Bundeli Folk-Song Genre Classification with kNN and SVM

Ayushi Pandey
Dept. of Computational Linguistics
The EFL University
ayuship.09@gmail.com

Indranil Dutta
Dept. of Computational Linguistics
The EFL University
indranil@efluniversity.ac.in

Abstract

While large data dependent techniques have made advances in between-genre classification, the identification of subtypes within a genre has largely been overlooked. In this paper, we approach automatic classification of within-genre Bundeli folk music into its subgenres; Gaari, Rai and Phag. Bundeli, which is a dominant dialect spoken in a large belt of Uttar Pradesh and Madhya Pradesh has a rich resource of folk songs and an attendant folk tradition. First, we successfully demonstrate that a set of common stopwords in Bundeli can be used to perform broad genre classification between standard Bundeli text (newspaper corpus) and lyrics. We then establish the problem of structural and lexical similarity in within-genre classification using n-grams. Finally, we classify the lyrics data into the three genres using popular machine-learning classifiers: Support Vector Machine (SVM) and kNN classifiers achieving 91.3% and 85% and accuracy respectively. We also use a Naïve Bayes classifier which returns an accuracy of 75%. Our results underscore the need to extend popular classification techniques to sparse and small corpora, so as to perform hitherto neglected within genre classification and also exhibit that well known classifiers can also be employed in classifying 'small' data.

1 Introduction

Bundeli is spoken in regions of Madhya Pradesh and Uttar Pradesh, in a region known as Bundelkhand, which encompasses several administrative districts in India. While the 2001 census identifies 3,070,000 Bundeli speakers, Ethnologue estimates 20,000,000 speakers[1]. Inspite of the large population, Bundeli, often considered a dialect of Western Hindi, would be considered an under-resourced language because of the lack of textual and literary resources that are available. Bundelkhand, however, is home to a rich tradition of lyrical styles and genres that are both performative and poetic. In addition to the major genres; Gaaris, Rais and Phags, several other genres can be found in this region, including Sora, Hori, and Limtera. With the exception of several Phags[2], Bundeli is a resource poor language, in that there are no available sources of textual material. Using a bag-of-words technique on the lyrics corpus, and Term Frequency-Inverse Document Frequency *tfidf* scores, we report on the performance of a k-Nearest Neighbour (kNN) classifier. We demonstrate that a 10-fold kNN cross validation exhibits an accuracy of nearly 85% in classifying within the Bundeli folk genres when a k of 2 neighbours is used. Using the SVM classification, we achieve an accuracy of over 90%. Following a brief description of related classification techniques applied to song genre classification within Indian and Western music, in general, in section 2 below, we provide a detailed account of methods used for creating news and folk song lyrics corpora in section 3. Following that, in section 4, we emphasize the need to use commonly-removed stopwords towards affecting a classification between news and song lyrics corpora. In this section, we also report on the feasibility of using probability density functions on word level

[1] http://www.ethnologue.com/language/bns
[2] http://www.kavitakosh.org

D S Sharma, R Sangal and J D Pawar. Proc. of the 11th Intl. Conference on Natural Language Processing, pages 133–138,
Goa, India. December 2014.

n-grams to better understand lexical and structural similarity within-genres. In section 5, we present detailed classification analyses on the song lyrics corpus and show the extent of accuracy that can be achieved when popular machine learning techniques are employed to classify within folk genres, that exhibit a great deal of lexical and structural overlap. In Section 6, we present the conclusions our analysis and also future directions for further research.

2 Related Work

Song genre classification as one of the primary tasks of music information retrieval, has been approached from analysis and classification of audio signal features (Kini et al., 2011; Jothilakshmi and Kathiresan, 2012; Tzanetakis and Cook, 2002); retrieval of lyrics based features (Howard et al., 2011; Mayer et al., 2008) and approaches which use both audio and lyrics based features for genre classification (Mayer and Rauber, 2011; Neumayer and Rauber, 2007). Within the Western context, both popular, classical, and folk music genres have been classifed using common machine learning algorithms. The techniques used for classification include both stochastic and probabilistic methods; Hidden Markov Models (Chai and Vercoe, 2001), Machine-learning etc. However, within the Indian context most all work on song genre classification has been restricted to audio feature vector extraction and classification. More precisely, various classification techniques such as Gaussian Mixture Models (GMM), k-nearest neighbour (kNN) classifiers and Support Vector Machine classifiers (SVM) have been employed to classify between Hindustani, Carnatic, Ghazal, Folk and Indian Western genres (Kini et al., 2011) and north Indian devotional music (Jothilakshmi and Kathiresan, 2012). The only instance of lyrics based classification has been explored in the context of Bollywood music in an effort to identify specific features of Bollywood song lyrics using Complex Networks (Behl and Choudhury, 2011). However, classification of folk genres has not received any attention so far. In this paper, we propose a two-fold approach, first, we suggest that classification between broad text genres such as news and songs can be successfully accomplished using common stopwords in Bundeli. And second, we also demonstrate that big data based machine learning approaches could be successfully used to classify, relatively small corpus of Bundeli folk songs into specific genres; namely, Gaari, Rai and Phag.

