author’s gender is conveyed through specific syntactical and morphological patterns and, use of emotions words.

Schler et al. [18] analyzed a corpus obtained from blogs of male and female bloggers and exploited the significant differences in writing style to determine an author’s gender and age based on the blog’s vocabulary. Using stylistic features such as parts-of-speech, function words, blog word, and hyperlinks which summed to a total of

502 features in all and, a Multi-Class Real Winnow (MCRW) learning algorithm. Their analysis concluded that most teenage bloggers are female while older bloggers are predominantly male and, while female bloggers discuss their personal lives, male bloggers write more about politics, technology, and money.

# Data and Methodology

|  |  |
| --- | --- |
| 100  101  102  103  104  105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130  131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149 | 150  151  152  153  154  155  156  157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182  183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199 |

Names of rhythm and blues (R&B) artists were scraped from 1Billboard’s Hot 100 archive and lyrics to songs of these artists retrieved from 2Musixmatch’s database using the API available. Songs were collected by obtaining all tracks for every album of an artist and removing songs by bands and, songs collaborated with other artists. Each artist was manually annotated obtaining a dataset including an artists’ gender proved difficult. A total of 41,157 lyrics were obtained of which 17,572 are by female artists and 23,585 by male artists.

In the classification experiments conducted, different models were created to exploit the different feature groups identified. The objective of the model in this work is to predict the gender of an artist from the artist’s lyrics. 10,000 songs were used to train the classification models and 31,157 songs used to train the **word2vec (COME BACK HERE)** model used to generate word embeddings. For the classification models, the dataset was split into an 80% training and 20% test set for evaluation which results in a training set of 8,000 songs and a held-out test set of 2,000 songs.

## Pre-processing

In other to build successful models and extract meaningful features, proper pre-processing is performed to remove noise in data. The pre-processing steps taken are:

* Replacing apostrophes: as expected, in lyrics, most especially in the R&B genre, some words are shortened and combined by artists using an apostrophe. Without proper conversion, these words would not give the necessary information needed. Some examples of words handled are: *‘lone, y’all, I’mma, ‘cause.*
* Removing stop words: Frequently occurring words such as *“the*”*, “to”*, *“a”* are filtered from the lyrics as they usually have little lexical content and their presence in a text fails to distinguish it from other texts [8].
* The lyrics of every song obtained from the Musixmatch database is appended with

*“\*\*\*\*\*\*\* This Lyrics is NOT for Commercial use \*\*\*\*\*\*\*”*. This phrase was also removed from the lyrics of every song in the dataset.

* Tokenization: lyrics are converted into tokens by breaking into words to obtain a list of words.

## Feature Extraction

As stated earlier, this work exploits different feature classes to determine the most important feature for correctly predicting the gender of an artist.

* Lexical**:** As lexical features, bag of words (a count of each word in the lyrics), bigrams, trigrams, character trigrams and character fourgrams where incorporated. In addition, numeric features such as the average token length and number of function words were included summing. The bag of words in this work, the bag of words is used as the baseline and the others as the lexical feature group.
* Syntactic: The NLTK (reference this) POS (Part of Speech) tagger was used to annotate the tokens. The four features considered are (i) the frequency of all POS tags (ii) the frequency of adjectives (iii) the frequency of verbs and (iv) the frequency of personal pronouns. The features were selected as they are expected to some the most informative in distinctively classifying gender. **(For examples, the artist tends to use a personal pronoun of the opposite gender------think before adding this)**
* Semantic:Word2Vec [9] was used to obtain semantic information from the lyrics, by generating word embeddings from a separate dataset of 66,050 songs using the Python module – Gensim **(referenceeee)**. Word embeddings are distributed word representations that embed every word into a low dimensional real-valued vector space which are assumed to convey semantic information of words [10]. Such dense representations can be learned from data, and they have proven to be effective in improving the performance of many natural language processing tasks [11]. In training the model, the size of the dense vector to represent each word was set to 200, the window which is the maximum distance between a target word and its neighboring word was set to 10 and the workers (threads) was also set to 10. For every song’s lyrics, the mean of the vector of each token was obtained and used as the feature.

**(give more explanation here in a more easily comprehendible form)**

* Sentiment Lexicon: Two existing sentiment lexicons were employed (i) AFINN (Nielsen, 2011) was used to calculate the overall polarity of a song. The overall polarity of a song is classified as either positive, negative or neutral (ii) Opinion lexicon (Hu and Liu, 2004**)** was used to obtain the number of positive and negative lexicon words in every song. The number of positive and negative lexicons were average over the length of lyrics**.**

## Experiment

A binary classification experiment was conducted using Multinomial Naïve Bayes as the classification algorithm. Naïve Bayes is a machine learning algorithm whose classification efficiency is proved in applications such as document categorization and e-mail spam filtering. It learns through a document classification algorithm and is based on simple usage of the Bayes’ rule [12]. The multinomial Naïve Bayes models the distribution of words in a document as a multinomial. It was selected because, it has proved to be good enough for text classification as in ([13, 14] look for papers to reference for this).The hyper-parameters (alpha and fit prior) for the classifiers were tuned for each model exploiting a feature group by means of a cross-validated grid search on all training data. During optimization of these parameters, the values of the additive smoothening parameter, alpha, was varied between 0.5 and 6 while the boolean, fit\_prior, that indicates if prior class probabilities should be learned is tuned to obtain either true or false. The optimized parameters obtained were used in training the final models which were tested on the held-out test set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Features** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| BoW(baseline) | 77.60 | 78.35 | 77.60 | 77.45 |
| Lexical | 83.77 | 83.97 | 83.77 | 83.75 |
| Syntactic | 57.15 | 57.37 | 57.15 | 56.82 |
| Sentiment | 67.29 | 67.99 | 67.29 | 66.96 |
| Semantic | 78.44 | 79.36 | 78.44 | 78.27 |
| Combined | **84.11** | 84.15 | 84.11 | 84.10 |

Table 1: Experimental results on held-out test set

# Results

In this section, the results obtained from the experiments conducted are presented. Metrics of each model for each feature group and the combined model with all features is shown and compared to the baseline**.** In addition, some insights obtained from the data are discussed.

