Exploiting Features for Gender Prediction of Artists Using Lyrics

Babafemi Oyinlola
32340424

Abstract

Text classification is a common task in natural language processing which is widely performed using the Naïve Bayes algorithm for classifying. It is recognized that male and female artists sing differently hence, this work explores four feature groups (lexical, semantic, syntactic and, sentiment) to predict the gender of artists. Experiments conducted show that a model trained with only lexical features obtained from lyrics performed best with an accuracy and F1 score of 86.06%.

1 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, 2018). 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 (Brants, 2003), annotation (Nam et al., 2012), automatic generation of song lyrics (Pudaruth et al., 2014) 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 lyrics 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.

2 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 (Sebastani, 2002). This work garners from previous text classification research. In 2014, Fell and Sporleder (Fell and Sporleder, 2014) conducted experiments to analyze lyrics and detect the genre of music a song belongs using the Weka (Hall et al., 2009) 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. (2016). 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. (2006) 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 as parts-of-speech, function words, blog words 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.

3 Data and Methodology

Names of rhythm and blues (R&B) artists were scraped from Billboard's[1] Hot 100 archive and the lyrics to songs of these artists retrieved from Musixmatch's[2] 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. The gender of each artist was manually annotated because, obtaining a dataset including artists' gender proved difficult. A total of 41,157 songs 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 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 resulted in a training set of 8,000 songs and a held-out test set of 2,000 songs.

3.1 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 stopwords: 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 (Bird et al., 2009).
- The lyrics of every song obtained is appended with *"******* This Lyrics is NOT for Commercial use *******"* which 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.

3.2 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 to six features. The bag of words is used as the baseline and the others as the lexical feature group.
- Syntactic: The NLTK[3] 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 were expected to contain some informative in distinctively classifying gender (R&B artists tend to use a personal pronoun of the opposite gender).
- Semantic: Word2Vec (Mikolov et al., 2013) was used to obtain semantic information from the lyrics, by generating word embeddings from 31,157 songs using the Python open source library – Gensim[4]. 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 (Li et al., 2015). 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 (Cao and Lu, 2017). In training the model, the Continuous bag of Words (CBoW) model was applied, the size of the dense vector to represent

[1] https://www.billboard.com
[2] https://developer.musixmatch.com
[3] Natural Language Toolkit – https://www.nltk.org
[4] https://radimrehurek.com/gensim

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.

- 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**.**

3.3 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 (Ramasubramanian and Singh, 2017). The multinomial Naïve Bayes models the distribution of words in a document as a multinomial. It was selected because, it has been proved to be good enough for text classification as in (Buzic and Dobsa, 2018; Rennie et al., 2003). 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.

4 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.

4.1 Metrics obtained from the Held-out Test Set

Table 1 shows the 10-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 77.34%. The model trained using only lexical features yields an accuracy of 86.06% while the second best (model with semantic features) a 77.34% accuracy. Surprisingly, performance of the model with the combined feature set was not as good as model with only lexical features. The accuracy and F-score obtained is 1.15% less at 84.91% 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 give 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 most important feature set in this classification task. An explanation for this could be that although male and female artists might have similar words in lyrics, these words and arranged differently for each gender when composed as songs.

Features	**Accuracy**	**Precision**	**Recall**	**F_1-Score**
BoW(baseline)	77.34	78.17	77.34	77.17
Lexical	**86.06**	**86.09**	**86.06**	**86.06**
Syntactic	56.96	57.27	56.96	56.51
Sentiment	67.29	67.99	67.29	66.96
Semantic	77.37	78.46	77.37	77.15
Combined	84.91	84.95	84.91	84.91

Table 1: Experimental results on held-out test set

4.2 Other 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 words are common to both sexes which is not surprising for R&B artists. Additionally, the overall sentiment of each song was obtained using AFINN 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

Figure 2: Word cloud of songs by male music artists

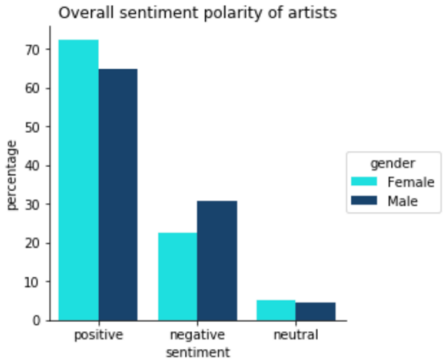

Figure 3: Overall polarity of songs by gender

5 Conclusion and Future Work

In this paper models were developed and tested to classify the gender of artists based on their lyrics. Data was obtained for R&B music artists from an online music database and the gender of artists manually labelled. The models were developed by exploiting four feature groups (lexical, semantic, syntactic and sentiment) and a combined set of all features. From the groups, the model trained using lexical information yields the best results proving to be the most important feature set in predicting the gender of an artist. Also, as suggested by other works, selecting Naïve Bayes as the classification algorithm for text classification tasks is a good choice.

In future work, it will be interesting to calculate the proportion and contribution of each feature in the model. Also, future research will explore if the same conclusion can be drawn when datasets with a mix of artists across various genres are used in training the models. Perhaps, a different feature group will give better results.

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.

Thorsten Brants. Natural Language Processing in Information Retrieval. In *Proceedings of the 14th Meeting of computational Linguistics in the Netherlands*, 2003.

Juhan Nam, Jorge Herrera, Malcom Slaney, Julius Smith, "Learning sparse feature representations for music annotation and retrieval," presented at the ISMIR, 2012.

Sameerchand Pudaruth, Sandiana Amourdon, and Joey Anseline. 2014. Automated generation of song lyrics using CFGs. In *Contemporary Computing (IC3), 2014 Seventh International Conference on.* IEEE, 613-616.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1-47.

Chris Hall, Richard Liptop, Martin Sliwinski, Micheal Katz, Charles Derby, Joe Vergheses. Cognitive activities delay onset of memory decline in persons who develop dementia. Neurology 2009; 73:356-61.

Alexander Sboev, Tatiana Litvinova, Dmitry Gudovskikh, Roman Rybka, Ivan Moloshnikov: Machine learning models of text categorization by author gender using topic-independent features, Procedia Comput. Sci. 101, 135-142 (2016)

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*. AAAI Press, March.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffery Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems,* pages 3111-3119.

Yitan Li, Linli Xu, Fei Tian, Liang Jiang, Xiaowei Zhong and Enhong Chen. Word embedding revisited: A new representation learning and explicit matrix factorization perspective. In *IJCAI* 2015.

Shaosheng Cao and Wei Lu 2017. Improving word embeddings with convolutional feature learning and subword information. In *AAAI*.

Finn Arup Nelson. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'making Sense of Microposts': Big things come in small packages.*

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining,* KDD'04, pages 168-177, New York, NY. Association for Computing Machinery.

Karthik Ramasubramanian & Abhishek Singh. "Machine Learning Using R", Apress, Berkeley, CA, 2017.

Dalibor Buzic and Jasminka Bobsa. Lyrics classification using Naïve Bayes. In 2018 *41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1011-1015.

Jason Rennie, Lawrence Shih, Jaime Teevan and David Karger. 2003. Tackling the poor assumptions of Naïve Bayes text classifiers. *Proceedings of ICLM*.

Aston Killin. 2018. The origins of music: Evidence, theory and prospects. *Music and Science* 1.

Steven Bird, Ewav Klein and Edward Loper. NLTK Book. O'Reilly Media, Sebastopol, CA, 2009.