Table of Contents

1. The abstract

In this paper we present a study on emotional music classification using information from lyrics and music. We show that the methods used are able to classify the lyrics significantly better than random. To enhance the results from the lyrics categorization we collected information from the music as well. Features like *tempo, key (major/minor)* and a few more were used with the hope to improve the classification.
The musical factors as well as the lyrics factor have been evaluated separately, and then later combined.

2. The introduction

During the past decade the availability of music on the internet has exploded. This introduced many new research areas related to music and classification of the music. For example a common feature on many websites is to be able to enter a search word, like a name or a category of music and in return get a set of songs that in some way have a connection to the search word.

The goal for this project was to use the knowledge we had gained from a previous course taken in language technology[R1] and apply it on some research previously done in the language technology field, and with the many lyrics and music sharing sites available on the internet today it was compelling to try and make a project out of this. So that's when the lyrics and music classification came to mind.

The idea was to first start by annotating the lyrics of a set of songs and experimenting with different methods to get a satisfying result, and then finally adding features extracted from the music to see if music and lyrics combined would contribute to a more accurate classification.

3. Methods

To get started and to get some insight on what has been done earlier on the subject we did some research and with the help of our project supervisor we found two articles that we could use as reference for our project. [A1][A2]
Due to the restricted amount of time there was no strictly set goal for the project, but there was the prospect that both lyrics and music could be analyzed both separately and later put together to see if that would enhance the result.

The first step was to manually annotate a set of songs. To get a big enough corpus 200 songs were tagged with appropriate emotions. This was done by retrieving the lyrics for a song, and then each verse or chorus was assigned the emotion perceived by reading through that chunk of text. It was also

possible for a chunk to get more than one emotion assigned to it. To speed up this step we did a hundred songs each that later were put together to make up the corpus.

When the corpus was annotated a script was used to get a training set and some test sets to experiment on. For this part a short and effective python script was written to produce output files in the format needed to use LibShortText.[Lib1] LibShortText is a tool for Short-text Classification and Analysis, and provided us with the tools needed to experiment and analyze data in this project.

To get a generalized result we used cross validation for the test sets. And by numbering the 200 songs and using the last figure of their number, 0, 1, 2 and so on, we got ten different test sets to evaluate. Since the songs also was divided in sets of a hundred from the annotation part of the project we could also split them up and see if there was any difference in the result comparing the tagging of emotions from the two of us.

Using LibShortText and our test files we experimented with different options, for example with or without stopword removal, stemming/ no stemming and using unigram or bigram.

The next step in the project was to try and add some features from the music and see if that would contribute and give us a better result with the classification. After discussing different options it was decided to use the EchoNest[Echo1] API to extract information from the music. The EchoNest features used were tempo (in beats per minute, or BPM), mode (i.e., major or minor key), energy, loudness, danceability, and valence (i.e. mood of the song). The features energy, danceability and valence were proprietary attributes determined by EchoNest.

As such detailed categories as ‘remorseful’, ‘wistful’, etc. were considered too fine-grained to be properly classified by musical features, and there was considerable overlap between categories, we decided to group the features into three types: mood, energy, and love.

The information extracted was converted into libsvm-format and used as extra features. This was used for the combined lyrical and musical features.

It was also converted into ARFF format for Weka. We used Weka’s SimpleLogistic algorithm to classify the musical features and evaluate how well they did by themselves.

4. Results

4.1. Categories

The following categories were originally considered when tagging the lyrics:
sad, regretful, happy, wistful, angry, party, crazy, upbeat, love, soft, romantic, strong.

Some of the categories were removed, this because they were confused with others too often and the we got the best result when using only these categories: sad, regretful, happy, wistful, angry, crazy, upbeat, love.

Table 1. Cross validation results on text features

	crazy	love	angry	sad	wistful	upbeat	happy
crazy	28	27	10	8	21	48	15
love	1	213	24	34	74	40	18
angry	10	43	24	9	13	16	5
sad	3	47	21	23	41	16	2
wistful	2	143	10	34	117	58	22
upbeat	5	98	11	26	113	89	17
happy	0	101	2	5	43	27	47

Later, the categories were grouped into mood, energy and love, as follows (label names italicized):

MOOD	ENERGY	LOVE
happy:	*strong:*	*love:*
happy	party	love
upbeat	upbeat	romantic
sad:	*soft:*	*other*
sad	soft	
regretful	wistful	
wistful		
	other	
other:		
angry		
crazy		
strong		

4.2. Results lyrics

Listed below are the results retrieved by running LibShortText with the following command and options, ./text-train.py train -f -P 3 -F 0 -N 1 -L 2. The different parameters correspond to:

-P stands for preprocessor options, we used 3 which means no stopword removal, stemming and bigram.