* 1. **Metrics obtained from the Held-out Test Set**

Table 1 shows the 5-fold cross validation results for each model. We compare the results obtained for all models with the baseline to understand their performance. As shown in Table 1, the baseline performs better than the models with sentiment and syntactic features with an accuracy of 76.6%. The model trained using only lexical features yields an accuracy of 83.77% with the second best (model with semantic features) having a 78.44% accuracy. Evidently, the model with the combined feature set performs best with an accuracy up to 84.11% which is only <1% better than the model with lexical features. These results indicate that the grammatical construction of lyrics in songs and the relationship between words in lyrics are not very relevant in predicting the gender of an artist as the model trained yields results that are only slightly better than guessing. Although better than syntactic features, the overall sentiments of songs as positive, negative or neutral, perform better by more than 10%. Amongst the feature groups considered, lexical information obtained from bigrams, trigrams and character n-grams prove to be the of key importance in the classifying task. An explanation for this could be that although make male and female artists might have similar words in lyrics, these words are constructed and arranged differently for each gender when composed into songs.

**TALK ABOUT SIGNIFICANCE TESTS**

## Other information and findings

Figures 1 and 2 display word cloud images of words used by female and male music artists respectively. Further analysis show that *love, baby, know, oh* and *like are the* most frequently occurring words found in the corpus for female artists. Similarly *love, know, baby, like* and *got* are the most occurring words used my male artists which is quite interesting because four of these the most frequently used words by both sexes which is not surprising for R&B artists. Additionally, the overall sentiment of each song was obtained using AFINN (Nielsen, 2011) and classified as positive, negative and neutral. While a higher percentage of songs were classified as positive with 72.42% of songs by female artists and 64.81% of songs by male artists. 22.51% of songs by female artists were classified as negative while 30.72% of songs by male artists classified as negative. Only a small proportion of songs were classified to have neutral sentiments of which 5.06% of these were songs by female artists and 4.46% of these were songs by male

artists.

Figure 1: Word cloud of songs by female artists

A picture containing text, newspaper

Description automatically generated

Figure

Figure

|  |  |
| --- | --- |
| 200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249 | 250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299 |

|  |  |
| --- | --- |
| 200  201  202  203  204  205  206  207  208  209  210  211  212  213  214  215  216  217  218  219  220  221  222  223  224  225  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240  241  242  243  244  245  246  247  248  249 | 250  251  252  253  254  255  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285  286  287  288  289  290  291  292  293  294  295  296  297  298  299 |

A blue and white text

Description automatically generated

Figure 2: Word cloud of songs by male music artists

**A screenshot of a cell phone

Description automatically generated**

Figure 3: Overall polarity of songs by gender

# Conclusion and Future Work

In this paper models were developed and tested to classify the gender of artists based on their lyrics. These models were built by exploiting four feature groups and a combination of all features. Of all feature groups, group providing lexical information yielded the highest metrics

Future work should include calculating the proportion/contribution of each feature in the model to obtain exact values.

Limitations would be the availability of data but time to train it too long so we reduced it. Proper tuning can be done further.

|  |  |
| --- | --- |
| 300  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345  346  347  348  349 | 350  351  352  353  354  355  356  357  358  359  360  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390  391  392  393  394  395  396  397  398  399 |

References

Michael Fell and Caroline Sporleder. Lyrics-based analysis and classification of music. In *COLING*, volume 2014, pages 620-631, 2014.

Hamid Eghbal-zadeh, Markus Schedl, and Gerhard Widmer. Timbral modelling for music artist recognition using i-vectors, In *EUSIPCO*, 2015.

Lucia Martin Gomez, Maria Navarro Caceres. Applying data mining for sentiment analysis in music. In *Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection – 15th International Conference, pages 312 – 325.* PAAMS, 2017.

Alfred. V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling, volume 1*. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual.* American Psychological Association, Washington, DC.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. [Alternation](http://dl.acm.org/citation.cfm?doid=322234.322243). *Journal of the Association for Computing Machinery*, 28(1):114-133. https://doi.org/10.1145/322234.32224.

Association for Computing Machinery. 1983. *Computing Reviews*, 24(11):503-512.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. [Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing](http://aclweb.org/anthology/P16-1001). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1–11. <https://doi.org/10.18653/v1/P16-1001>.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mary Harper. 2014. [Learning from 26 languages: Pro- gram management and science in the babel program](http://aclweb.org/anthology/C14-1001). In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, page 1. <http://aclweb.org/anthology/C14-1001>.

Alexander V. Mamishev and Murray Sargent. 2013. *Creating Research and Scientific Documents Using Microsoft Word*. Microsoft Press, Redmond, WA.

Alexander V. Mamishev and Sean D. Williams. 2010. *Technical Writing for Teams: The STREAM Tools Handbook*. Wiley-IEEE Press, Hoboken, NJ.

|  |  |
| --- | --- |
| 400  401  402  403  404  405  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435  436  437  438  439  440  441  442  443  444  445  446  447  448  449 | 450  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495  496  497  498  499 |