Designing Punjabi Poetry Classifiers Using Machine Learning and Different Textual Features

Jasleen Kaur[1] and Jatinderkumar Saini[2]
[1]Department of Computer Engineering, PP Savani University, India
[2]Symbiosis Institute of Computer Studies and Research, India

Abstract: *Analysis of poetic text is very challenging from computational linguistic perspective. Computational analysis of literary arts, especially poetry, is very difficult task for classification. For library recommendation system, poetries can be classified on various metrics such as poet, time period, sentiments and subject matter. In this work, content-based Punjabi poetry classifier was developed using Weka toolset. Four different categories were manually populated with 2034 poems Nature and Festival (NAFE), Linguistic and Patriotic (LIPA), Relation and Romantic (RORE), Philosophy and Spiritual (PHSP) categories consists of 505, 399, 529 and 601 numbers of poetries, respectively. These poetries were passed to various pre-processing sub phases such as tokenization, noise removal, stop word removal, and special symbol removal. 31938 extracted tokens were weighted using Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme. Based upon poetry elements, three different textual features (lexical, syntactic and semantic) were experimented to develop classifier using different machine learning algorithms. Naive Bayes (NB), Support Vector Machine, Hyper pipes and K-nearest neighbour algorithms were experimented with textual features. The results revealed that semantic feature performed better as compared to lexical and syntactic. The best performing algorithm is SVM and highest accuracy (76.02%) is achieved by incorporating semantic information associated with words.*

Keywords: *Classification, naïve bayes, hyper pipes, k-nearest neighbour, Punjabi, poetry, support vector machine, word net.*

Received April 7, 2017; accepted July 8, 2018
https://doi.org/10.34028/iajit/17/1/5

1. Introduction

"Poetry: the best words in the best order."-Samuel Taylor Poetries are beautiful composition formed with assistance of different elements such as words, rhyme, rhythm, meter, and imagination. Being a human, we can easily differentiate between normal texts from poetic text. But processing poetic text through the use of computer is very difficult. With the introduction of Unicode encoding in World Wide Web, poetic text on the web is increasing day by day. So, this poetic text need to be classified for its easy and fast retrieval.

An Automatic poetry classifier takes a poem as an input and identifies its category as an output. An Automatic poetry classification is treated as a text classification problem [25]. In this work, poetries are classified on the basis of its content. For this classification work, vocabulary of poem is used. As poetries are imaginative piece of writing, same word may have different meaning and different meaning, depending upon the context in which it has been used. As India is having a rich literature, this poetry classification system is having an application in classifying literature pieces, poetry, according to the theme of poetry.

Punjabi language belongs to Indo-Aryan language family and holds 10th rank among most spoken language of the world and at 3rd position in most spoken language in India [10, 17, 19]. A lot of research has been carried out in automated classification of poems written in foreign languages, especially English. But this area still needs to be explored in Indian languages. No work has been reported on 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 *et al*. [3] tried to automatically categorize poems based on their emotional content. For this experiment, they have used a Quevedo's poetry written in Spanish. Accuracy reported by Decision Tree is 56.22%, which is increased to 75.13% by using resample filter. Hamidi *et al*. [7] proposed a meter classification system for Persian poems based on features extracted from uttered poem. The results show 91% accuracy in three top meter style choices of the system. Jamal *et al*. [8] represents classification of Malay pantun using Support Vector Machines (SVM). 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. Kumar and Minz [15] worked to find the best classification algorithms among the K-Nearest Neighbour (KNN), Naïve Bayesian (NB) and SVM

with reduced features for classification of poems. The results showed that SVM has maximum accuracy (93.25 %). Alsharif *et al*. [1] tried to classify Arabic poetry according to emotion associated with it. Four machine learning algorithms are compared: NB, SVM, Voting Feature Intervals (VFI) and Hyperpipes. The best precision achieved was 79% using Hyperpipes. Can *et al*. [4] investigated two fundamentally different machine learning text categorization methods, SVM and NB, for categorization of Ottoman poems according to their poets and time periods. The result shows that SVM outperformed NB. Lou *et al*. [16] used SVM to classify poems in English by combining tf-idf and Latent Dirichlet Allocation. All this work has been done for English.

