BolLy: Annotation of Sentiment Polarity in Bollywood Lyrics Dataset

G. Drushti Apoorva and Radhika Mamidi(✉)

NLP-MT Lab, KCIS, IIIT-Hyderabad, Hyderabad, India
drushti.g@research.iiit.ac.in, radhika@iiit.ac.in

Abstract. This work presents a corpus of Bollywood song lyrics and its metadata, annotated with sentiment polarity. We call this *BolLy*. It contains lyrics of 1055 songs ranging from those composed in the year 1970 to the most recent ones. This dataset is of utmost value as all the annotation is done manually by three annotators and this makes it a very rich dataset for training purposes. In this work, we describe the creation and annotation process, content, and the possible uses of the dataset. As an experiment, we have built a basic classification system to identify the emotion polarity of the song based solely on the lyrics and this can be used as a baseline algorithm for the same. *BolLy* can also be used for studying code-mixing with respect to lyrics.

1 Introduction

Bollywood Hindi songs constitute a major part of music sales [25] in India and these are widely available in digitized form on the Internet. People choose to listen to different types of songs depending on their mood. This highlights the fact that classification of music on the basis of emotions they evoke [23] would help in varied tasks. Research in this field would most definitely require annotated datasets. Our work provides this resource.

In the dataset we present, the songs have been classified into two classes but for the definition of these classes, we have taken into consideration a wide variety of emotions. The size of *BolLy* is also substantially large as compared to other such datasets existing for Bollywood songs with their primary language being Hindi. It helps address the differences in Hindi and Western music appropriately [17]. A unique feature of this resource is that it is in the Devanagari script. This avoids the preprocessing cost of text normalisation. It contains the lyrics and metadata of 1055 Bollywood songs. This would be very useful for research in fields related to lyrics analysis and emotion polarity detection.

The dataset presented in this work has varied applications. One of the major applications is to extract emotion polarity from song lyrics which can be used in the creation of various systems such as automatic playlist generation and recommendation systems. They would also aid in music library management. 'BolLy' dataset contains Hindi as the major language and also has the usage of

K. Hasida and W. P. Pa (Eds.): PACLING 2017, CCIS 781, pp. 41–50, 2018.
https://doi.org/10.1007/978-981-10-8438-6_4

English or other Indian languages which is an emerging trend in the songs being composed. Hence, another application can be in the field of code mixing.

Our work is a step towards providing quality dataset for research in fields mentioned above. The purpose of this dataset is:

- to encourage research in the field of emotion extraction, particularly in lyrics.
- to provide a reference dataset for evaluating research.
- to help new researchers get started in the field of Hindi lyrics analysis.
- to provide a kickstart to studies on code mixing in Bollywood songs.

This paper is organised as follows: Sect. 2 elaborates the related work in this domain while Sect. 3 talks about the dataset, its creation and annotation. As mentioned earlier, there can be various applications of such a dataset. An experiment of the usage of this dataset for song classification based on emotion polarity, which can be used as a baseline for such studies, is described in Sect. 4. Section 5 discusses the further work possible on this dataset and has the conclusion of the work presented.

2 Related Work

One of the major contributions of the dataset proposed is in the field of music classification. Some of the important work done towards classification of music is mentioned. This is followed by discussion of related work specific to available datasets in this context or the taxonomies used for annotation.

The work done in music classification involves those making use of lyrics [6], audio [3,12,13,15] or a multimodal approach [2,9,10]. The classification techniques based on lyrics have been investigated in various languages. Affective lexicon and fuzzy clustering method were used for Chinese song lyrics in the work presented in [6]. In [5], it was shown that audio features do not always outperform lyric features using a dataset of 8,784 English song lyrics for the given task.

2.1 Lyrics Datasets

Since abundant resources are available for English song lyrics, we would restrict our discussion to datasets created specifically for Hindi and Indian languages. The dataset available for Bollywood song lyrics in [16] has a total of 461 song lyrics classified as positive or negative based on certain moods in each class.