3 Corpus Creation

The three genres Gaari, Phag and Rai were chosen because they are commercially available in the form of analog audio tapes. These audio tapes were digitized and converted to MP3 format for transcription and future analyses. As far as stylistic register go, gaaris are marriage songs that are sung in repeated choruses. The chorus is periodically repeated after each verse. Phags are lyrical poems, showing rich lexical diversity and rhythmic meter. Rais are dance songs, sung to the beat of the increasing speed of the *Dholak* (local percussion instrument). In Rais, we see the usage of repeated lines more than any other genre. In Gaaris and Rais, a significant level of semantic overlap can be predicted. Both use simple, conversational terms, and owing to their content being repeated, they pose a challenge for classification. These patterns can be evinced from repetitive choruses as in Figure 1, below.

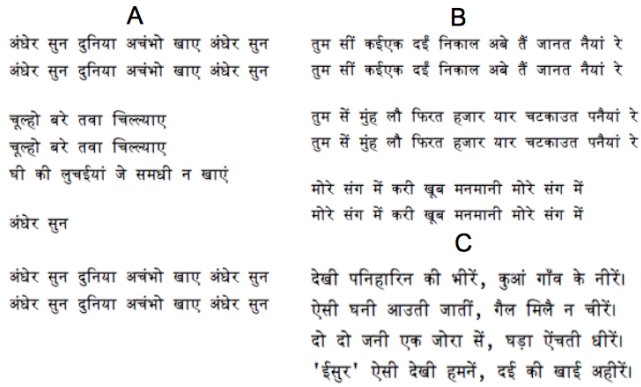

Figure 1: Panel A depicts Gaari with repeated chorus. Panel B shows Rais with repeated lines. Panel C shows the lyrical form of the Phag

Therefore, within-genre classification remains a problem at the textual level because there is significant lexical overlap between Gaaris and Rais.

3.1 Song Corpus

The songs come from various sources. Our first source was a collection from an oral tradition of singing from regions near Sagar, Madhya Pradesh. Online videos and audio cassestes from a Jhansi based casette company, Kanhaiya Casette, became our second source. Phags were mined from a web

resource for poetry and prose which contains collections of the famous Bundeli poet Isuri [3]. Isuri's phags, however, are also available in a few textbooks on Bundeli folk culture. The lyrics were orthographically transcribed by listening to the song files in MP3 format in Devanagari. Owing to poor audio quality of the songs, they required a native Bundeli speaker for transcription, the first author being one. Orthographic normalisation was maintained for words that used non-contrastive phonemic or morphological features. For example, the function word "se" and "sein" meaning "from" was normalised to "sein". Similarly for "unhonein" and "unne", meaning third person ergative was normalised to "unne". Where necessary, similar normalization procedures were used to homogenize orthographic variation. Since there was no available list of stopwords, we adopted the corresponding stoplist from Hindi. We narrowed the scope of the exhaustive stopword list by selecting only the most frequent stopwords occurring in songs.

3.2 News corpus

Before analysing the songs within themselves, we needed to establish a differentiation between a generalised corpus and a song corpus. The only available online resource was the publication of a Delhi-based region specific newspaper[4]. Although the newspaper was published in Bundeli, there was no normalisation of dialectal variation. The articles featuring from Mahoba region alone were the ones that could be employed for comparison with our song corpus. This website was not properly designed using standardised web-designing techniques, and automatic web-scraping techniques. So the data was hand-mined by copy-pasting the relevant region-specific articles. Details of the song and news corpora are presented in Table 1. While 98 news articles formed the news corpus, the song corpus consisted of 37, 39 and 40 Gaaris, Rais, and Phags, respectively.