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*. [23]. Poetries are classified on the basis of subject and accuracy reported by SVM classifier is 56.80 %. But no such poetry classifier is developed for Punjabi poetry. Our work is first of its kind for Punjabi language.

3. Understanding Punjabi Poetry

Poetry is an art form in which human language is used for its aesthetic qualities in addition to its notional and semantic content.

3.1. Elements of Poetry

Poetry is made up of different elements such as sound device, diction, form, rhyme, and rhythm. Diction: The kind of language chosen for a poetry, its range of vocabulary, is its diction. A good vocabulary, gained through reading good poetry and prose, enhances poetic expression. Diction depends upon the language in which poetry is written. Clarity: The purpose of all writing is to communicate effectively. Literature pieces, especially poetry, create word pictures. By making proper selection of words, poet can converse more effectively and efficiently.

3.2. Selection of Features

Selection of linguistic features used for experimentation was based on different elements of poetries. Following features of poetries were studied and used for this research work.

3.2.1. Lexical Feature

Diction is the main crafting element of poetry, especially for content based classification of poetries. For content based classification of poetry, words used in the poetries were chosen and used as feature. This feature is named as ‘Lexical’ because in natural Language processing terminology, words are termed as lexicons.

3.2.2. Syntactic Feature

Word class categorization of lexicons was utilized for experimentation. This feature was named as syntactic features because words with their part of speech tags were used. In this research work, we have used Punjabi part of speech tagger [18].

3.2.3. Semantic Feature

As Poetry is imaginative piece of writing, so poet may use word with multiple meaning. So, in order to identify the correct meaning/sense, semantic part of linguistic is considered and this feature is included in the research. Algorithm used for extracting synonym information is as follows.

Algorithm Synonym_info(Tokenized poetry P_t)

For input token from P_t, create a bag B_t.
For each input token t, search in Punjabi synset database.
{
If token is mapped to one synset ID,
then that sense s is added to poetry sense bag P_s.
elseIf token t is mapped to more than one synset ID
{
1. For each synset ID, create a bag B_s of
Synonym words and example sentences.
2.Measure the overlap between each B_s and B_t.
3 Select Bs with maximum overlap and add that
to tokenized sense poetry Bag P_s
}
}
Return (Poetry Sense bag P_s)

For estimating the semantic information associated with tokens, Punjabi WordNet was used [22]. In this research work, synonym information associated with tokens was used because there is high probability that poet uses synonyms to convey their thoughts or feelings through poetry [2] and synonym information was proven to perform better as compared to other lexical relations on classification task in Indian language [24, 26, 27]. The idea behind using similarity measure is to capture the belief that there will be high overlap between words in B_t and related word found from Word Net semantic relations and examples. These extracted synonyms, followed by category it belongs to, were used as feature in poetry classification.

4. Methodology

An automatic poetry classification is a multistage process. Architecture of Punjabi poetry classifier is shown in Figure 1.

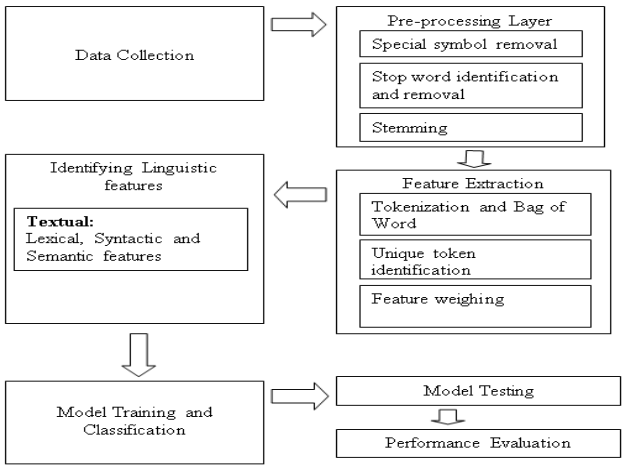

Figure 1. Architecture of Punjabi poetry classifier.

It consists of 5 layers: Data collection, Pre-processing, Feature Extraction, Identifying different linguistic features, Feature Selection, Model Training and classification, Model testing and Performance Evaluation.