-F is the feature representation where we used binary, 0, this is the default setting.

-N 1 means that instance-wise normalization will be done before training/test.
-L specifies what classification to be used, we got the best results when using L2-loss support vector classification.

10 fold cross-validation:
fold 0 acc 32.9113924051
fold 1 acc 33.4763948498
fold 2 acc 36.6666666667
fold 3 acc 26.2443438914
fold 4 acc 27.397260274
fold 5 acc 30.0546448087
fold 6 acc 17.6470588235
fold 7 acc 21.7105263158
fold 8 acc 33.152173913
fold 9 acc 31.1557788945

average: 29.0416240842

4.3. Results music features

0,000 happy
0,574 sad
0,174 wistful
0,000 regretful
0,556 upbeat
0,000 angry
0,333 crazy
0,277 love

Weighted Avg. 0,356

4.4. Results music and lyrics with grouping

	Mood	Energy	Love
Text	55%	41%	58%
Music	65%	59%	58%

4.5. Combined result music and lyrics

Mood: 56.6%
Energy: 42.6%
Love: 59.9%

5. Analysis & Discussion

5.1. Lyrics

At first when we used twelve different categories many of them got mixed up because they did not differ enough, for example romantic and love was hard to distinguish from one another when just looking at the text. By concatenating some of the categories that got mixed up a lot and having fewer categories the results improved.
To further improve the results for this part of the project we could have used more than just one person to tag the same lyrics, this would have contributed to a more generalized tagged corpus.