The work presented in [20] uses another dataset for Bollywood song lyrics that contains lyrics and the metadata of 300 songs composed between the years 1995 and 2015 with details about the usage of foreign words in different decades. A Telugu dataset of 300 song lyrics has been used in the work towards a multimodal approach for sentiment analysis presented in [1]. Thus, to the best of our knowledge, the dataset presented by us, *BolLy*, is the largest of its kind amongst other resources for the language and the only one that contains instances of code-mixing.

2.2 Taxonomy for Classification

A proper taxonomy is vital to mood classification of songs. The Russell's Circumplex Model of 28 Affect Words [19] is one such two dimensional model that classifies moods using dimensions denoted as 'pleasant-unpleasant' and 'arousal-sleep'. It is seen that a lot of research groups have adapted Russell's model or used subsets of it for their work [7,22,26]. Our work makes use of this model to decide the scope of the tags used for annotation. This is discussed in detail in Sect. 3.2 (2) that deals with annotation guidelines.

3 The Dataset

3.1 Creation of the Dataset

Bollywood lyrics can be easily extracted from many websites. They are commonly available transliterated in the Roman script. For simplicity of processing, we extracted them from a source where Hindi words were available in Devanagari script, i.e. consisting of utf-8 characters. The lyrics of the songs were extracted with metadata including the movie or album they belong to, the singers and the year of release.

After the initial collection of data, it was cleaned to further reduce preprocessing required. Repetitions of lines or words in the lyrics were represented using numbers, for example, a line was followed by 'X2' if it was to be repeated twice. In certain cases, these numbers were in Devanagari. All such representations were removed and the line or word in question were copied as many times mentioned.

In Bollywood, a lot of songs are remixed into different versions. Sometimes the same songs are sung by different singers. There are quite a few instances wherein the same song occurs in different moods, such as both happy and sad, in a movie. All such songs appeared multiple times in the dataset. These songs differ in terms of audio features and evoke varied emotions when listened to. As there was no difference in their lyrics and our work focuses solely on the emotions evoked by song lyrics, the multiple occurrences of such songs were removed from the dataset, and only one file was retained.

Following are the features of the dataset:

- originally procured 1,082 song lyrics.
- 1,055 song lyrics in the final dataset, after removal of duplicates.
- song lyrics/files with metadata amounting to 2.6 MB data.
- 712 songs annotated as positive in the final dataset.
- 343 songs annotated as negative in the final dataset.
- total number of tokens in the dataset are 2,17,285.
- average number of tokens in a song are rounded off to 211.
- total number of tokens in positive songs are 1,51,362.
- average number of tokens in a positive song can be rounded off to 218.
- total number of tokens in negative songs are 65,923.
- average number of tokens in a negative song can be rounded off to 196.

3.2 Annotation

Principles of Annotation. Three levels of granularity are described for existing methods of sentiment analysis in [11]. On the basis of the level defined, the task is to identify if positive or negative sentiment is expressed at that level. They can be carried out at the level of the whole document as in [24], or at sentence level or at the level of entities and aspects [4,21]. It is possible that different smaller parts of a song's lyrics evoke different emotions. We aim to identify whether the whole song's lyrics evoke a positive or a negative emotion. Hence, it is best for us to look at document level classification. The annotators were asked to go through the whole song before tagging them. This results in the tag corresponding to the polarity of the general mood evoked by it.

Annotation Process. Each song in the dataset was annotated as positive or negative by three annotators, all of whom were university students in the age group 20–24 and were native speakers of Hindi. Songs evoke a certain emotion or mood, and these can be classified as those with positive or negative valence according to Russell's Circumplex Model, as shown in Fig. 1. The songs that evoke emotions ranging from 'aroused' to 'sleepy' including 'calm', 'satisfied', 'delighted', 'excited', etc. are to be tagged as positive. Negative tags are to be given to songs evoking moods such as 'angry', 'annoyed', 'miserable', 'depressed', etc., all of them spanning from 'alarmed' to 'tired'. Each song is annotated with the tag given by majority of the annotators for it.