4.1. Data Collection

First Step involved in Punjabi Poetry Classification was corpus building. As there is no publicly available poetry corpus in Punjabi, so the corpus of Punjabi poetry was populated manually. Total 2034 poetries were collected from various online sources such as http://www.punjabi-kavita.com/, http://www.punjabizm.com/,http://punjabimaaboli.com / [20, 21]. More than 100 categories were available for data collection, but to equalize the number of poetries in categories and to reduce the ambiguity in assigning poetries, 4 categories were framed and named as: Nature and Festival (NAFE), Linguistic and Patriotic (LIPA), Relation and Romantic (RORE) and Philosophy and Spiritual (PHSP). This entire poetry dataset is shown in Table 1 and it is present in Unicode format [28].

Table 1. Punjabi poetry dataset used.

S. No.	Category	Sub-Category	Count
1.	NAFE	Nature and Indian festival related poetries	505
2.	LIPA	Patriotic and language related poetries	399
3.	RORE	Love based and Relation based poetries	529
4.	PHSP	Philosophical and religious poetries	601

4.2. Pre-Processing Poetries

Pre-processing step involves passing poetries through various sub phases of text classification such as tokenization, stop word removal, noise removal sub phase and stemming. Various punctuation marks like: Comma (,), Dandi (।), Double Dandi (॥) , sign of interrogation (?), Sign of Exclamation (!) and Punjabi numerals are common occurring symbols in poetry. As there was no publicly available list of Stop Words in Punjabi language, manual collection of stop words was done. 256 unique stop words were identified from poetries as well as news articles [9, 12, 13]. This stop word list was used for stop word removal sub phase. Another important phase involved was stemming. In this research work, Punjabi tokens were extracted to their root form using stemming rules defined by Gupta [5].

4.3. Features Extraction

Feature extraction indicates extracting useful features from textual data. In this research work, unigram 'Bag of Word' (BOW) was created. This leads to Vector Space Document Model (VSDM) of poetries. For creating unigram BOW model, Poetries were tokenized and unique tokens were extracted from poetries.

From 2034 poetries, total 31938 tokens were extracted and used for building the model. These extracted tokens were passed to the different selected features based on poetry elements. Term frequency and term frequency-inverse document frequency was used to weigh the tokens.

4.4. Model Training and Classification

Initially, 10 different machine learning algorithms were experimented on 240 poetries with the objective to find the best machine learning algorithms. 10 different algorithms experimented were Adaboost, Bagging, C4.5, Decision Tree, Hyperpipes, K-nearest Neighbour, Naive Bayes, PART, SVM, Voting Feature Interval and ZeroR. Four algorithms NB, K-Nearest Neighbour (KNN), SVM and Hyperpipes (HP) [11, 14] performed better as compared to other algorithms. So, these four algorithms were selected for further experimentation on entire Punjabi poetry corpus using Weka [6].

4.5. Performance Evaluation

The next section presents the results of different algorithms for Punjabi poetry domain.

5. Results and Analysis

Selected four machine learning algorithms were trained and tested on 2034 poems. Results obtained using different linguistic features are presented in next subsection.

5.1. Lexical Feature

In Lexical feature, each token was considered as a feature and weight to it was assigned using Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). Performance of lexical feature with different weighting schemes (TF and TF-IDF) was shown in Table 2. As it can be observed from Table 2 that using TF weighing scheme, SVM (with accuracy of 72.04%) outperformed all other machine learning algorithm. Whereas, with

accuracy of 40.08%, KNN is the worst performer. Performance accuracy of KNN, HP and NB remains same on using TF-IDF weighing scheme. But, 6% decrease is observed in case of SVM.

Table 2. Performance of baseline classifiers using lexical feature.

Algorithm	Weighing Scheme							
	TF				TF-IDF			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
HP	**62.75**	**0.66**	**0.63**	**0.64**	**62.75**	**0.66**	**0.63**	**0.64**
KNN	**40.08**	**0.83**	**0.43**	**0.53**	**40.08**	**0.83**	**0.43**	**0.53**
NB	**59.69**	**0.62**	**0.63**	**0.62**	**59.69**	**0.62**	**0.63**	**0.62**
SVM	**72.04**	**0.73**	**0.72**	**0.72**	**66.43**	**0.75**	**0.70**	**0.72**

The confusion matrix for SVM classifier (with TF weighing scheme) and category-wise results are shown in Tables 3 and 4, respectively.

