Automatic Lyrics-based Music Genre Classification in a Multilingual Setting

Sam Howard[1] **, Carlos N. Silla Jr.**[1] **, Colin G. Johnson**[1]

[1]School of Computing
University of Kent
Canterbury, United Kingdom

{sjh61,cns2,C.G.Johnson}@kent.ac.uk

Abstract. *A large amount of research has been undertaken with regard to the classification of lyrics into genres, but most of this work has featured solely English lyrics. This study investigates the implications of classifying a multilingual database and the effectiveness of a number of techniques and algorithms for doing so. Part of this involves the creation of a high-quality dataset for use in this research. This paper finds that there are significant challenges in preprocessing multilingual text, and that traditional techniques like stemming and stop words may actually do more harm than good in such circumstances. It also finds that classes with strong language bias may be more likely to perform better than those with multiple languages.*

1. Introduction

Due to the digital music explosion of the late 1990's and early 2000's a new field has emerged. Music Information Retrieval [Downie, 2003] covers a broad range of topics, including topics as diverse as the human perceptions of music and intellectual property rights, but it is perhaps best known for research into the classification and clustering of music.

A great number of studies have been performed with regard to the classification of audio features extracted from songs into genres, largely using Machine Learning algorithms [Tzanetakis and Cook, 2002]. A number of studies have also looked into the potential for lyric features to assist or replace audio features [Neumayer and Rauber, 2007, Mayer et al., 2008a, Mayer et al., 2008b].

Most studies have focused on datasets with only one language represented, which is unrealistic in the real world. This paper aims to extend these studies by examining the effectiveness of a number of machine learning techniques and algorithms with regard to the classification of songs written in Spanish and Portuguese using solely lyric features.

2. Multilingual Dataset Creation

One problem with music classification is the appropriate classification of songs into genres. While no perfect solution is possible [Lippens et al., 2004], the Latin Music Database (LMD) [Silla Jr. et al., 2008], an existing dataset consisting solely of audio features, goes some way towards solving this by using experts to classify the songs by how they would be danced to.

Due to this, and it's variation in language, it provides a good basis for a multilingual lyrics database. The LMD contains over 3000 songs in 10 different music genres.

The lyrics are predominantly written in Spanish or Portuguese, but also contain a small number of English and Spanglish (Mixed English-Spanish) songs.

Using the LMD as a starting point, the next step was how to obtain the lyrics for the database. Although all the meta-data (performing artist, song name) is available from the LMD, current automated lyrics fetchers cannot handle the unusual (non-english/non-mainstream) songs. For this reason, we manually fetched the lyrics to create the Latin Music Lyrics database (LMLyD) which contains lyrics for 500 songs - around a sixth of the songs in the LMD. All ten genres are represented equally, with 50 songs in each. Since the lyrics have been inputted by a human a reasonable quality can be assured. The current version of the LMLyD contains songs in either Portuguese or Spanish.

3. Automatic Text Classification and Machine Learning Concepts

Once the lyrics database has been created and it is available, it is possible to deal with lyrics by using text classification techniques [Sebastiani, 2002]. In order to put lyrics through a classification algorithm they must first be processed into a suitable format.

In this work we have used the 'Bag of Words' approach to pre-process the lyrics. This approach consists of keeping track of how many instances of each word present in the data is used. This results in a set of Word Vectors which can be used to train and test the machine learning algorithms. All words are considered with no regard to their position in the text.

For the purpose of automatic multilingual lyrics classification, we have selected the following machine learning algorithms: Naive Bayes (NB), SMO and J48, all of which are available in the Weka framework [Witten and Frank, 2005].

4. Results and Analysis

In this section we first present the computational results by the individual classifiers with the respective analysis of their results and then we present further analysis on the language and pre-processing impacts. All the experiments reported in this section used a stratified ten-fold cross-validation procedure [Witten and Frank, 2005].

The results from testing with Naive Bayes (NB) are shown in the second column of Table 1. Overall NB performed well, correctly identifying the genre of 60.3% of the lyrics tested. However, there was an unexpected variation in performance amongst the genres. Particularly of note are Tango, which was correctly classified 94.1% of the time, and Bolero, which was correctly classified only 9.8% of the time.

Bolero's poor performance can be explained by the fact that while all of the other genres contain songs predominantly written in a single language, Bolero is almost equally split between songs written in Spanish and songs written in Portuguese. Because NB treats each variable independently, it has effectively been tuned to pick up on lyrics written in *both* Spanish and Portuguese, rather than *either*.