4 News and Song Corpus Classification

Most common text classification methods, including music genre classification, begin by removing common stopwords derived from a generalized corpus. Token frequencies of a set of 7 common stopwords was used for the purpose of this classification. One way ANOVA with Broad Genre as predictor shows a significant main effect $F[1,26]=5.438; p<0.05$. As the boxplot in Figure 2 shows, stopword token frequencies are significantly higher in news as compared to songs. Thus, stopwords when included in any corpus, can help classify news from songs, even though, more commonly, stopwords are excluded during the preprocessing stages of standard classification-based approaches.

Type of text	Number
News Articles	98
Gaaris (G)	37
Rais (R)	39
Phags (P)	40
Total songs (G+R+P)	116

Table 1: Details of the news and song corpora

We also performed a test of lexical diversity to classify between the news and song corpus. Using Python scripts we calculated the type-token ratio of the two broad genres. The news corpus had a lexical diversity of 0.028% while the song corpus had nearly half the lexical diversity of the news corpus, i.e., 0.014%.

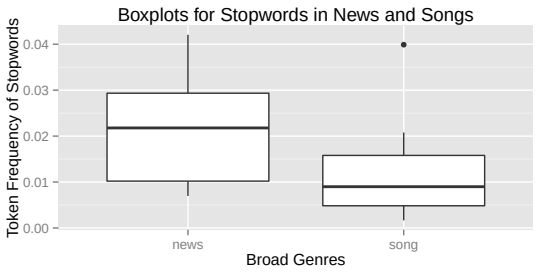

Figure 2: Higher stopword token frequencies in the Bundeli news corpus compared to the songs

4.1 Kernel Density Estimates

Kernel Density Estimates (KDEs) of ngram terms are generated using Equation 1 below. KDE is a non-parametric estimate of the probability density function of random variables, in this case the counts of the ngrams. KDEs allow for better inferences about the population, based on a real and finite data sample. KDEs make it possible for us to examine the probability density function of ngrams for the various genres. Based on the

[3] http://www.kavitakosh.org
[4] http://www.khabarlahariya.org/?cat=64

KDEs we make inferences about the possible presence of categorical features in the sample based on smoothed bins and probability density. Equation 1 shows the KD function estimator. In our estimates, we use a normal kernel such that K(x) = ϕ(x), where ϕ is the standard normal density function.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\frac{(x - x_i)}{h} \tag{1}$$

4.2 Structural Description using KDEs

After making a collection of the entire corpus, we employed n-gram methods to examine the structural distribution of the data with respect to the genres. Therefore, using Python scripts, we first generated unigrams, bigrams and trigrams for all the 3 genres. Then, we used R Studio to generate KDEs to identify the n-gram based structural description of Gaaris, Rais and Phags. A frequency range was used such that it captured the most frequent n-gram terms in Gaaris, Rais and Phags.

The figures show a representation of the n-gram structure of the songs. As can be seen in the three figures, unigrams and bigrams show a high degree of overlap between the three genres in the same frequency range. The genres cannot be differentiated using the structural description detailed by the unigrams and bigrams. In the case of trigrams, although the peaks of the distribution differ, the probability density is extremely low. These terms would easily be eliminated in the sparsity calculation and cannot be termed as predictors of the genre-variation. The KDEs show a structural description of the songs, establishing the existing overlap in the three genres.

5 Classification using Machine Learning Techniques

kNN, SVM and Naïve Bayes classifiers were the machine-learning techniques used for the purpose of this classification. The pre-processing techniques included collection and compilation of the corpus made of text files. The corpus was cleaned of stopwords, punctuation marks and white-spaces. The sparsity was set to 85%. A term-document matrix was created from the corpus where each song file is converted to a vector space where term frequency-inverse document frequencies (*tfidf*) are stored. The *tf* (term frequency)