Table 3. Confusion matrix for SVM classifier using lexical feature.

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

It can be observed from the confusion matrix given in Table 3 that 432 poetries are correctly classified as PHSP poetries and 306 poetries are classified as LIPA poetries.

Table 4. Category-wise results for SVM classifier weighed using TF weighing.

	Accuracy	Precision	Recall	F-measure
NAFE	85.64	0.766	0.732	0.749
LIPA	90.80	0.843	0.767	0.803
RORE	81.67	0.680	0.701	0.691
PHSP	79.56	0.672	0.720	0.695

From Table 4, it can be observed that LIPA is the best classified category with accuracy of 90.80% and precision of 0.843. And PHSP is the worst classified category having accuracy 79.56% and precision of 0.672. PHSP and RORE were the most confusing categories. It was possible due to overlapping of poetic words in different categories.

From Figure 2, it can be observed that SVM is the better performer as compared to HP, NB and KNN using lexical features and TF weighing. Second best performer is HP, followed by NB and KNN.

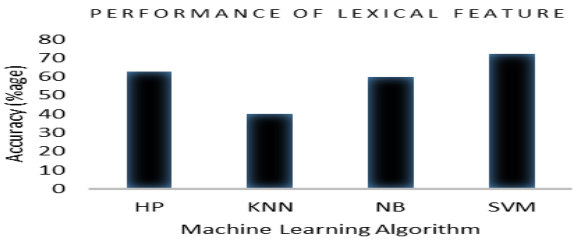

Figure 2. Performance of machine learning algorithms using lexical feature.

Rakshit *et al.* [23] had tried to build subject based poetry classifier that work on Bangla poetry, they had used different linguistic features (lexical, semantic, orthographic, stylometric) for poetry classification. Using lexical features weighed with TF-IDF, accuracy reported was 56.80 (%). So, it can be interpreted from the results that lexical features works well for building Punjabi poetry classifier as compared to Bangla poetry classifier.

5.2. Syntactic Feature

In syntactic feature, each token, followed by its Part Of Speech (POS) tag, was considered as a feature and weight to it was assigned using TF and TF-IDF. Total 32396 tokens, followed by its POS tags, were extracted from 2034 poetries. An increase in number of tokens can be observed as compared to lexical feature phase because in lexical feature, each token was considered irrespective of its POS tag whereas tokens were divided on the basis of their POS tags. Results of different machine learning algorithms using syntactic features weighed using TF and TF-IDF is presented in Table 5.

It can be observed from the Table 5, using TF as weighing scheme, maximum accuracy achieved by SVM classifier was 72.15%. Whereas, KNN was the poor performer. The same kind of performance pattern has been reported by using TF-IDF weighing scheme.

Table 5. Performance of baseline classifiers using syntactic feature.

Algorithm	Weighing Scheme							
	TF				TF-IDF			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
HP	64.72	0.38	0.67	0.49	64.72	0.66	0.67	0.66
KNN	46.16	0.66	0.92	0.66	46.16	0.38	0.92	0.49
NB	60.01	0.61	0.65	0.63	60.02	0.61	0.65	0.63
SVM	72.15	0.76	0.72	0.74	70.65	0.75	0.72	0.73

As it can be interpreted from Tables 2 and 5, addition of POS tags boosts the performance of machine learning algorithms. Category-wise results of SVM classifier (with TF weighing scheme) were depicted in Table 7 and confusion matrix was presented in Table 6.

Table 6. Confusion matrix for SVM classifier using syntactic feature.

	NAFE	LIPA	RORE	PHSP
NAFE	368	15	62	60
LIPA	16	300	17	66
RORE	54	14	364	97
PHSP	42	33	80	446

From confusion matrix shown in Table 6, it can be observed that with addition of POS tags, more poems were correctly classified in PHSP category (in

comparison with Table 3). Whereas, scenario was bit different for remaining three categories: NAFE, LIPA and RORE. 97 RORE poetries were misclassified as PHSP poetries and 80 PHSP poetries were classified as RORE poetries. So, there is maximum overlapping of words in RORE and PHSP categories. From the confusion matrix shown in Table 6, it can be observed that PHSP was the most confusing category of poems because poetries from different categories are often classified in PHSP category. Figure 3 shows the performance of different machine algorithms using syntactic features weighed with TF.