The annotations were carried out in a controlled environment in which the annotators were not allowed to listen to the audio of the song presented. Hence, the annotation is solely on the basis of lyrics. Also, the songs were presented to the annotators without any of the metadata associated with it to prevent any preconditioning. The number of positive and negative tags given by each annotator can be seen in Table 1.

Table 1. Number of positive and negative tags given by each annotator.

Annotator	Positive tags	Negative tags
1	721	334
2	728	327
3	710	345

By the end of annotation of the final dataset of 1055 songs, 712 are tagged as positive while the rest 343 are annotated as negative. From the original dataset of 1,082 songs, 27 were removed as they were duplicates. The annotation for these 27 songs were used to check for consistency of the individual annotators. While annotators 1 and 2 had annotated the duplicate instances of only 1 song out of 27 differently, the third annotator's tagging was inconsistent for 2 songs. These small numbers can be ignored in such a large dataset as they show that the annotators were rarely inconsistent in their task.

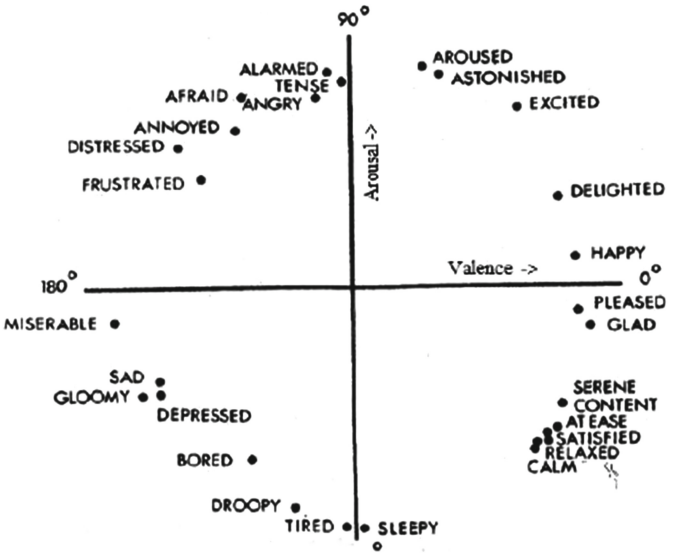

Fig. 1. Russell's Circumplex Model [19] classifying 28 affect words on the basis of positive and negative valence and arousal.

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 songs 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.

There are many statistical measures that can be used for measuring the reliability of agreement between the annotators given the number of data points, number of annotators and the number of tags they can be given. Cohen's kappa, Fleiss' kappa and Scott's pi are some of them. Out of these, Cohen's kappa and Scott's pi work only for cases where there are two annotators. Fleiss' kappa, a generalisation of Scott's pi statistic, is useful when there are more annotators and thus is most suited for our work.

The Fleiss' kappa obtained for the annotations for our dataset is 0.79. This corresponds to 'substantial agreement' [8] according to the interpretation of Fleiss' kappa shown in Table 2.

4 Experimentation

4.1 Theory

We conducted a few experiments to show one of the applications of the dataset. This involved classification of the song lyrics to extract sentiment polarity expressed by them. This can be used as a baseline experiment for sentiment

Table 2. Interpretation of Fleiss' kappa values for inter-annotator agreement [8].

κ	Interpretation
<0	Poor agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

analysis tasks for Hindi song lyrics. We make use of Naive Bayes [14] and Support Vector Machine classifiers for these experiments. These are discussed here briefly.

Naive Bayes methods are supervised learning algorithms based on the Bayes' Theorem with the assumption that every pairs of features are independent. This leads to using the classification rule $P(y|x_1...x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$, given a class variable y and dependent feature vector x_1 through x_n. There are different variations of this depending on the distribution of the data points.

Another supervised learning algorithm is SVM or Support Vector Machine. This makes use of a hyperplane to identify the decision boundary. Given a dataset of n points of the form $(x_1, y_1)....(x_n, y_n)$ where y_i is either 1 or -1 depending on which class the data point x_i belongs to. The hyperplane can be defined as the set of points x satisfying $w.x - b = 0$ where w is the normal vector to the hyperplane and parameter $\frac{b}{||w||}$ expresses the offset of this hyperplane along w from the origin.

