SUKHAN: Corpus of Hindi Shayaris annotated with Sentiment Polarity Information

Salil Aggarwal **Abhigyan Ghosh** **Radhika Mamidi**

Language Technologies Research Centre
KCIS, IIIT Hyderabad
Telangana, India
salil.aggarwal@research.iiit.ac.in
abhigyan.ghosh@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Shayari is a form of poetry mainly popular in the Indian subcontinent, in which the poet expresses his emotions and feelings in a very poetic manner. It is one of the best ways to express our thoughts and opinions. Therefore, it is of prime importance to have an annotated corpus of Hindi shayaris for the task of sentiment analysis. In this paper, we introduce SUKHAN, a dataset consisting of Hindi shayaris along with sentiment polarity labels. To the best of our knowledge, this is the first corpus of Hindi shayaris annotated with sentiment polarity information. This corpus contains a total of 733 Hindi shayaris of various genres. Also, this dataset is of utmost value as all the annotation is done manually by five annotators and this makes it a very rich dataset for training purposes. This annotated corpus is also used to build baseline sentiment classification models using machine learning techniques.

1 Introduction

Sentiment analysis is simply '*the task of extraction and analysis of subjective information present in some natural language data with the use of natural language processing*' [1]. But, the task of sentiment analysis becomes challenging for languages having annotated corpus only in some limited domains. One such language is Hindi and shayari is one of its domains which has no annotated dataset for the task of sentiment analysis. Shayari is a very rich tradition in South Asia. It has generally 2 to 4 lines which have some kind of deep meaning in them. It is mainly written in languages like Hindi, Urdu, Bangla, Nepali and Punjabi. Whether you are sad, alone, happy or in love, you can use shayari to express your feelings and thoughts. That's why, it is very important to have an annotated corpus of shayaris for the task of sentiment analysis. No such annotated corpus of Hindi shayaris currently exists in literature. SUKHAN is the first corpus of Hindi shayaris with annotated sentiment polarity information existing in literature as per our knowledge. It is written in Devanagari script and hence avoids the pre-processing cost of text normalization. We have also conducted various baseline experiments in order to compare the performance of various classifiers on the annotated corpus.

This paper is divided into 5 sections. Section 2 discusses related work in this area. Section 3 elaborates on the source and the creation of the corpus. Section 4 elaborates on the annotation process and the annotation scheme used for getting sentiment labels. Inter-annotator agreement has also been calculated and the details for interpretation of values for the Fleiss Kappa index have also been mentioned. Section 5 describes the experimental setup for training a model using various classifiers which helps in establishing the baseline for sentiment classification of these Hindi shayaris. All the results and conclusions using the annotated corpora have been briefly discussed in Section 6.

2 Related Work

So far, no work has been done to run sentiment analysis on Shayaris, neither in Hindi nor in any other Indian language, where similar constructs occur. However, work on sentiment analysis of Hindi poems has been done previously (Pal and Patel, 2020) but never on shayaris. Sentiment Analysis has also been done on Odia poems by Gaurav Mohanty and Mamidi (2018). Music classification has been carried out using lyrics (Hu et al., 2009), audio (Lu et al., 2005) and even multi-modal features (Laurier et al., 2009) for

[1] Source: Wikipedia

Proceedings of the 17th International Conference on Natural Language Processing, pages 228–233
Patna, India, December 18 - 21, 2020.

English. Similar work has been carried out for mood classification of Telugu (Abburi et al., 2016) and Hindi songs (Patra et al., 2016)

Nowadays, research mainly focuses on social media and very little attention is given to the traditional literature of which shayari is an important part. Automatic analysis of poetry is done for poems written in various languages like English, Chinese, Arabic, Malay, and Spanish. Barros et al. (2013) tried to categorize poems based on their emotional content. In the case of traditional literary works such as poetry, a lexicon creation methodology has been discussed for analyzing classical Chinese poetry (Hou and Frank, 2015). Hamidi et al. (2009) has also proposed a meter classification system for Persian poems based on features extracted from uttered poem. Alsharif et al. (2013) tried to classify Arabic poetry according to emotion associated with it.

