A survey on classifiers for sentiment classification of song lyrics

GROUP31

Akash Patel
921119-3256
19/11/92
akash@kth.se

Christoph Kaiser
920416-T336
16/04/92
ckai@kth.se

Lisa Schmitz
920606-T203
06/06/92
lschmitz@kth.se

Nikolaos Tatarakis
891016-T457
16/10/89
nta@kth.se

Abstract

This paper explores the potential accuracy of the analysis of song lyrics. Datasets with various labels were tested for their ability to categorize lyrics as emitting a certain mood. The focus lies on the comparison of different categorisations and classifiers. The identification of emotions in lyrics is a problem which has no satisfactory solution yet due to the ambiguous and subjective nature of lyrics.

1 Introduction

Music is perhaps considered the most impactful art medium in terms penetration and cultural relevance. Most people are familiar with the feeling of how a song can influence our mood to be everything between melancholy to euphoric. This control over people's feelings is something which can be used for many different purposes. For example music providers like Spotify offer playlists labeled with a certain mood and also play similar songs if requested. But industries are not the only area of application. Researchers see a use for it in edutainment and even psychological therapy [6]. Unfortunately, the task of predicting the correct associated mood is not an easy one due to the complexity of how emotion is transferred in songs. Obviously emotion is encoded both in the audio and the lyrics of a song [12]. This paper compares methods to identify emotion by solely analysing the text of song lyrics. In order to do this, different variants of a text analyser were tested. The modification includes using different categories of expressive emotions and classifiers as well as various sizes of the test and the trainingsset.

1.1 Contribution

The goal of our study was to explore the limitations and advantages of different classifiers which are trained with one single dataset. Since nearly every study which aims to categoriese lyrics according to emotions uses its own version of labeling and a specific set of lyrics, the results of the studies are not adequately comparable. This is why we applied different classifiers on the same dataset. We expected to gain a better understanding of suitable emotion categories and the efficiency of certain classifiers in combination with n-grams.

1.2 Outline

Since our work deals with different approaches of categorising and classifying song lyrics, previous work should be taken into count. The related work is therefore presented in Section 2. We based our text analyser on the results of these previous studies. Section 3 explains the method we used to realise and implement the analyser in detail. We used different variations of our text analyser, modifying the categorisation, the size of the trainingsset and the classifier. The results we were able to gather are described in Section **??**. Moreover, problems that came up during the research are mentioned in this section. The results are summarised in Section 4 and possible further research areas are touched upon briefly.

2 Related work

Our work is mainly based on the paper of Youngmoo E. Kim et al. [6]. It gives an overview of recent approaches of emotion recognition in lyrics. Most of the presented approaches are content-based and therefore relevant references for our work. Not only do they deal with different categorisations of mood but also treat variations of classifiers. This work provides a good insight into what has already been done and what worked well. Therefore it can be seen as the foundation we built our work on. We realised some of the presented methods and compared them to each other. To solve the task of finding appropriate emotion labels, research approaches by Yang et al. [11] and by Downie et al. [5] were taken into account. These lables are used by the lyric classifiers which have been developed as part of this work. Furthermore, the classifiers work according to the bag-of-words model which is explained by Li et al. [7].

3 Method

The basic model we used for our text analyser is the bag-of-words model [8]. The functionality is illustrated in the figure 1. In this model a database with labeled data such as lyrics are stored. The next step would be to pre-process the lyrics and create a dictionary with n-grams based on it. The database is divided into two parts: a training and a test set. Using this data, a classifier is trained with a training set and its accuracy can then be evaluated with the classification of the test set.

3.1 Implementation

The following section will discuss the individual parts of the text analyser in greater detail whilst providing some relevant scientific background to the classification methods used.

Database The database was one of the obstacles of this research. Due to copyright issues it was not possible to access an already available database and we had to built one on our own. Some earlier researches use `Allmusic.com` as a basis for the database [6], since it provides emotional labels for songs. For this reason a program has been written to extract lyrics along with their emotional labels.