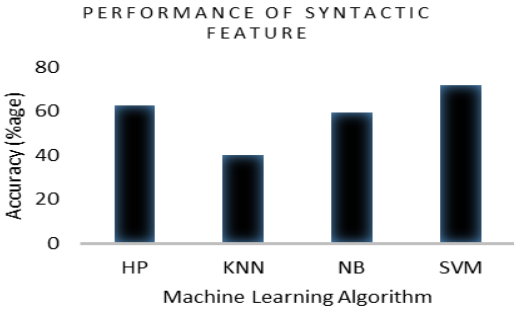

Figure 3. Performance of machine learning algorithms using syntactic feature.

From Figure 3, it can be observed that using syntactic feature and TF weighing scheme, SVM outperformed all other machine learning algorithms. HP is the second best performer, followed by NB and KNN.

Table 7. Category wise results for SVM classifier using TF weighing.

	Accuracy	Precision	Recall	F-measure
NAFE	85.58	0.768	0.729	0.748
LIPA	90.17	0.831	0.753	0.790
RORE	82.02	0.691	0.693	0.692
PHSP	79.63	0.672	0.739	0.704

Category-wise precision, recall and f-measure values are depicted in Table 7. Precision values for LIPA category was 0.831 and for PHSP category, precision reported was 0.672.

5.3. Semantic Feature

For integrating semantic information, Punjabi Word Net was used. Table 8 provides the result of different machine learning algorithms by incorporating synonym information of tokens. It can be observed from Tables 2, 5, and 8 that addition of semantic information of words leads to improvement in accuracy of machine learning algorithms. Inclusion of semantic information boosts the performance of poetry classification task.

Table 8. Performance of baseline classifiers using semantic feature.

Algorithm	Weighing Scheme							
	TF				TF-IDF			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
HP	64.77	0.65	0.64	0.64	62.52	0.65	0.64	0.52
KNN	45.97	0.90	0.38	0.52	40.16	0.82	0.44	0.64
NB	60.32	0.63	0.62	0.62	60.09	0.66	0.64	0.64
SVM	76.02	0.76	0.75	0.75	75.50	0.82	0.79	0.80

Confusion matrix of SVM (with TF weighing scheme) and category-wise results are shown in Tables 9 and 10, respectively.

Table 9. Confusion matrix for SVM classifier using semantic feature.

	NAFE	LIPA	RORE	PHSP
NAFE	372	16	51	66
LIPA	16	312	17	55
RORE	31	8	406	84
PHSP	43	33	70	455

It can be observed from Tables 2, 5, and 8 that addition of semantic information in classification enhances the accuracy of NAFE, LIPA, PHSP and RORE categories. With accuracy of 91.42 %, LIPA is the best classified category, followed by NAFE, RORE and PHSP. Whereas, RORE and PHSP are the most confusing categories because of overlapping of words used for writing both kind of poetries. Spiritual poetries generally consist of devotion towards religion and so word set for this category and PHSP category is often overlapping.

Table 10. Category-wise results of SVM classifier using TF weighing.

	Accuracy	Precision	Recall	F-measure
NAFE	87.38	0.805	0.737	0.769
LIPA	91.42	0.846	0.780	0.811
RORE	85.54	0.746	0.767	0.757
PHSP	81.48	0.689	0.757	0.722

Figure 4 shows the performance of different machine algorithms using semantic features weighed with TF.

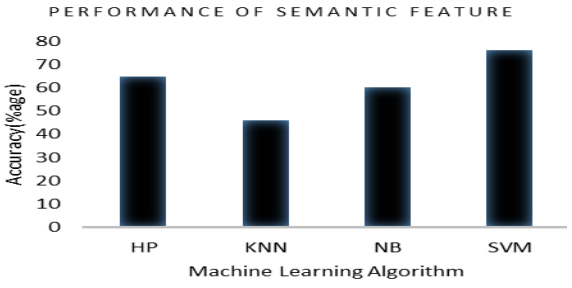

Figure 4. Performance of machine learning algorithms using semantic feature.

