

Masterstudium: Information Engineering and Knowledge Management

Diplomarbeitspräsentationen der Fakultät für Informatik





Jerome Penaranda

Technische Universität Wien Institut für Softwaretechnik und interaktive Systeme Arbeitsbereich: Information und Software Engineering Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber und Mag. Robert Neumayer

## Abstract

The organization of large quantities of music is a common problem in an era, in which there is an increase in the spread of digital music. A well-tried means is the classification in appropriate musical genres. We propose the use of text categorization techniques to classify music in the form of song lyrics, which are present in the internet. In addition, different features, both content-based and structure-based features, are extracted from the song lyrics. With these features a classifier is trained, which then assigns the appropriate music genre to the respective lyrics. Support Vector Machines and Naive Bayes Classifiers are primarily used in such classifications. We present experiments comprising the evaluation of the classification process and the combination of different features to increase the classification accuracy. On the basis of these experiments, we study how many lyrics are necessary to get good results, which overall performance we can expect for classification and which feature combinations are suitable for the classification of song lyrics.

### **Basic Concept**

- 1. song lyrics are extracted from the internet -> arranged into a taxonomy of *n* musical genres
- 2. a musical genre is represented by *m* song lyrics with artists of this genre
- 3. features are extracted from the song lyrics
- 4. text classifiers (SVM, Naive Bayes) are trained

# **Data Sets**

# Sing365-Corpus

1281 lyrics, 6 musical genres

genre	number
$\operatorname{Hip-Hop}/\operatorname{Rap}$	236
Rock	217
Pop	204

- 5. new song lyrics are classified into one of the *n* musical genres

#### **Features**

- bag-of-Words features: each feature corresponds to a single word found in the lyric
- part-of-speech features: are extracted from the lyrics by assigning part-of-speech tags (noun, verb, preposition, etc.) to the individual words
- language feature: represents the language of the song lyrics
- rhyme features: represent various rhyme patterns identified by analyzing the words at the end of each line

Nr	Feature Type	POS-Labels	Feature Set
1.	POS Features	NN, NNP, NNS	$\#  ext{ of nouns}$
2.		VVB, VVD, VVG,	#  of verbs
		VVI, VVN, VVNJ,	
		VVGJ,VVGN,VVZ	
3.		PNR	$\#$ of rel_pronouns
4.		II	#  of prepositions
5.		RR, RRR, RRT	#  of adverbs
6.		A, AN, THE	$\#  ext{ of articles}$
7.		PN, PND, PNG	$\#  ext{ of pronouns}$
8.		VM, VBB, VBD,	$\#  ext{ of modals}$
		VBG, VBI, VBN,	
		VBZ, VDB, VDD,	
		VDG, VDI, VDN,	
		VDZ, VHB, VHD,	
		VHG, VHI, VHZ	
9.		JJ, JJR, JJT	#  of adjectives

$\mathbf{Nr}$	Feature Type	Feature Set
1.	Rhyme Features	#  of AA
2.		#  of AB
3.		#  of AABB
4.		#  of ABAB
5.		#  of ABBA
6.		#  of words
7.		# of number ofwords
8.		#  of wordpool
9.		$\#  ext{ of chars}$

- lyrics from the website: www.sing365.com
- use of global genres
- musical genres are assigned to the lyrics according to the mappings found in www.allmusic.com

гор	Z04
Reggae	192
R&B	229
Country	203
Total	1281

genre	number		
Acid Punk	19	••••	104
Alternative	478	Hardcore	184
Ambient	15	Hip-Hop	613
Avantgarde	110	Indie	334
Blues	23	Industrial	22
BritPop	56	Metal	572
Christian Rock	38	New Metal	98
Classic	513	Pop	860
Country	134	Post Punk	26
Dance	13	Punk Rock	1390
Dance Hall	10	R&B	254
Electronic	143	Reggae	49
Emo	254	Rock	715
Experimental	10	Ska	37
Folk	46	Slow Rock	501
	40 41	Soundtrack	27
Garage Goth Metal		Speech	53
	46	Trip-Hop	52
Grunge	120	World	4
Hard Rock	24	Total	7884
		-	

genre	number	
Avantgarde	120	
Blues	23	
Classic	513	
Country	180	
Electronic	286	

# Parallel-Corpus

- 7884 lyrics, 37 musical genres
- corpus was built according to the mp3 files of a private music collection
- lyrics from the websites : lyrc.com.ar, sing365.com and oldielyrics.com
- use of musical genres and styles
- genres are assigned according to the genres also assigned to the respective mp3 file

#### Parallel-Corpus (1.level genres)

- musical genres and styles of the Parallel-

Corpus are merged into global genres

Hip-Hop	613
Pop	860
Reggae	59
Rock	4892
R&B	254

## **Rhyme Detection**

Rhyme features are extracted from English song lyrics by identifying rhyme patterns using phonemes. In case of non-English lyrics rhyme detection is performed by comparing the final syllables of words at the end of each line. The rules of extraction of the individual word endings are as follows:

1. If the last character is a consonant, then all characters are scanned to the left, until it gets to a vowel. The consonant(s) then make up the word ending with the vowel. If, however, the character that is before a vowel is likewise a vowel, then this character is also added to the word ending. Examples of such are: concentrat-ion, measurem-ent, light-ing, etc.