Even though the pre-labeling of the songs were helpful, with 289 emotional labels there were simply too many categories to use. Therefore we had

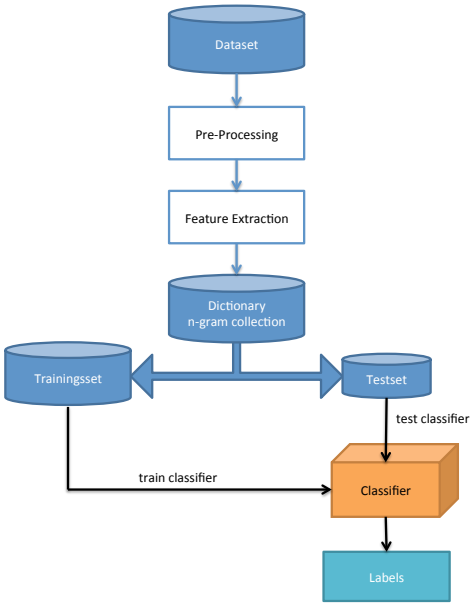

Figure 1: A description that makes browsing the paper easy and clearly describes what is in the picture. Make sure that the text in the figure is large enough to read and that the axes are labelled.

to find super-categories for existing labels. This means that a some emotional labels are closely related to each other and therefore can be grouped together to a more generalised emotion. For example the emotion *happy* could be used as a super-category for *cheerful* and *fun*.

Distinct ways of labeling have been investigated in previous research. Downie et al. suggest in [5] to use five clusters of emotions. Whereas Yang and Lee only suggests a binary distinction into positive and negative emotions [11]. Both variants were tested during our research.

Pre-processing The lyrics have been pre-processed the same way for each of the variants of the text analyser. Stop-words are filtered out using the Natural Language Toolkit and special characters are deleted. In a next step, every word of the lyric is leammtized which means that it is transferred into its basic form. For example "wait" and "waiting" would be interpreted as the same word after the lemmatizing has been done. The pre-processing is performed to avoid misclassifications.

Feature extraction The feature extraction is conducted for all lyrics. The procedure consists of transferring the lyrics into n-gram representation. We

considered only 1-, 2- and 3-grams as reasonable parameters which is why only these have been tested in our study. One reason for this was that 4-grams were considered as too specfifc and would not lead to accurate results.

Classification The Natural Language Toolkit and the SciKit-Learn library provide the most commonly used text classifiers. In the following section the tested classifiers will be introduced.

- *Naive Bayes* Let $\mathcal{C} = (c_1, \ldots, c_n)$ be the categories and n the number of categories. Given a document $\mathcal{D}$ and its word list $\mathcal{W} = (w_1, \ldots, w_m)$ with m being the number of words of a certain document, the Naive Bayes classifier determines the category of a document as follows:

  $$c^*_{NB} = argmax_{c_j \in \mathcal{C}} P(c_j) \prod_{i=1}^{d} P(w_i|c_j)$$

  with $P(c_j)$ being the *a priori* probability of a class c_j and $P(w_i|c_j)$ being the conditional probability of the word w_i given class c_j.

  Naive Bayes will only work well for large enough datasets [2].

- *Multinomial Naive Bayes* Multinomial Naive Bayes extends the Naive Bayes algorithm to handle multinomial distributed data [9]. In text analysis this is particularly useful where events represent the occurrence of a word in a document.

  A feature vector $w = (w_1, \ldots, w_n)$ is given where w_i counting the number of times a word i was observed in a particular instance. The likelihood of observing a histogram $\mathbf{w}$ is then given by,

  $$P(\mathbf{w}|c_j) = \frac{(\sum_i w_i)!}{\prod_i w_i!} \prod_i p_{ji}^{w_i}$$

  Here p_{ki} is part of a multinomial distribution p_k and the probability that a word i is observed.

- *Bernoulli Naive Bayes* In contrast to the Multinomial NB the Bernoulli NB uses binary term occurrence rather than term frequencies which allows it to penalize non-occuring terms. Although there may be multiple features, each one is assumed to be a binary valued variable. Thus samples are represented as binary feature vectors. The likelihood of a feature vector given a category is given by,

$$P(\mathbf{w}|c_j) = \prod_i p_{ji}^{w_i}(1 - p_{ji})^{1-w_i}$$

The model is particularly useful for shorter texts where the lack of certain features can tell you more than solely analyzing occurring features [3].

- *Logistic Regression* Logistic regression is a linear model for classification where the probabilities describing the possible outcomes of a single event are modeled using a logistic function. This model is then used to classify new input data [2].

- *Support Vector Machine* If input data are overlap and have similar features, SVMs can be used. They essentially perform non-linear separations by transforming input data to higher dimensions with so called Kernels. This is then combined with a Lagrange constraint to find hyperplanes which separates classes with the largest distance [4].

