PuPoCl: Development of Punjabi Poetry Classifier Using Linguistic Features and Weighting

JASLEEN KAUR[1]
JATINDERKUMAR R. SAINI[2]

Narmada College of Computer Application
Ganesh Shrishti, Vadadla, Bharuch
GJ SH 164, Gujarat 392011, India
[1]sidhurukku@yahoo.com
[2]saini_expert@yahoo.com

Abstract. Analysis of poetic text is very challenging from computational linguistic perspective. For library suggestion system, poetries can be characterized on different measurements, such as writer, time period, sentiments, emotions and topic. In this paper, subject based Punjabi poetry classifier was developed using Weka tool. Four different categories were manually populated with 2034 poems (NAFE, LIPA, RORE, PHSP) categories consists of 505, 399, 529 and 601 numbers of poetries, respectively. After tokenization of 2034 poetries, 45667 features were extracted and passed to noise removal sub phase. A total of 31938 features were extracted, after removal of noise, and weighted using term frequency and the entire process is repeated for tf-idf weighting scheme also. Two types of Linguistic features namely: lexical features and syntactic features of poetries were explored to develop classifier using machine learning algorithms. Naive Bayes, Support Vector Machine, Hyper pipes and K-Nearest Neighbour algorithms were experimented with 31938 lexical features and 30396 syntactic features. Result shows that SVM outperformed all other classifiers using TF and TF-IDF weighing schemes whereas KNN is the worst performer. With addition of POS tags with words, accuracy of SVM is increased by 1%. Result also revealed that with testing time of 0.19 seconds, SVM is the most efficient machine learning algorithm for Punjabi poetry classification, using TF-IDF scheme.

Keywords: Classification, Lexical, Part of Speech, Poetry, Punjabi, Syntactic.

(Received March 14th, 2017 / Accepted October 31st, 2017)

1 Introduction

Poems are artistic form of writing. As said by Wordsworth, "Poems are the spontaneous overflow of powerful feelings". With the help of rhythm, meter, imagination and words, poetry becomes imaginative piece of writing. By making use of memory and senses gifted to human being, they can easily differentiate poems from simple text. But processing such imaginative pieces of writing using machine is very challenging. These artistic pieces of writing can be classified using various metrics such as poet, historic period, emotions associated with the text, theme focussed by poet in the poem.

In this paper, content based classification of poetries is carried out. This paper focuses on subject based analysis of poetries written in Punjabi language. With the help of words/lexicons/tokens used in poetries, poems are categorised into predetermined classes.

A lot of research has been carried out in automated classification of poems written in foreign languages, especially English. However, this area still needs to be explored in Indian languages. No work has been reported on Punjabi poetry. Punjabi is the tenth most spoken language of the world and the third most spoken lan-

guage in India [17]. Punjabi language belongs to Indo-Aryan language family (Kaur, J. and Saini, J.R. (2014)) and has more than 130 million speakers worldwide including India, Pakistan, Canada and United Kingdom. In this work, we are building Punjabi poetry classifier (PuPoCl).

In this work, the emphasis is on the vocabulary of poem that determines its subject. Understanding the vocabulary of poem to determine its class is the most tedious task. Because poetries are imaginative pieces of text, poet may use different words in different context in poetry. Furthermore, the exact interpretations that humans assign to poems depend on their individual experiences. The computational linguistic analysis of poetry is very challenging. This work is important, not only for better understanding of rich literature but also has application in library for making recommendations to readers based on their literary taste. Our work is the first of its kind to do automatic classification of Punjabi poetry.

2 Related work

Automatic analysis of poetry is done for poems written in various languages like English, Chinese, Arabic, Malay, and Spanish. Brief review of same is given in this section.