From Figure 4, it can be interpreted that SVM outperformed all other machine learning algorithm using semantic feature and TF weighing.

5.4. Performance of Machine Learning Algorithms Using Textual Features

Results of three different textual features are depicted in Figure 5.

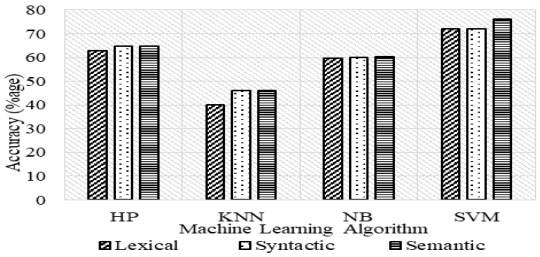

Figure 5. Performance of textual features.

It can be observed from Figure 5 that out of three different textual features, use of semantic features in Punjabi poetry classification task enhances the performance of machine learning algorithms. Out of four different machine learning algorithms, highest raise can be seen in SVM. Whereas, NB does not show any significant change in its performance using three different textual features. In case of HP and KNN, performance of syntactic feature of text and semantic features are almost same.

6. Conclusions

Computational linguistic analysis of Punjabi poetry is done to categorize poems based on its subject. For this experimentation, 4 different categories (NAFE, LIPA, RORE, PHSP) were manually populated with 2034 poems. These 2034 poems were passed through various text pre-processing phases: tokenization, stop word removal, special symbol removal, stemming. Tokenized unigram words extracted from these sub phases are used as vector space representation. Total 31398 extracted words were used for training and testing of 4 machine learning algorithm using weka toolset. These extracted words and different textual features were used for building classifiers. With TF weighing of all extracted tokens and using lexical feature, SVM outperformed all other machine learning algorithms with accuracy of 72.04 %. Whereas, performance of SVM was boosted with addition of syntactic and semantic information. With addition of POS tags, accuracy reported by SVM was 72.15 %. By incorporating Punjabi WordNet information, Accuracy reported by SVM is 76.02 %.

References

[1] Alsharif O., Alshamaa D., and Ghneim N., "Emotion Classification in Arabic Poetry using Machine Learning," *International Journal of Computer Application*, vol. 5, no. 16, pp. 10-15, 2013.

[2] Article Poetry Analysis accessed from https://en.wikipedia.org/wiki/Poetry_analysis, Last Visited, 2015.

[3] Barros L., Rodriguez P., and Ortigosa A., "Automatic Classification of Literature Pieces by Emotion Detection: A Study on Quevedo's Poetry," *in Proceedings of Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, pp. 141-146, 2013.

[4] Can E., Can F., Duygulu P., and Kalpakli M., "Automatic Categorization of Ottoman Literary Texts by Poet and Time Period," *Computer and Information Science-II*, pp. 51-57, 2012.

[5] Gupta V., "Automatic Stemming of Words for Punjabi Language," *Advances in Signal Processing and Intelligent Recognition Systems*, vol. 264, pp. 73-84, 2014.

[6] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I., "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10-18, 2009.

[7] Hamidi S., Razzazi F., and Ghaemmaghami M., "Automatic Meter Classification in Persian Poetries using Support Vector Machines," *in Proceedings of IEEE International Symposium on Signal Processing and Information Technology*, Ajman, pp. 563-567, 2009.

[8] Jamal N., Mohd M., and Noah S., "Poetry Classification Using Support Vector Machines," *Journal of Computer Science*, vol. 8, no. 9, pp. 1441-1446, 2009.

[9] Kaur J. and Saini J., "A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language," *International Journal of Data Mining and Emerging Technology, Indian Journals*, vol. 5, no. 2, pp. 114-120, 2015.

[10] Kaur J. and Saini J., "A Study and Analysis of Opinion Mining Research in Indo-Aryan, Dravidian and Tibeto-Burman Language Families," *International Journal of Data Mining and Emerging Technology*, vol. 4, no. 2, pp. 53-60, 2014.