- *Linear Support Vector Machine* The main difference with linear SVM compared to kernel SVMs is that the input data is linearly separable and thus a non-linear kernel to transform data to an higher dimension is not necessary. They are particularly useful for data which have a lot of features and thus easier and faster to separate with a hyperplane [4].

- *Stochastic gradient descent* This classifier combines multiple binary classifiers to separate features in input data. These classifiers are then updated using a version of gradient descent to update the parameters and minimize the error. In contrast to normal gradient descent the stochastic version updates the classification parameters for every training example it encounters with respect to the gradient of the error. In the batch gradient descent one has to run through the entire training set before calculating the gradient and updating the parameters, which is usually slower [1].

3.2 Experimental setup

In order to compare the performance of the different classifier variants we changed various parameters. The effect of different amounts of emotional categories have been tested, as well as variations of the n-grams and classifiers.

3.3 Experimental results

In the following abstract the accuracy results of different classifiers will be presented and interpreted.

Pos/neg categories As mentioned earlier, a classifier was implemented which lables given lyrics as either *positive* or *negative*. Overall the research shows an accuracy rate of about 67% for the best performing classifier Linear SVC as seen in 2. Furthermore, from this figure we see that the best classifiers, Linear SVC and Logistic regression are in fact linear. This can be explained by having highly individualized data with large feature space. Thus more complicated classifiers like kernel SVMs are not necessary. However, the NB and Multinomial NB do not perform well at all. This might be due to *sparseness*, meaning the training data is not large enough to represent the frequencies of certain n-gram phrases with $n > 1$.

These results imply that the use of only two categories - positive and negative - seem to have let to an overgeneralisation of emotions. Even the categories itself which were provided by `Allmusic.com` could not clearly be mapped to one of these two categories. To illustrate the difficulty we discovered, a small study was conducted. Using a questionnaire people were asked to evaluate the categories from `Allmusic.com` as evoking either a negative or positive feeling and to do the same with some randomly picked song lyrics. The results are shown in figure **??** It can be seen that there is a significant discrepancy of opinion across participants. This emphasises the personal perception of emotion. Due to this, it cannot be expected to gain a higher accuracy by a binary classifier in emotion categorisation.

Five clusters A second model of categorisation used in this research has been proposed by Downie et al. [5]. They used five cluster of emotion to categorise lyrics as shown in table 1. The results for this categorisation can also be seen in figure 3. The Multinomial NB provides the highest accuracy with about 32% . Even though it is significantly better than guessing there are many cases of misclassification. A reason for this is that the size of the dataset is only half the size than the one for the binary pos/neg categorisation. This could not be avoided because the five cluster model introduced in [5] takes only a small subset of emotion labels provided by `Allmusic.com` into account.

Variations in n-grams Another parameter we modified was the assignment of n for the n-grams. The effect on the accuracy of the classifiers is illustrated by the figure 2 and figure 3 and . In the pos/neg category case

Clusters	**Mood Adjectives**
Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 1: Clusters of mood adjectives used in the MIREX Audio Mood Classification task [5].

as seen we observe a strong decline in accuracy for Naive Bayes (NB) and Multinomial NB which might be due to the fact that as we increase the number of n-grams, phrases will be less independent since they contain same words. Along with convergence at 2-gram and high run time after 3-gram the evaluation was halted.

In the cluster case the reason for the accuracy decline for all classifiers with increasing n-grams might be due to the fact that phrases can contain words which are common to several clusters. Phrases would therefore be unreliably classified to one of the clusters. Furthermore the same issues with the NB classifier in the two category version of the text analyser can be observed in this case.

Variations in the training and test set In order to reflect a more precise picture of the individual classifiers accuracy, the whole database was randomised before partioning into a training and test set. This was performed multiple times and the results were consequently averaged. The accuracies presented in figure 2 are a result of this method.

Variations in classifiers We tested the different classifiers described in 3.1. The performance of different classifiers for the two mood variant are shown in figure 2 and for the cluster variant are presented in figure 3. As it can be seen in the figures, the ... classifier yields the best average results for both categorisation variants. The reason why the linear classifier is working well is that the tested data is highly individual. There is a significant variance between song lyrics and this is why a linear distinction is the best choice for the data in the present case.