Barros L. et.al, 2013 [2] tried to automatically categorize poems based on their emotional content. For this experiment, they have used a Quevedo's poetry written in Spanish. A reference classification of the same (Bleuca Categorization) is also used during the experimentation. Decision Tree is built using Weka toolkit for classification problem. The accuracy of this classifier is 56.22%, which is increased to 75.13% by using resample filter. This experiment is done to determine whether a classifier with information about emotions detected in a given Quevedo's poem can able to reproduce Bleuca's Categorization. Hamidi S. et.al 2009 [7] proposed a meter classification system for Persian poems based on features extracted from uttered poem. In the first stage, the utterance has been segmented into syllables using three features, pitch frequency and modified energy of each frame of the utterance and its temporal variations. In the second stage, each syllable is classified into long syllable and short syllable classes which is a convenient categorization in Persian literature. In this stage, the classifier is support vector machine (SVM) classifier with radial basis function kernel and employed features are the syllable temporal duration, zero crossing rate and PARCOR coefficients of each syllable. The sequence of extracted syllables classes is then compared with classic Persian meter styles using dynamic time warping, to make the system robust against syllables insertion, deletion or classification. The system has been evaluated on 136 poetries utterances from 12 Persian meter styles gathered from 8 speakers, using k-fold evaluation strategy. The results show 91% accuracy in three top meter style choices of the system. Jamal N. et.al 2012 [8] represents classification of Malay pantun using SVM. Pantun is traditional Malay poetry. The capability of SVM through radial basic function (RBF) and linear kernel functions are implemented to classify pantun by theme, as well as poetry or non-poetry. A total of 1500 pantun are divided into 10 themes with 214 Malaysian folklore documents used as the training and testing datasets. Term frequency- inverse document frequency (TF–IDF) used for both classification experiments. The highest average percentage of 58.44% accuracy was found for the classification of poetry by theme. The results of each experiment showed that the linear kernel achieved a better percentage of average accuracy compared to the RBF kernel. Kumar and Minz 2012 [14] worked to find the best classification algorithms among the K-Nearest Neighbour (KNN), Naive Bayes (NB) and SVM with reduced features for classification of poems. Information Gain Ratio is used for feature selection. The results showed that SVM has maximum accuracy (93.25 %) using 20 % top ranked features.

Alsharif et.al 2013 [1] tried to classify Arabic poetry according to emotion associated with it. The problem was treated as a text categorization problem, classifying poems into four classes: Retha, Ghazal, Heja and Fakhr. Four machine learning algorithms are compared: Naive Bayes, SVM, VFI (Voting Feature Intervals) and Hyperpipes. The best precision achieved was 79% using Hyperpipes with non-stemmed, non-rooted, mutually deducted feature vectors containing 2000 Features. Can et. al 2012 [3] investigated two fundamentally different machine learning text categorization methods, Support SVM and NB, for categorization of Ottoman poems according to their poets and time periods. Dataset comprises of the collected works (divans) of ten different Ottoman poets. The result shows that SVM, with almost 90% accuracy, is a more accurate classifier compared to NB in categorization tasks. Lou et al. 2015 used SVM to classify poems in English into 3 main categories and 9 subcategories by combining TF-IDF and Latent Dirichlet Allocation. All this work has been done for English. Figure 1 summarizes various poetry classification works done in foreign languages and Indian languages.

A lot of research has been reported in various foreign languages but scenario is bit different for Indian languages. Not much work has been reported for Indian

languages. Bangla poetry classification is done by Rakshit et.al 2015 [18]. Poetries are classified on the basis of subject and accuracy reported by SVM classifier is 56.8% and this work is extended to poet identification using stylometric analysis of poetries. But no such poetry classifier is developed for Punjabi poetry. Our work is the first of its kind for Punjabi language.

3 Methodology

This section presents the detailed process followed in this paper to find the best machine learning algorithm for Punjabi poetry classification task. The system view of poetry classifier is presented in Figure 2. It consists of text classification process steps viz. corpus building, text pre-processing, feature extraction, feature selection, model building and model evaluation.