3 Dataset

Due to the unavailability of annotated Hindi Shayari corpus with sentiment polarity information, the dataset was constructed manually. The advent of UTF-8 encoding has led to text in Indian scripts increasing day by day on the web. We collected shayaris from numerous online sources such as:

https://poetrytadka.com,

https://shayarifm.com/

https://shayarilovers.in/

https://bestnow.in/

These websites consist of many Hindi shayaris from various categories. We have used Hindi shayaris written using Devanagari script only. A total of 845 shayaris were mined. We have not used the title of shayari because the title of shayari usually does not have a strong association with the theme of the shayari. That's why, title was not used for extracting emotions from shayaris in order to avoid wrong results. The name of the shayar[2] also do not carry any sentiment information and therefore does not serve the task at hand and therefore is not used in baseline experiments. **Table 1** provides details on the initial statistics of the dataset before annotation.

[2]Person who writes shayaris

4 Annotation

4.1 Principles of Annotation

Three levels of granularity are described for existing methods of sentiment analysis. So, the task of sentiment analysis can be carried out at three different levels (Liu, 2012). On the basis of the level defined, the task is to identify if positive or negative sentiment is expressed at that level. It can be done at an aspect level (Hu and Liu, 2004), sentence level or at the level of the whole document (Turney, 2002). In the case of shayaris, it is possible that the different parts of the shayaris elicit different emotions. Since the task is to identify sentiment of the shayari as a whole, annotation is carried out only at an overall document level. The annotators were asked to go through the whole shayari before tagging them. This results in the tag corresponding to the polarity of the general mood evoked by it.

4.2 Annotation Process

We hired 5 annotators from different parts of India for the process of annotating the shayaris. These annotators were chosen from different regions in order to eliminate the chances of any kind of regional bias. These annotators were university students in the age group of 20–24 and were native Hindi speakers with sound background in linguistics and they speak and write in Hindi language on a daily basis. Each Hindi shayari was annotated by these 5 annotators. Each annotator was provided with the corpus and they had to annotate each and every shayari independently. They were not allowed to have any kind of communication with other annotators during the whole annotation process. Also, they were not given any kind of information regarding the shayar because there are some shayars who only write some particular type of shayaris which mostly evoke some particular kind of emotion like love, anger, hatred etc in the reader's mind. So revealing the name of shayar might result in some preconditioned bias which could have resulted in wrong annotation.

Shayaris are a very sophisticated form of language. At the same time, they can generate different kinds of thoughts and emotions in the mind of reader. So, a proper method is required for annotating the shayaris based on these emotions. Here, Russel's Circumplex Model (Russell and Pratt, 1980) serves as an appropriate reference

Sr. No.	Description	Initial Value
1	Initial shayari Count	845
2	Total number of words (tokens)	18978
3	Average number of words (tokens) per shayari	~ 23
4	Total number of unique words	2851

Table 1: Initial Dataset Statistics

for emotion identification. In this model, human emotions are plotted on a 2D plane of sentiment polarity and arousal as shown in Figure 1. For a given poem, the identified emotions were collected and based on these emotions, the sentiment polarity of the shayaris was decided. Initially, shayaris were classified into 5 categories:

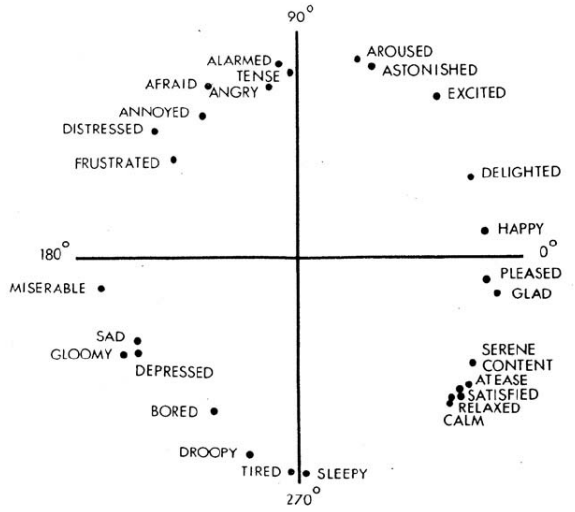

Figure 1: Russell's Circumplex Model of Affect

1. **Strongly Positive**: For the whole shayari, if the shayar is using only positive language such as expression of support, motivation, admiration, positive attitude, cheerfulness, forgiving nature, positive emotional state, etc, then the emotional states identified are tending to the positive side of Russell's model. These type of shayaris were classified as strongly positive.