[11] Kaur J. and Saini J., "Automatic Punjabi Poetry Classification Using Machine Learning Algorithms with Reduced Feature Set," *International Journal of Artificial Intelligence and Soft Computing*, vol. 5, no. 4, pp. 311-319, 2016.

[12] Kaur J. and Saini J., "POS Word Class based Categorization of Gurmukhi Stemmed Stop Words," *in Proceedings of 1st International Conference on Information and Communication Technology for Intelligent Systems*, Ahmedabad, pp. 3-10, 2015.

[13] Kaur J. and Saini J., "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronicle," *in Proceedings of ACM Symposium WIR'16*, Indore, pp. 32-37, 2016.
[14] Kaur J. and Saini J., "Punjabi Poetry Classification: The Test of 10 Machine Learning Algorithms," *in Proceedings of International Conference on Machine Learning and Computing*, Singapore, pp. 1-5, 2017.
[15] Kumar V. and Minz S., "Poem Classification using Machine Learning," *in Proceedings of International Conference on Soft Computing for Problem Solving*, Jaipur, pp. 675-682, 2012.
[16] Lou A., Inkpen D., and Tan C., "Multi-Category Subject-Based Classification of Poetry," *in Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference*, Florida, pp. 187-192, 2015.
[17] Punjabi language. from https://simple.wikipedia.org/wiki/Punjabi_language, Last Visited, 2015.
[18] Punjabi Part of Speech Tagger accessed from http://punjabipos.learnpunjabi.org/ Last Visited, 2015.
[19] Punjabi Poetry. Accessed from http://www.punjabi-kavita.com/, Last Visited, 2015.
[20] Punjabi Poetry. Accessed from http://www.punjabizm.com./ Last Visited, 2015.
[21] Punjabi Poetry, from http://punjabimaaboli.com/, Last Visited, 2015.
[22] Punjabi WordNet. Accessed from http://wordnet.thapar.edu/wordnetcms/public/wordnet/wordnet.php?langid=19&id=2, Last Visited, 2016.
[23] Rakhsit G., Ghosh A., Bhattacharyya P., and Haffari G., "Automated Analysis of Bangla Poetry for Classification and Poet Identification," *in Proceedings of 12th International Conference on Natural Language Processing*, Trivandrum, pp. 247-253, 2015.
[24] Sarmah J., Sahara N., and Sarma S., "A Novel Approach for Document Classification using Assamese WordNet," *in Proceedings of International Global Wordnet Conference*, Japan, pp. 324-329, 2012.
[25] Sebastiani F., "Machine Learning In Automated Text Categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1-47, 2002.
[26] Singh S. and Siddiqui T., "Utilizing Corpus Statistics for Hindi Word Sense Disambiguation," *The International Arab Journal of Information Technology*, vol. 12, no. 6A, pp. 755-763, 2015.
[27] Sinha M., Reddy M., and Bhattacharya P., "Hindi Word Sense Disambiguation," *in Proceedings of International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems*, Delhi, 2004.
[28] Unicode Table., from http://www.tamasoft.co.jp/en/general-info/unicode-decimal.html, Last Visited, 2015.

Jasleen Kaur had done Bachelor of Technology, Computer Science and Engineering from Guru Teg Bahadur Khalsa Institute of Engineering Technology, Malout, Punjab and Master of Technology (Computer Engineering) from Punjabi University, Patiala, Punjab. She had completed her PhD. from Uka Tarsadia University, Bardoli, Gujarat. She has published 15 papers in various International Journals and had more than 60 citations. She had publications with Indersicence Publishers, Springer and ACM digital Library.

Jatinderkumar Saini is Ph.D. from VNSGU, Surat. He secured First Rank in all three years of MCA and has been awarded Gold Medals forthis. Besides being University Topper, he is IBM Certified Database Associate (DB2) as well as IBM Certified Associate Developer (RAD). Associated with more than 50countries, he has been the Member of Program Committee for more than 50 International Conferences (including those by IEEE) and Editorial Board Member or Reviewer for more than 30 International Journals (including many those with Thomson Reuters Impact Factor). He has more than 55 research paper publications and nearly 20 presentations in reputed International and National Conferences and Journals. He is member of ISTE, IETE, ISG and CSI.