Due to unavailability of readymade Punjabi poetry corpus in public domain, dataset collection and identifying different categories was a major challenge. The corpus of Punjabi poetry was populated manually. Total 2034 poems were collected from various online sources such as `http://www.punjabi-kavita.com/`, `http://www.punjabizm.com/`, http://punjabimaaboli.com/. Considering the data collected, initially five different categories (Nature, Festival, Romantic, Religious, and Philosophical) were identified. One major challenge encountered was the overlapping of words among five identified categories, and ambiguity in assigning any poem to any category. So, in order to solve these problems, poetries in five categories were reorganised into four categories. These four categories are NAFE, LIPA, PHSP and RORE.

NAFE stands for Nature and Festival. This category consists of nature related poetries. And many Punjabi festivals are connected with nature phenomenon, both the categories are emerged. LIPA stands for Linguistic and Patriotic. This category includes patriotic and linguistic poetry. To avoid the confusion in words, patriotic category is combined with linguistic category which consists of poetry related to Punjabi language. 'RORE' stands for Romantic and Relation. This category consists of romantic poetry as well as poetry related to different relations. 'PHSP' stands for Philosophical and Spiritual. This category includes poems related to philosophy and religious poetry. All these poetries were converted in Unicode format for further processing [19]. As can be observed here, each of the four categories chosen in this paper was a combination of two categories.

Total 2034 poetries were used for computational classification of Punjabi poetry. NAFE, LIPA, RORE, PHSP categories consists of 505, 399, 529 and 601 numbers of poetries, respectively. For subject based analysis of poetry, the vocabulary of poetry was used as feature. Poetries were submitted to tokenization sub phase, where individual words were extracted from poems. These extracted words were used to create 'bag of words' from poetries. These words were passed for noise removal and pre-processed to remove noise, present in form of special symbols, Punjabi numerals and stop words, from them [9]. These extracted words were considered as linguistic features. In this work, two types of linguistic features were selected to build classifiers:

1. Lexical features: Each word type was considered as a feature and weight to it was assigned using term frequency (TF) and term frequency-inverse document frequency (TF-IDF).

2. Syntactic features: Each word type followed by its part of speech tag (POS) was considered as a feature and weight to it was assigned using TF and TF-IDF. Part of speech tags were generated using Punjabi part of speech tagger [17].

After tokenization, a total of 45667 tokens were extracted from 2034 poetries. These 45667 words were pre-processed to remove special poetry symbols such as Comma (,), Dandi, Double Dandi, Punjabi numerals (from 1 to 10 written in Punjabi) and 184 stop words identified by Kaur J. and Saini JR, 2015 [10]. After pre-processing, 31938 words are extracted from 2034 poetries that are further used for building classifier. TF and TF-IDF were used to weigh these extracted words. In order to build the model different machine learning algorithms (Support Vector Machine (SVM), Naive Bayes (NB), k-nearest neighbour (KNN) and Hyperpipes (HP)) were trained and tested. These machine learning algorithms were divided into two broad categories: lazy learners and eager learners, based on their execution pattern [5]. These machine learning algorithms were simulated using Weka tool [6]. Weka is data mining software in Java, developed by University of Waikato, New Zealand. As indicated by numerous independent researches carried out on the domain of poetry classification (in foreign languages) [8], [14], [1] and text classification in Indian languages [12], SVM, KNN, NB and HP performed well. So, these algorithms were chosen to build the model. Also, it has been shown that in absence of any other parametric data SVM could be used for classification purposes [4]. Performance of these models is reported in terms of Accuracy (%) and Efficiency (in terms of training time and testing time) in the next section.