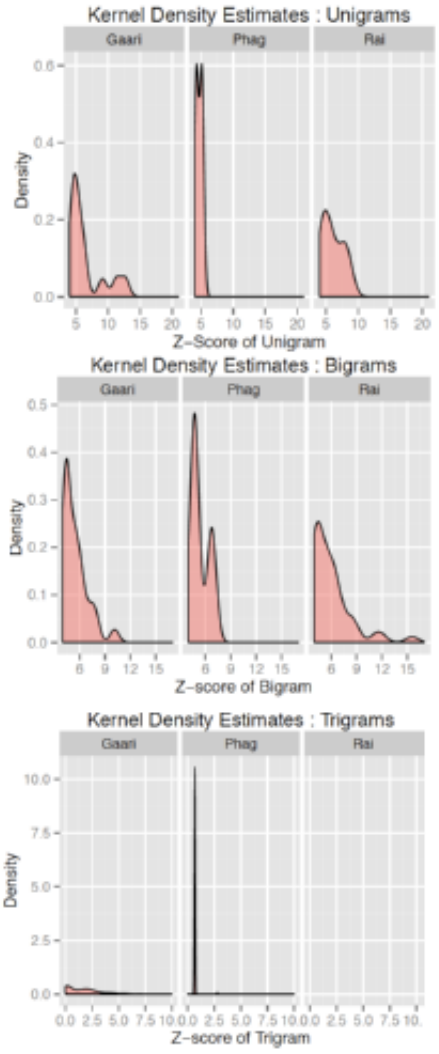

Figure 3: Kernel Density Estimates for Unigrams, Bigrams and Trigrams

scores were calculated with Equation 2 and the *inverse document document frequency (idf)* scores were calculated with equation 3. The product of the tf and idf scores are used to train a classifier with 10-fold cross-validation.

$$tf(t,d) = 0.5 + \frac{0.5 \times f(t,d)}{max\{f(w,d) : w \epsilon d\}} \tag{2}$$

$$idf(t,D) = log\frac{N}{|d \epsilon D : t \epsilon d|} \tag{3}$$

The tf-idf scores are a product of the tf and idf scores from equations 3 and 4.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{4}$$

5.1 kNN Classification

A 10-fold kNN classifier is trained on a term document matrix where each song file is converted to

Test error	CV Mean error	CV Std.error	K
0.2173913	0.1588889	0.03085654	1
0.1739130	0.1600000	0.02772588	2
0.1739130	0.2155556	0.04641319	3
0.1739130	0.2588889	0.04930379	4
0.1739130	0.2733333	0.05612547	5
0.2173913	0.3322222	0.03668163	6
0.2173913	0.3222222	0.05670418	7
0.2173913	0.3322222	0.05772196	8
0.2173913	0.3444444	0.05062030	9
0.2608696	0.3688889	0.03988323	10

Table 2: 10-fold Cross-Validation results for a kNN classifier: Reported Test errors, Cross-Validation (CV) Mean errors, CV Standard errors and the associated k-neighbour

a vector space where term frequency-inverse document frequencies (*tfidf*) are stored.

For training, the rows and columns of the training set and the genre names for just the training set, and for testing, the rows and columns excluding that of genre names is passed to the kNN model. A sparsity of 85% is maintained for the training dataset.

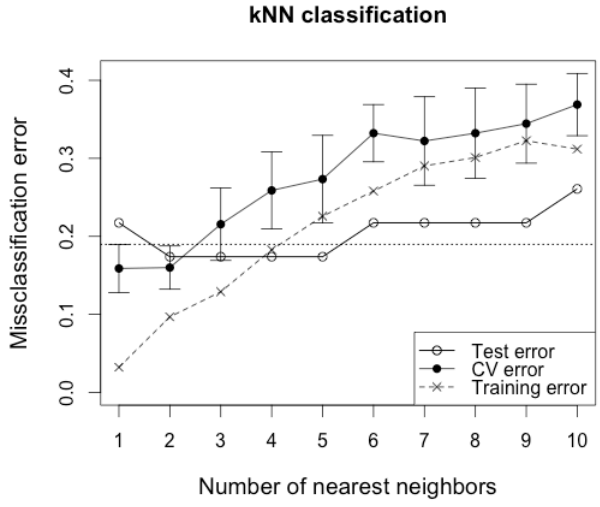

Figure 4: Misclassification errors from kNN training, testing and cross-validation

A first pass kNN classification yields an accuracy of 86.95% when k=2 nearest neighbours are used to measure the distance between the *tfidf* scores with a precision of 0.818 and an F-score of 0.9. A 10-fold cross-validation is then employed to choose the optimal k-neighbour, results of which are given in Table 2. Based on the lowest test, CV mean and CV Standard errors an optimal K=2 is found to be the best kNN classifier for our dataset. The results of the k=fold cross-validation can be seen in Figure 4.