2. **Strongly Negative**: For the whole shayari, if the shayar is using only negative language such as expression of hate, judgement, fear, anger, failure, criticism, negative attitude, etc. These types of shayaris were classified as strongly negative.

3. **Positive**: For the most of the shayari, if the shayar is using positive languages but also using negative language at some instants, then those types shayaris were classified as positive.

4. **Negative**: For the most of the shayari, if the shayar is using negative languages but also using positive language at some instances, then those shayaris were classified as negative.

5. **Neutral**: If the shayar is using both positive and negative language at equal intervals, then it is hard to tell what type of sentiment is present in the shayari. Those types of shayaris were classified as neutral.

Since the scope of this paper is to only determine a shayari as positive or negative, shayaris present in category 5 were discarded. Those present in Category 1 and 3 were annotated as positive and those present in Category 2 and 4 were annotated as negative. A total of 381 shayaris were annotated as positive whereas 352 shayaris were annotated as negative.

4.3 Inter-annotator Agreement

Inter-annotator agreement is a measure of how well the annotators can make the same annotation decision for the same category. Given the task in hand, it is fair to assume that annotation of the shayaris based on the emotions evoked by reading the lyrics is a very subjective opinion. Thus, inter-annotator agreement becomes an important factor in validating the annotators' work. The Fleiss' kappa obtained for the annotations for our dataset is 0.83. This corresponds to 'almost perfect agreement' according to the interpretation of Fleiss' kappa shown in **Table 2**.

5 Baseline for Sentiment Classification

In order to establish baseline results for the annotated corpus, a few experiments were conducted. The task was to classify Hindi shayaris as carrying positive or negative sentiment by training

Sr. No.	Range	Interpretation
1	$\leq$0	Poor agreement
2	0.01-0.20	Slight agreement
3	0.21-0.40	Fair agreement
4	0.41-0.60	Moderate agreement
5	0.61-0.80	Substantial agreement
6	0.81-1	Almost perfect agreement

Table 2: Fleiss Kappa values for inter-annotator agreement.

appropriate classification models. Initially three different classifiers were employed for this task and the results of each were compared. Term frequency-inverse document frequency (TF-IDF) (Jones, 1972) features were used to create a vector representation for an entire shayari. We also explored usage of character level n-grams as TF-IDF features to evaluate the performance of these classifiers.

5.1 Experimental Setup

The dataset was split into a ratio of 4:1 for the purpose of training and testing. For the baseline experiments, TF-IDF features for word n-grams and character n-grams were used for the task of sentiment classification. All the experiments were conducted using 'scikit-learn' (Pedregosa et al., 2011), an open source Python library[3]. Precision, Recall and F1-score are the three evaluation metrics which were calculated using **5-fold cross-validation**.

For baseline experiments Naive Bayes, Logistic Regression and Support Vector Machine (Cortes and Vapnik, 1995) were the classifiers used for baseline experiments. The 733 shayaris of the SUKHAN corpus were also passed through various pre-processing phases. Then using TF-IDF weights, vector representations were obtained for each shayari. TF-IDF is basically a technique to quantify a word in documents. We generally compute a weight to each word which signifies the importance of the word in the corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

[3]http://www.scikit-learn.org

6 Observations

TF-IDF was calculated for unigrams, uni-bigrams and uni-bi-trigrams. **Table 3** illustrates the results of the same for the three aforementioned classifiers. Even though 733 shayaris is a sizable corpus for the task in hand, it does not show a significant increase in accuracy especially with added bi-gram and tri-gram features. This is because most bi-grams and tri-grams occur sparsely in the entire corpus. Here, on an average, Linear-SVM performed better than all of the other classifiers. In order to tackle the problem of sparsity, we conducted experiments using n-grams at character level. For the baseline, 2-6 and 3-6 character n-grams were used to calculate character level TF-IDF features. The results of the same are illustrated in **Table 4**. On the basis of F1-score, Linear-SVM performed better than the other classifiers.

7 Conclusion

Shayaris and emotions share a very strong bonding. Each shayari is penned by a shayar with a lot of emotions, feelings and values. Different poetry elements such as diction, rhyme, rhythm, and imagery make shayari different from a normal piece of text. That's why In this research area, we have created a Hindi shayari corpus which would help to create an automatic system for categorization of shayaris based on polarity identification questionnaire and emotional states present in it. SUKHAN is the first corpus of Hindi shayaris of diverse themes, with shayaris, manually annotated as either having positive or negative sentiment values. The corpus was manually annotated with 733 Hindi shayaris scripted in the Devanagari script and based on the emotions they evoke. We also trained four different types of classifiers on our data. Classification models have been built using TF-IDF word-level features. Linear-SVM performed better as compared to other classifiers and the results of the experiments should serve as a good baseline.