S. no	Language		Dataset(size, classes)	Algorithm	Performance	Remarks
1	Foreign	Spanish	Quevedo's poetry 185 poems, 4 classes	Decision tree using Weka tool.	Accuracy: 56.22%	Accuracy increased to 75.13% using resample filter
2.		Persian	136 poetries utterances from 12 Persian meter styles	SVM with radial basis kernel	Accuracy: 91%	meter classification system
3		Malay	Malay pantun. 1500 pantun, 10 classes	SVM with TF-IDF	Accuracy: 58.44%	Classify pantun by theme, as well as poetry or non-poetry.
4		English	www.poetseer.com, www.poemhunter.com (400 poems, 8 classes)	SVM, NB, KNN	Accuracy: 93.25% with SVM	Using Gain ratio for feature selection
5		Arabic	Four categories: Retha, Ghazal, Heja and Fakhrtotal 1231 poems, 4 classes	NB, SVM, VFI (Voting Feature Intervals) and Hyperpipes	Accuracy: 79% with Hyperpipes	Emotion based classification of Arabic poetry
6		ottoman	the collected works (divans) of ten different Ottoman poets	SVM and NB	Accuracy: 90% with SVM	Classification according to poet and time period
7		English	7214 poems,3 main classes and 9 sub classes	SVM with TF-IDF and LDA	Precision: 0.848 with SVM	Multi label classification
8	Indian	Bangla	1341 poems, 4 classes	SVM, NB	Accuracy: 56.80% with SVM	Classification according to Subject and poet

Figure 1: Summary of Poetry Classifiers

Learning	Classifier	Features	Weighing scheme TF		
			Accuracy (%)	Efficiency (sec)	
				Training time	Testing time
Lazy	HP	LEX	62.75	3.17	1.22
		LEX+POS	64.72	1.45	1.95
	KNN	LEX	40.08	1.00	4.41
		LEX+POS	46.16	4.86	6.10
Eager	NB	LEX	59.69	35.25	19.30
		LEX+POS	60.01	35.14	20.78
	SVM	LEX	**72.04**	28.03	0.48
		LEX+POS	**72.15**	37.82	0.62

Figure 2: System view of Poetry Classifier

4 Results and Discussions

The goal of this work was to find the best machine learning algorithm that works well for Punjabi poetry classification. So, to achieve this objective, different classifiers (SVM, KNN, NB, and HP) were trained and tested using different linguistic features. The performance of classifiers is evaluated using 10-fold cross validation. Training time and testing time are reported in seconds and represented as 'seconds'.

Figure 3 shows the result of classifiers using TF as weighing scheme. The features column consists of two values: LEX and LEX+POS. LEX indicates building classifiers using LEXical features only and LEX+POS indicates building classifier model using lexical features followed by its part of speech tag. The confusion matrix for SVM classifier is shown in Table 1.

Learning	Classifier	Features	Weighing scheme TF		
			Accuracy (%)	Efficiency (sec)	
				Training time	Testing time
Lazy	HP	LEX	62.75	3.17	1.22
		LEX+POS	64.72	1.45	1.95
	KNN	LEX	40.08	1.00	4.41
		LEX+POS	46.16	4.86	6.10
Eager	NB	LEX	59.69	35.25	19.30
		LEX+POS	60.01	35.14	20.78
	SVM	LEX	**72.04**	28.03	0.48
		LEX+POS	**72.15**	37.82	0.62

Figure 3: Performance Evaluation of Classifiers using TF as weighing scheme.

Table 1: Confusion matrix for SVM classifier.

	NAFE	LIPA	RORE	PHSP
NAFE	370	14	63	58
LIPA	16	306	18	59
RORE	54	10	371	94
PHSP	43	33	93	432

As indicated in Table 1, by weighing features using term frequency, SVM (with accuracy of 72.04%) outperformed all other machine learning algorithm. Whereas, KNN is the worst performer, with accuracy of 40.08%. Slight change (0.9%) in the performance of SVM can be observed from the Table 1 by using POS tags. It can be observed from Table 1 that SVM performed better as compared to all other machine learning algorithms.

It can be observed from the confusion matrix that LIPA is the best classified category with accuracy of 76.69% and RORE is the worst classified category having accuracy 70.13%. PHSP and RORE are the most confusing categories because 94 PHSP poetries are often mis-classified as RORE poems. It is possible due to overlapping of poetic words in different categories. Figure 4 shows the result of classifiers using TF-IDF as weighing scheme.