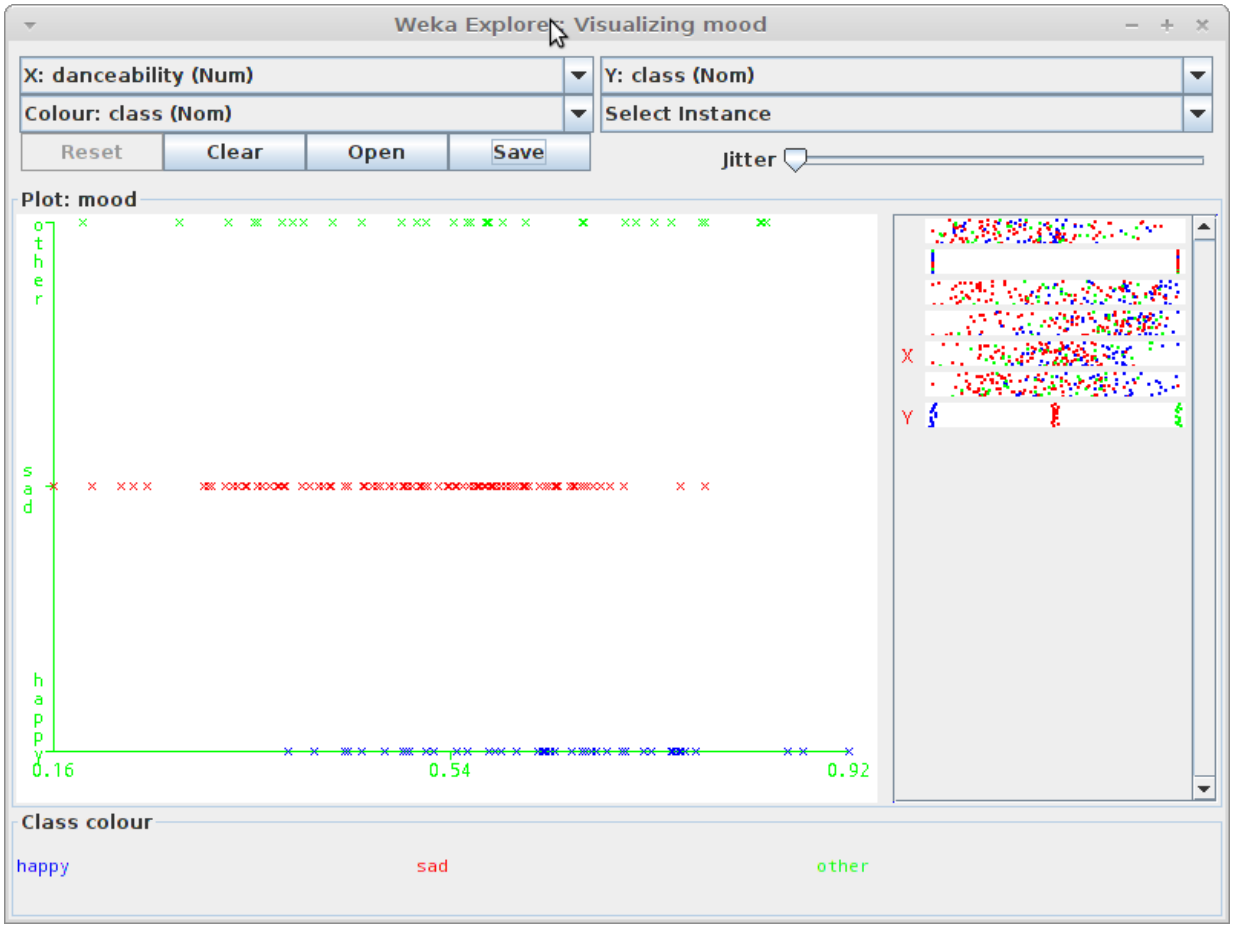
5.2. Music features

The results we got when just looking at the music varied a lot, as seen in section 4.3. For example sad and upbeat were very good whereas happy and angry were really bad. The reasons for this may be many, but we believe that tempo and the other features that we looked at are more distinguished in categories like upbeat which make it fairly easy to categorize them. But in categories like happy the music and its features can vary a lot and thereby making it hard to categorize them.

In addition, there was considerable overlap between categories, with categories such as ‘sad’ and ‘wistful’ and ‘happy’ and ‘upbeat’ being applied to the same lyrics.

5.3. Lyrics and Music with grouping

As can be seen from section 4.4, the results for both the lyrical and the musical features improved considerably after the grouping. The figure below shows how one of the musical features works: ‘sad’ songs have mostly low danceability, and ‘happy’ songs have mostly high danceability, with both types having a clear cutoff point. Songs labeled ‘other’ (neither happy nor sad) have danceability scores across the spectrum.



We also expected for the results to improve if we used both the lyrics and the music together. The results did improve from just using lyrics themselves, as seen in section 4.4, but only by about 1 or 2 percentage points in all cases, which may not be a statistically significant amount. Also, in the case of Mood and Energy, they performed significantly worse than the musical features by themselves, which tells us that libshorttext isn't using a good classifier for the musical features. This was the same whether we used LogisticRegression or SVM, so it isn't clear why this is so.

5.4. Comparison to State-of-the-art

Below are the average numbers from the paper we used as a reference for our project[A1].

AVERAGE:

Lyrics: 0.4766

Music: 0.3371

Lyrics&Music: 0.5439

If we use these numbers to compare with the average we got we find that we did slightly worse in the

lyrics part, but for the music part we actually did better.

6. Conclusions

Unfortunately we didn't get the final results that we were hoping for, being that the combined lyrics and musical features provided us with a better classification than the individual classification of lyrics and music. The main reason to why we didn't get the results we were hoping for was that we ran out of time, since it was only an eight week school project. But overall we are pleased with the results we got from the individual classifications, and it is our opinion that given a little more time we could have gotten better results for the combined classification.

7. References

[R1] Language Technology EDAN20, LTH.
http://kurser.lth.se/lot/?val=kurs&kurskod=EDAN20

[A1] Rada Mihalcea, University of North Texas. Carlo Strapparava, FBK-irst. ***Lyrics, Music, and Emotions***.
www.aclweb.org/anthology/D12-1054

[A2] Yunqing Xia, Center for Speech and language Tech. RIIT, Tsinghua University. Linlin Wang, State Key Lab of Intelligent Tech. and Sys. Dept. of CST, Tsinghua University. Kam-Fai Wong, Dept. of SE&EM, The Chinese University of Hong Kong. Mingxing Xu, Dept. of CST, Tsinghua University. ***Sentiment Vector Space Model for Lyric-based Song Sentiment Classification.***
www.aclweb.org/anthology/P/P08/P08-2034.pdf

[Lib1] LibShortText
http://www.csie.ntu.edu.tw/~cjlin/libshorttext/

[Echo1] EchoNest
http://www.echonest.com/