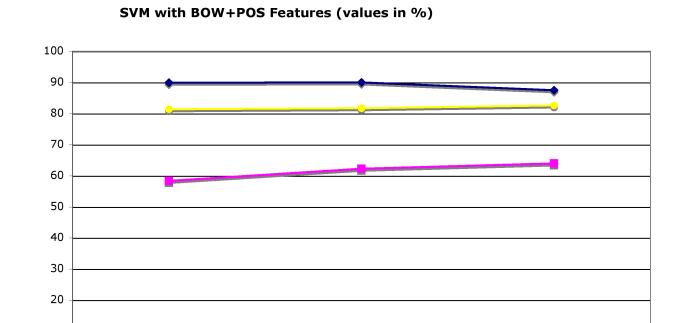
2. If the last character is a vowel, then all characters will be scanned to the left, until it gets to a consonant. This consonant is not added to the word ending. Examples of such are: dram-a, cinem-a, etc.

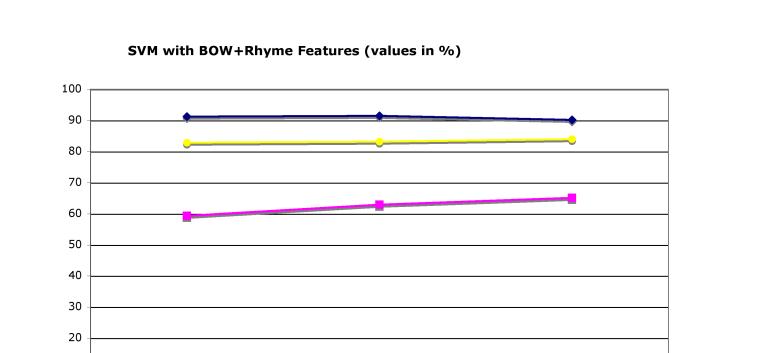
### **Process of the experiments**

- 1. features are extracted from song lyrics
- 2. different features are combined into various feature sets
- 3. chi-square value is computed for each feature -> features are arranged according to size
- 4. feature vector is created using the first *n* features with the highest chi-square value, notation Cn describes the strategy of selecting *n* features (n= 600, 800, 1000)
- 5. classifier is trained (SVM, Naiver Bayes)
- 6. 10-fold cross validation is conducted for evaluation

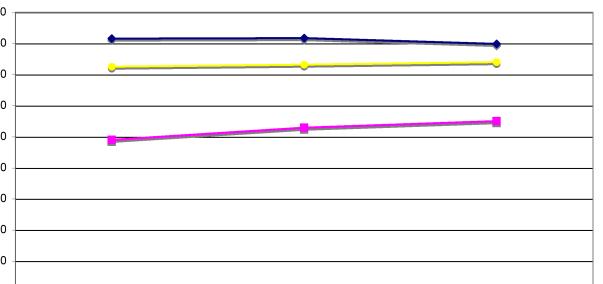
# **Evaluation**

- accuracy of up to 91% using Support Vector Machines and the Sing365-Corpus
- up to 65% accur. using SVM and the Parallel-Corpus
- improvement of results of the Parallel-Corpus by modifying the Corpus -> Parallel-Corpus



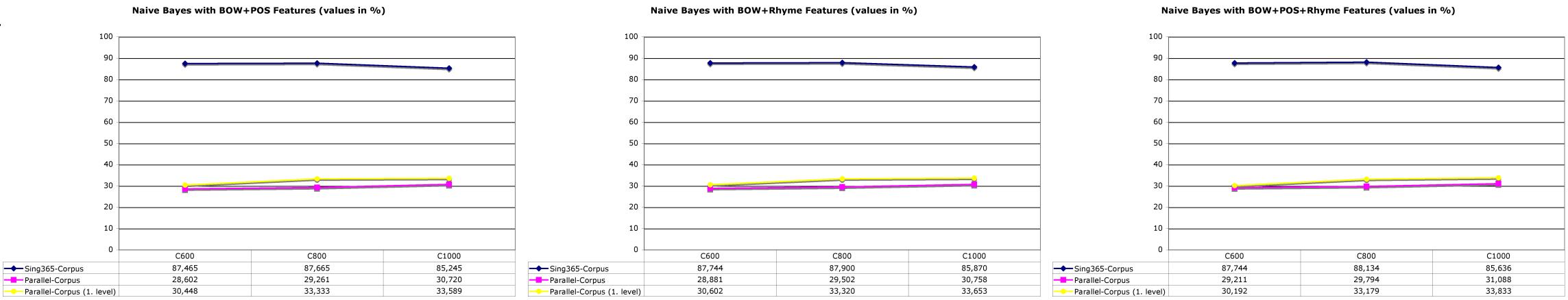






#### (1.level): 83%

- C600 C800 C1000 C600 C800 C1000 C600 C800 C1000 90,007 89,929 87,431 91,178 91,432 91,556 91,710 90,163 89,805 ----Sing365-Corpus ------Sing365-Corpus ------Sing365-Corpus 58,358 62,265 63,952 59,310 62,874 65,119 59,005 62,899 65,030 SVM consistently outperforms Naive Bayes Parallel-Corpus ----- Parallel-Corpus ----- Parallel-Corpus 81,371 81,730 82,602 82,794 83,153 83,961 82,500 83,128 83,974 ---- Parallel-Corpus (1. level) Parallel-Corpus (1. level) Parallel-Corpus (1. level) Parallel-Corpus (1. level) - Parallel-Corpus (1. level) -Parallel-Corpus (1. level)
- stopword removal improves performance
- worse results using Stemming
- genres with high significance (e.g. Classic containing non-english lyrics) are classified very accurately
- better results achieved with Rhyme features than with POS features
- the use of Phonemes is more suitable for rhyme detection than other approaches
- better results of Parallel-Corpora using SVM and higher feature dimensions



Parallel-Corpus (1. level) Parallel-Corpus (1. level) Parallel-Corpus (1. level)