5.2 SVM Classification

We use the e1071 package in R to perform a SVM classification on the songs dataset (Meyer et al., 2014). The term-document matrix used for kNN classification is passed to the SVM algorithm. We partition the corpus into training and test sets. An 80-20% partition is used for training and test, respectively. Thus, the training dataset contains 93 songs, and the test dataset has 23 songs. The training data contains the genre-labels but the test dataset does not. When a 5-fold cross-validated SVM is trained, the SVM performs classification with a diagonal accuracy of 91.3% and kappa accuracy of 85.7%. Table 3 shows the confusion matrix as predicted by the SVM Classifier.

	True		
Predictions	Gaari	Rai	Phag
Gaari	11	2	0
Rai	0	3	0
Phag	0	0	7

Table 3: True positives and false positives for Naïve Bayes Classification

5.3 Naïve Bayes Classification

We use the e1071 package in R to perform a Naïve Bayes classification on the songs dataset (Meyer et al., 2014). The term-document matrix used for kNN classification is passed to the Naïve Bayes algorithm. We partition the corpus into training and test sets. An 80-20% partition is used for training and test, respectively. Thus, the training dataset contains 93 songs, and the test dataset has 23 songs. The training data contains the genre-labels but the test dataset does not. When trained, the Naïve Bayes performs classification with a diagonal accuracy of 78.2% and kappa accuracy of 68.4%. Table 4 shows the confusion matrix as predicted by the Naïve Bayes Classifier.

6 Conclusions and Further Research

In this paper, we explored both statistical and machine-learning techniques to perform lyrics-

	True		
Predictions	Gaari	Rai	Phag
Gaari	6	0	0
Rai	5	5	0
Phag	0	0	7

Table 4: True positives and false positives for Naïve Bayes Classification

based classification within the genres of Bundeli folk music. Using different sources, we created a corpus of 116 Bundeli folk songs to perform classification. To separate lyrics from standard Bundeli texts, we performed a broad-genre classification using stopwords and lexical diversity measures. Finally, we extended existing machine-learning techniques to successfully classify the three genres. Our findings report that popular methods of classification that are employed on 'big data' can be used to perform within-genre classification. Our results indicate that the SVM and kNN Classifiers perform better than Naïve Bayes classifier.

The present research can be extended to classify more genres in Bundeli folk-music. The models can be further expanded to include genre-classification from other dialects of Western Hindi like Awadhi, Bagheli and Braj. The lyrics-based approach can be combined with an audio-feature vector analysis to build a multi-modal classfication system.

Acknowledgements

We are deeply thankful and appreciative of three anonymous reviewers for providing us with comments and suggestions that helped us better formulate our research initiative and reposition our efforts to apply 'big data' machine learning techniques to our 'small data' classification problem.

References

Aseem Behl and Monojit Choudhury. 2011. A corpus linguistic study of bollywood song lyrics in the framework of complex network theory. In *International Conference on Natural Language Processing*. Macmillan Publishers, India.

Wei Chai and Barry Vercoe. 2001. Folk music classification using hidden markov models. In *Proc. of International Conference on Artificial Intelligence*.

Sam Howard, Carlos N. Silla Jr., and Colin G. Johnson. 2011. Automatic lyrics-based music genre classification in a multilingual setting. In *Thirteenth Brazilian Symposium on Computer Music*, 31st August–3rd September 2011.

S. Jothilakshmi and N. Kathiresan. 2012. Automatic music genre classification for indian music. In *International Conference on Software and Computer Applications (ICSCA 2012)*.

S. Kini, S. Gulati, and P. Rao. 2011. Automatic genre classification of north indian devotional music. In *Proceedings of the National Conference on Communications (NCC)*, pages 1–5, Jan 2011, Bangalore, India.

Rudolf Mayer and Andreas Rauber. 2011. Music genre classification by ensembles of audio and lyrics features. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 675–680, Miami (Florida), USA, October 24-28. http://ismir2011.ismir.net/papers/PS6-4.pdf.

Rudolf Mayer, Robert Neumayer, and Andreas Rauber. 2008. Rhyme and style features for musical genre classification by song lyrics. In *In Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR'08*.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch, 2014. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3.

Robert Neumayer and Andreas Rauber. 2007. Integration of text and audio features for genre classification in music information retrieval (accepted for publication. In *Proceedings of the 29th European Conference on Information Retrieval (ECIR'07*, pages 724–727.

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