4.2 Methodology

The song lyrics were available as files which were converted into separate lists of lyrics and their corresponding tags for positive and negative songs, called *PLyr* and *PTags* for positive songs and *NLyr* and *NTags* for negative songs, respectively. For splitting the dataset into training and testing set, we create lists *TrainLyr* and *TestLyr* with their respective tags in the lists *TrainTags* and *TestTags*. This split is in the ratio 9:1. *TrainLyr* and *TrainTags* are used to train the models for Multinomial Naive Bayes (MultinomialNB), Bernoulli's Naive Bayes (BernoulliNB) and Support Vector Machine (SVM). The latter is run over 5 folds. These trained models are used to predict the sentiment polarity of the data points in *TestLyr* and the predicted tags are stored in *PredTags*, which are then compared with *TestTags* for evaluation.

4.3 Experiments Conducted

The method explained above was conducted on the *BolLy* dataset. The dataset was split in the ratio of 9:1 for training and testing purposes. As the class is

imbalanced and there are more number of positively tagged songs, we set the class prior according to the class distribution. Without this, the classifier would have tagged all the test data points as positive as that would give a better result but that is not our aim. Specifying the class priors helps the classifier to take into account the features given and predict the classes for test data accordingly.

These experiments were carried out using an open source Python library, 'scikit-learn' [18]. The features used included bag of words, term frequency and term frequency-inverse document frequency which were all extracted using the modules from this library. The evaluation metrics such as accuracy, precision, recall and F1 measure were extracted. The SVM classifier was run for 10 folds over the dataset.

4.4 Results and Discussion

The average accuracies for the three classifiers with the data split in the ratio 9:1 are shown in the Table 3. Subsequent tables show the other evaluation metrics for each of the classifiers used (Table 4).

Table 3. Accuracies obtained for the classifiers

Classifier used	Accuracy (%)
MultinomialNB	69.61
BernoulliNB	71.57
SVM	75.49

Table 4. MultinomialNB classifier scores

Classifier	Class	Precision	Recall	F1 score (%)
MultinomialNB	Positive	**0.80**	0.62	**0.70**
	Negative	0.46	**0.67**	0.54
BernoulliNB	Positive	**0.80**	**0.77**	**0.79**
	Negative	0.56	0.61	0.58
SVM	Positive	0.74	**0.99**	**0.84**
	Negative	**0.90**	0.27	0.42

These results show that Support Vector Machine works the best for this experiment. These experiments can be used as a baseline as the features used are very basic and not tuned specific to the dataset or the classification problem. Incorporating these would definitely give better results which is why the results shown here would work well as a baseline. Precision can be looked at as a classifier's exactness while recall would give a measure of its completeness. If we look at all the evaluation metrics, we see that barring two cases, the models give better precision and recall for the positive class.

5 Future Work and Conclusion

Detection and classification of sentiments and opinion mining is a challenging task from the point of view of computational linguistics. This is because opinions are not realised uniformly in the syntax, semantics and the pragmatics of language. Even more difficult is this task in song lyrics as there is no fixed grammar or structure that is followed. Our work is a contribution towards encouraging work in this direction.

As mentioned, a variety of applications are possible once the sentiment or mood of songs are identified. Some of them include automatic playlist generation, digital music library management, song suggestion systems, etc. The dataset presented, *BolLy* would indeed thrust forward the studies conducted in the related fields and prove to be better than existing resources in Hindi.

The experiment shown is just a basic implementation of an opinion extraction task and gives a good baseline for specific and fine-tuned implementations for the same. It gives an insight as to how imbalanced datasets can be handled and what classification algorithms would serve our purpose better.

The insights derived would be helpful to build a better and more accurate system for opinion extraction from songs, using existing resources like SentiWordNet, subjectivity lexicon, language identifiers to identify and extract better features. Another possible work would be to assess the dynamic change of mood over the period of song and compare it with the general emotion polarity of the whole song.