Identifying the sentiment with the shayari is the first step towards identifying the emotions and thoughts which the shayari could evoke in the mind of the reader and this can further be used to build recommendation systems which are used by every major company in the e-commerce area.

Model	Features	Class	Precision	Recall	F1-Score
Linear-SVM	uni	Positive	0.852	**0.904**	0.876
		Negative	**0.907**	0.857	**0.880**
	uni-bi	Positive	**0.882**	0.859	0.870
		Negative	0.872	**0.896**	**0.884**
	uni-bi-tri	Positive	**0.888**	0.837	0.861
		Negative	0.856	**0.904**	**0.878**
Naive-Bayes	uni	Positive	**0.881**	0.850	0.864
		Negative	0.866	**0.896**	**0.880**
	uni-bi	Positive	0.869	0.871	0.870
		Negative	**0.882**	**0.880**	**0.881**
	uni-bi-tri	Positive	0.874	0.872	0.872
		Negative	**0.883**	**0.885**	**0.884**
Logistic regression	uni	Positive	**0.891**	0.819	0.853
		Negative	0.845	**0.909**	**0.875**
	uni-bi	Positive	**0.895**	0.810	0.850
		Negative	0.839	**0.914**	**0.875**
	uni-bi-tri	Positive	**0.898**	0.803	0.847
		Negative	0.834	**0.917**	**0.872**

Table 3: Sentiment Analysis with Word-Level TF-IDF Features

Model	Features	Class	Precision	Recall	F1-Score
Linear-SVM	(2-6) gram	Positive	0.873	0.864	0.868
		Negative	**0.876**	**0.886**	**0.880**
	(3-6) gram	Positive	**0.876**	0.862	0.868
		Negative	0.873	**0.888**	**0.880**
Naive-Bayes	(2-6) gram	Positive	**0.892**	0.808	0.845
		Negative	0.836	**0.901**	**0.869**
	(3-6) gram	Positive	**0.877**	0.819	0.844
		Negative	0.841	**0.893**	**0.865**
Logistic Regression	(2-6) gram	Positive	**0.882**	0.831	0.855
		Negative	0.851	**0.898**	**0.873**
	(3-6) gram	Positive	**0.880**	0.832	0.853
		Negative	0.852	**0.896**	**0.872**

Table 4: Sentiment analysis with Character-Level TF-IDF Features

References

Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth Gangashetti, and Radhika Mamidi. 2016. Multimodal sentiment analysis of telugu songs. In *SAAIP@ IJCAI*, pages 48–52.

Ouais Alsharif, Deema Alshamaa, and Nada Ghneim. 2013. Emotion classification in arabic poetry using machine learning. *International Journal of Computer Applications*, 65(16).

Linda Barros, Pilar Rodriguez, and Alvaro Ortigosa. 2013. Automatic classification of literature pieces by emotion detection: A study on quevedo's poetry. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 141–146. IEEE.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Pruthwik Mishra Gaurav Mohanty and Radhika Mamidi. 2018. Kabithaa: An annotated corpus of odia poems with sentiment polarity information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Saeid Hamidi, Farbod Razzazi, and Masoumeh P Ghaemmaghami. 2009. Automatic meter classification in persian poetries using support vector machines. In *2009 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 563–567. IEEE.

Yufang Hou and Anette Frank. 2015. Analyzing sentiment in classical chinese poetry. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 15–24.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760.

Yajie Hu, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*, pages 123–128.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. 2009. Music mood representations from social tags. In *ISMIR*, pages 381–386.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Lie Lu, Dan Liu, and Hong-Jiang Zhang. 2005. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18.

Kaushika Pal and Biraj V Patel. 2020. Model for classification of poems in hindi language based on ras. In *Smart Systems and IoT: Innovations in Computing*, pages 655–661. Springer.

Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. Multimodal mood classification framework for hindi songs. *Computación y Sistemas*, 20(3):515–526.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

James A Russell and Geraldine Pratt. 1980. A description of the affective quality attributed to environments. *Journal of personality and social psychology*, 38(2):311.

Peter D Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*.