The results from testing SMO are shown are shown in the third column of Table 1. SMO performed similarly to NB, correctly classifying 57.4% of cases. While on average

Classifier	Axe	Bachata	Bolero	Forro	Gaucha	Merengue	Pagode	Salsa	Sertaneja	Tango
NB	66.0%	68.0%	09.8%	41.2%	62.7%	68.6%	47.1%	72.5%	72.5%	94.1%
SMO	68.0%	58.8%	39.2%	45.1%	51.0%	72.5%	37.3%	62.7%	60.8%	78.4%
J48	56.0%	39.2%	11.8%	33.3%	37.3%	39.2%	35.3%	45.1%	19.6%	41.2%

Table 1: Results for the different classifiers

Language	NB	SMO	J48	Average
Spanish	75.0%	68.0%	40.0%	61.0%
Portuguese	58.0%	51.0%	37.0%	48.6%
Both	10.0%	39.0%	12.0%	20.3%

Table 2: Language impact in classification.

its performance is slightly below Bayes, songs from the Bolero genre are classified correctly four times more often. While Bayes is looking for all of the features found during training, only half of which can be found due to the bilingual nature of the genre, SMO depends only on which side of a hyperplane the sample falls. While there will be two distinct clusters of points representing Bolero songs within the N-dimensional space, these should both fall on the same side of the hyperplane, meaning that both languages within the genre can be classified more easily.

The results from testing J48 are shown in the fourth column of Table 1. J48 performed poorly compared to the other algorithms, classifying only 35.8% of lyrics correctly. As with NB, Bolero songs are especially poorly classified. This is due to the fact that the algorithm always seeks to keep single classes together by finding the features which separate classes best. Since Bolero is split across languages, it contains two mutually exclusive sets of features. Since none of these features can apply to a majority of Bolero songs, they are not used to divide the dataset, and the resultant classification is very poor. In fact, the poor performance as a whole may be due to the algorithm's reluctance when it comes to splitting Bolero, thus removing any language advantage.

4.1. Influence of Language

Upon examining the lyrics music genre classification results it seems that language plays a large role in classification. As seen in Table 2, genres consisting solely of Spanish songs have a notable advantage over those consisting of Portuguese songs, and mixed language genres tend to perform worst of all.

The reason for this is relatively simple – while there are five genres consisting solely of Portuguese songs, there are only four consisting where the same is true of Spanish. Since the number of classes to choose between after selecting only those of the same language is lower, Spanish genres are more likely to be classified correctly.

Both NB and J48 struggled with Bolero songs due to the equal split in language present within the genre. A potential solution for this is to break the genre up into two classes: SpanishBolero and PortugueseBolero. Doing so should prevent Naive Bayes for looking for the features of both languages within Bolero songs, and should allow J48 to split the tree along language lines.

4.2. Influence of Stop Words Removal

An interesting question one might ask is what is the impact of using preprocessing techniques for lyrics classification. While on first glance it may seem unlikely that they will do anything but good, some features of lyrics may cause this not to be true.

Unlike most text classified using machine learning algorithms, Lyrics contain a great deal of artistic expression. In general only the content of the words is useful for classification, but the style of language used in lyrics may also be varied amongst genres.

Stemming may remove a lot of these features - for example, a livelier genre may be more likely to use the present tense, whereas a slower genre might use the past tense. While one has 'Dancing' and the other 'Danced', only 'Dance' will be used.

Removing stop words could also adversely affect the results. While it is unlikely

	NB	SMO	J48
With Stop Words	58.9%	56.6%	31.8%
Without Stop Words	60.3%	57.4%	35.8%

Table 3: Results for each algorithm with and without stop words removal.

there are particular trends within genres for tracks of the same language to use similar numbers of stop words, the fact that there are multiple languages means that the stop words present will give a strong bias to classify a song into a genre with many tracks of the same language. Removing stop words might improve the classification due to content, but reduce the classification due to language.

In this work we limit our analysis to the impact of stop words removal, due to technical issues with the Spanish stemmer. As seen in Table 3, the use of stop words actually leads to poorer classification performance than not using stop words at all. The difference is small for Naive Bayes and SMO but greater for J48.

5. Conclusions

In this paper we have investigated the use of different classifiers for lyrics-based music genre classification. A number of experiments were run on a novel lyrics database, the results of which showed great potential for a reasonably high level of accuracy. The results also presented a number of issues, which can be interpreted to identify a number of areas where multilingual classification should be treated differently to single-language classification. We have also verified that the use of a common text classification technique such as stop word removal is harmful to lyrics classification performance.

References

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, 37(1):295–340.

Lippens, S., Martens, J., Leman, M., Baets, B., Meyer, H., and Tzanetakis, G. (2004). A comparison of human and automatic musical genre classification. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 4, pages 233–236.

Mayer, R., Neumayer, R., and Rauber, A. (2008a). Combination of audio and lyrics features for genre classification in digital audio collections. In *Proc. of the ACM 16th Int. Conf. on Multimedia*, pages 159–168.

Mayer, R., Neumayer, R., and Rauber, A. (2008b). Rhyme and style features for musical genre classification by song lyrics. In *In Proc. of the 9th Int. Conf. on Music Information Retrieval*.

Neumayer, R. and Rauber, A. (2007). Integration of text and audio features for genre classification in music information retrieval. In *Proc. of the 29th European conference on Information Retrieval*, pages 724–727.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47.

Silla Jr., C. N., Koerich, A. L., and Kaestner, C. A. A. (2008). The latin music database. In *Proc. of the 9th Int. Conf. on Music Information Retrieval*, pages 451–456.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.