Further, this dataset can be annotated and used for code-mixing tasks as well. It can also be used for cueing in song generation systems. This dataset once used for training a specific model, would be critical in dynamic collection of song lyrics and their annotation, thus making available a substantially large, rich corpus for similar tasks. This would help creating an invaluable resource.

References

1. Abburi, H., Akkireddy, E.S.A., Gangashetty, S.V., Mamidi, R.: Multimodal sentiment analysis of Telugu songs. In: Proceedings of the 4th Workshop on Sentiment Analysis Where AI Meets Psychology (SAAIP 2016), pp. 48–52 (2016)
2. Bischoff, K., Firan, C.S., Paiu, R., Nejdl, W., Laurier, C., Sordo, M.: Music mood and theme classification-a hybrid approach. In: ISMIR (2009)
3. Fu, Z., Lu, G., Ting, K.M., Zhang, D.: A survey of audio-based music classification and annotation. IEEE Trans. Multimedia **13**, 303–319 (2011)
4. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177. ACM (2004)
5. Hu, X., Stephen Downie, J., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX audio mood classification task: lessons learned. In: Proceedings of the 9th International Conference on Music Information Retrieval (2008)
6. Hu, Y., Chen, X., Yang, D.: Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: ISMIR (2009)

7. Katayose, H., Imai, M., Inokuchi, M.: Sentiment extraction in music. In: 9th International Conference on Pattern Recognition (1988)
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
9. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Seventh International Conference on Machine Learning and Applications, ICMLA 2008, pp. 688–693. IEEE (2008)
10. Laurier, C., Sordo, M., Herrera, P.: Mood cloud 2.0: music mood browsing based on social networks. In: Proceedings of the 10th International Society for Music Information Conference (ISMIR 2009), Kobe, Japan (2009)
11. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. **5**(1), 1–167 (2012)
12. Lu, L., Liu, D., Zhang, H.-J.: Automatic mood detection and tracking of music audio signals. Trans. Audio Speech Lang. Process. **14**, 5–18 (2006)
13. Lie, L., Liu, D., Zhang, H.-J.: Automatic mood detection and tracking of music audio signals. IEEE Trans. Audio Speech Lang. Process. **14**(1), 5–18 (2006)
14. McCallum, A., Nigam, K., et al.: A comparison of event models for Naive Bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48. Citeseer (1998)
15. Patra, B.G., Das, D., Bandyopadhyay, S.: Automatic music mood classification of Hindi songs. In: Proceedings of the 3rd Workshop on Sentiment Analysis Where AI Meets Psychology (2013)
16. Patra, B.G., Das, D., Bandyopadhyay, S.: Mood classification of Hindi songs based on lyrics. In: Proceedings of the 12th International Conference on Natural Language Processing (ICON 2015) (2015)
17. Patra, B.G., Das, D., Bandyopadhyay, S.: Multimodal mood classification - a case study of differences in Hindi and western songs. In: COLING (2016)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
19. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**, 1161–1178 (1980)
20. Shakoor, A.A., Sahebdin, W.B., Pudaruth, S.: Exploring the evolutionary change in bollywood lyrics over the last two decades. In: The Second International Conference on Data Mining, Internet Computing, and Big Data (BigData 2015), p. 46 (2015)
21. Szabó, M.K., Vincze, V., Simkó, K.I., Varga, V., Hangya, V.: A Hungarian sentiment corpus manually annotated at aspect level (2016)
22. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press, New York (1989)
23. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: ISMIR, vol. 8, pp. 325–330 (2008)
24. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics (2002)

25. Ujlambkar, A.M., Attar, V.Z.: Automatic mood classification model for Indian popular music. In: 2012 Sixth Asia Modelling Symposium (AMS), pp. 7–12. IEEE (2012)
26. Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I.-B., Ho, Y.-C., Chen, H.H.: Toward multi-modal music emotion classification. In: Huang, Y.-M.R., Xu, C., Cheng, K.-S., Yang, J.-F.K., Swamy, M.N.S., Li, S., Ding, J.-W. (eds.) PCM 2008. LNCS, vol. 5353, pp. 70–79. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89796-5_8