Drawbacks We were aware of the difficulty of the task of designing a sufficient lyrics classifier. In recent years a lot of research has been conducted

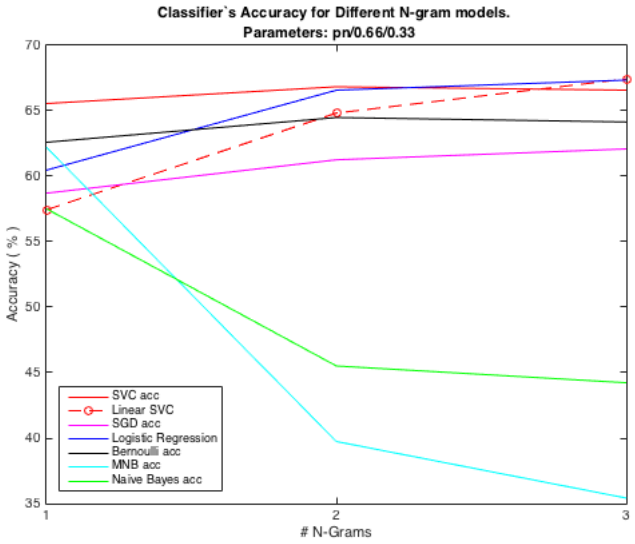

Figure 2: Accuracy of the classifiers with different n-grams for pos/neg classification.

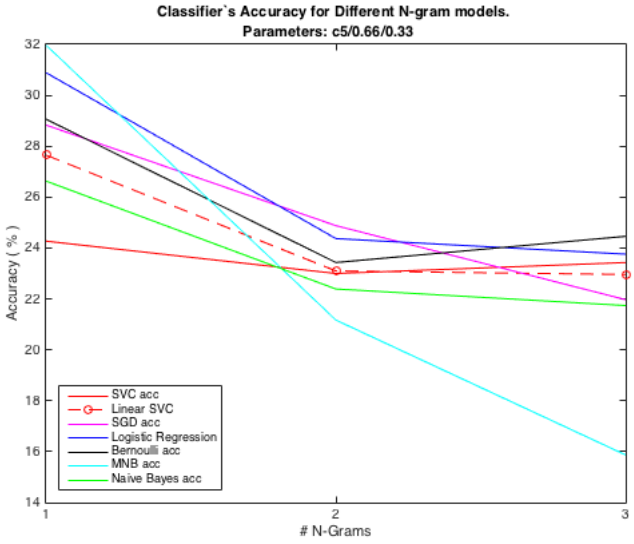

Figure 3: Accuracy of the classifiers with different n-grams for the five cluster classification.

in this area but there is still a lack of a reliable lyrics analyser. But during our research we were able to gain a better insight into the nature of this difficulty. Feelings and emotions are something very personal, and the perception of the emotion which is transferred by a certain lyric highly depends on the person who is perceiving said lyric. It is not unusual that two people strongly disagree on the general mood of a lyrical text [10]. This is why it is of supreme importance to select a representative model of categories.

4 Summary and Conclusions

From the results we can conclude that the presented text analyser can successfully classify lyrics to a positive or negative mood with a satisfactory degree. In the clustered emotions case we obtain a success rate higher than random guessing for 1-gram and 2-gram. The most accurate classifiers are Linear SVM, kernel SVM and Logistic Regression.

The greatest limitation to this project was the subjectivity of how song lyrics are perceived. Some people may identify a song as cheerful while others find it brooding. This definitely affects the labeling in the training set since we consider the labels an accurate reflection of the real world.

5 Contributions

We the members of project group31 unanimously declare that we have all equally contributed toward the completion of this project.

References

[1] Cs229 lecture notes.

[2] Generative and discriminative classifiers: Naive bayes and logistic regression.

[3] Text classification using naive bayes.

[4] support vector machine (and statistical learning theory)tutorial.

[5] XHJS Downie, Cyril Laurier, and MBAF Ehmann. The 2007 mirex audio mood classification task: Lessons learned. In *ISMIR 2008: Proceedings of the 9th International Conference of Music Information Retrieval*, page 462. Lulu. com, 2008.

[6] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010.

[7] Yong H Li and Anil K. Jain. Classification of text documents. *The Computer Journal*, 41(8):537–546, 1998.

[8] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[9] Jason D Rennie, Lawrence Shih, Jaime Teevan, David R Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 3, pages 616–623. Washington DC), 2003.

[10] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[11] Dan Yang and Won-Sook Lee. Music emotion identification from lyrics. In *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*, pages 624–629. IEEE, 2009.

[12] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H Chen. A regression approach to music emotion recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):448–457, 2008.