Results of different classification algorithms using different linguistic features and weighing schemes are shown in Figure 5. As it can be clearly observed from Figure 5 that SVM is the winning classifier under all criteria. Except for SVM, other three algorithms are having consistent performance under TF and TF-IDF weighing scheme.

From Figure 3, it can be interpreted that maximum accuracy reported by using lexical features to build the classifier model is 72.04% and 66.43% using TF and TF-IDF weighing scheme, respectively. SVM is the highest performer, HP is second highest performer, followed by NB and KNN is the worst performer with accuracy of 40.08% with both the weighing schemes. Rakshit G and Puspak B (2015) [18] had tried to build subject based poetry classifier that work on bangla poetry. They had classified bangla poetry using lexical features with its TF-IDF value and accuracy reported by SVM was 56.80% Rakshit G and Puspak B (2015) [18]. It can be interpreted from the results that lexical features works well for building punjabi poetry classifier as compared to bangla poetry classifier.

5 Conclusion

Computational linguistic analysis of Punjabi poetry is done to categorize poems based on its subject. For content based analysis, total 2034 poems are collected in four different categories. In this paper, experimentation is done using two types of linguistic features: Lexical and Syntactic features. Lexical features consist of words used in poetry whereas, words followed by their POS tag is used as syntactic feature. Term frequency and term frequency-inverse document frequency are used to weigh features. Result shows improvement in accuracy with addition of POS tags. Different classifiers are trained and tested using lexical features as well as syntactic features in Weka. SVM outperformed all other machine learning algorithms using different linguistic features, weighed by TF and TF-IDF. Using Lexical features, SVM achieved accuracy of 72.04% and 66.43% weighed by TF and TF-IDF, respectively. And using POS tags, accuracy reported by SVM is 72.15% and 70.65% with TF and

Learning	Classifier	Features	Weighing scheme TF-IDF		
			Accuracy (%)	Efficiency (seconds)	
				Training time	Testing time
Lazy	HP	LEX	62.75	4.31	2.03
		LEX+POS	64.72	0.45	6.58
	KNN	LEX	40.08	2.16	4.77
		LEX+POS	46.16	6.28	0.85
Eager	NB	LEX	59.69	37.68	22.78
		LEX+POS	60.02	38.40	23.34
	SVM	LEX	**66.43**	49.35	0.19
		LEX+POS	**70.65**	47.12	0.40

Figure 4: Performance Evaluation of Classifiers using TF-IDF as weighing scheme.

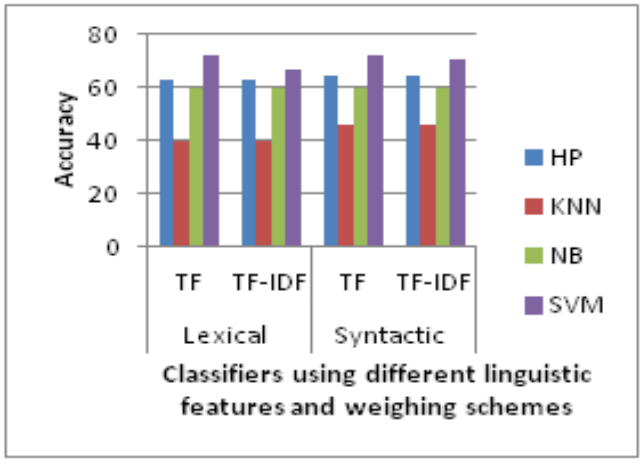

Figure 5: Performance of different classifiers using different linguistic features and weighing schemes.

TF-IDF weighing scheme. In comparison with TF-IDF, TF weighing scheme performed well for Punjabi poetry classification task. Testing time taken by SVM using lexical features is reported as 0.48 seconds and 0.19 seconds with TF and TF-IDF weighing scheme, respectively. SVM is also the most efficient machine learning algorithm in terms of testing time it takes.

References

1. Alsharif, O., Alshamaa, D. and Ghneim, N. (2013). Emotion Classification in Arabic Poetry using Machine Learning. International Journal of Computer Applications; 5(16); 10-15.

2. Barros, L., Rodriguez, P., and Ortigosa, A. (2013). Automatic Classification of Literature Pieces by Emotion Detection: A Study on Quevedo's Poetry. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII); 2-5 Sept. 2013; 141-146.

3. Can, E.F., Can, F., Duygulu, P., Kalpakli M. (2012). Automatic Categorization of Ottoman Literary Texts by Poet and Time Period. Computer and Information Sciences-II; 51-57.

4. Chandrakar O. S. and Saini J. R., "Empirical Study to Suggest Optimal Classification Techniques for Given Dataset", published in the proceedings of IEEE International Conference on Computational Intelligence & Communication Technology (CICT-2015) held during 13-14 Feb. 2015 at ABES Engineering College, Ghaziabad, India; ISBN: 978-1-4799-6022-4 & 978-1-4799-6024-8; published by IEEE Computer Society; Feb. 2015;

30-35; DOI: 10.1109/CICT.2015.2

5. Chatterjee D., "Eager learning versus Lazy learning methods in Classificiation", Thesis submitted in Jadavpur University, 2013.

6. Hall, M., Frank,E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H., (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations; 11(1).

7. Hamidi, S., Razzazi, F., and Ghaemmaghami, M.P. (2009). Automatic Meter Classification in Persian poetries using Support Vector Machines. IEEE International Symposium on Signal Processing and Information Technology (ISSPIT); 563-567.

8. Jamal, N., Mohd, M., & Noah, S. A. (2012). Poetry Classification Using Support Vector Machines. Journal of Computer Science; 8(9); 1441-1446.

9. Kaur J. and Saini J. R. (2015). A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language. International Journal of Data Mining and Emerging Technologies; ISSN: 2249-3212 (eISSN: 2249-3220); Indian Journals, New Delhi, India; 5(2); 114-120.

10. Kaur J. and Saini J. R. (2015). A Study of Text Classification Natural Language Processing Algorithms for Indian Languages. The VNSGU Journal of Science and Technology; ISSN: 0975-5446; Journal of The Veer Narmad South Gujarat University, Surat, Gujarat, India; 4(1), July 2015; 162-167.

11. Kaur J. and Saini J. R. (2017). Automatic Punjabi Poetry Classification Using Machine Learning Algorithms with Reduced Feature Set. International Journal of Artificial Intelligence and Soft Computing; 5(4); 311-319; [doi>10.1504/IJAISC.2016.081353].

12. Kaur, J. and Saini, J. R. (2015). POS Word Class based Categorization of Gurmukhi Stemmed stop words. International Conference in Information Communication Technology for Intelligent System, Smart Innovation in Smart Technology Springer; 2; 3-10.

13. Kaur, J. and Saini, J. R. (2014). A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families. International Journal of Data Mining and Emerging Technologies; 4(2); 53-60.

14. Kumar, V. and Minz, S. (2012). Poem Classification using Machine Learning.International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Advances in Intelligent Systems and Computing; 236; 675-682.

15. Lou, A., Inkpen, D. and Tan, C.,(2015). Multilabel Subject-Based Classification of Poetry, Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference; 187-192.

16. Punjabi language. Retrieved November 01, 2015, from `https://simple.wikipedia.org/wiki/Punjabi_language`

17. Punjabi POS tagger accessed on December 17, 2015 from http://punjabipos.learnpunjabi.org/default.aspx December 2015

18. Rakhsit, G., Ghosh, A., Bhattacharyya, P. and Haffari, G., (2015).Automated Analysis of Bangla Poetry for Classification and Poet Identification. 12th International Conference on Natural Language Processing, December 2015.

19. Unicode Table. Retreived on October 2015, from http://www.tamasoft.co.jp/en/general-info/unicode-decimal.html.