BHAAV (भाव) - A Text Corpus for Emotion Analysis from Hindi Stories

Anonymous NAACL submission

Abstract

In this paper, we introduce a new Hindi text corpus (BHAAV - which means 'emotions' in Hindi) for analyzing emotions that a writer expresses through his characters in a story, as perceived by a narrator/reader. The corpus consists of 20,304 sentences collected from 230 different short stories spanning 18 genres. Each sentence has been annotated by three native Hindi speakers with at least ten years of formal education in Hindi, with the goal of identifying one of the five different emotions - *anger*, *joy*, *suspense*, *sad*, and *neutral*. To our knowledge, this is the first and largest annotated text corpus for emotion analysis of stories in a low-resource language like Hindi. This paper discusses the scope of the corpus and its possible uses. We also provide a detailed analysis of the dataset and train baseline classifiers reporting their performances. The dataset will be made publicly available.

1 Introduction

Emotion analysis from text is the study of identifying, classifying and analyzing emotions (e.g., *joy*, *sadness*) as expressed and reflected in a piece of given text (Yadollahi et al., 2017). It's wide range of applications in areas such as - *customer relation management* (Bougie et al., 2003), *dialogue systems* (Ravaja et al., 2006), *intelligent tutoring systems* (Litman and Forbes-Riley, 2004), *analyzing human communications* (Kövecses, 2003), *natural text-to-speech systems* (Francisco and Gervás, 2006), *assistive robots* (Breazeal and Brooks, 2005), *product analysis* (Knautz et al., 2010), and *studying psychology from social media* (De Choudhury et al., 2013), has drawn considerable attention from the scientific community making it one of the important areas of research in computational linguistics.

Almost all the methods and resources developed in this domain deals with English language (Yadollahi et al., 2017), making our understanding of expression of emotions only limited to English text. This paper describes our attempt to develop a text corpus for emotion analysis from stories written in Hindi, which is one of the 22 official languages of India and is among the top five most widely spoken languages in the world[1]. Despite its wide usage, there are no text based resources for emotion analysis in Hindi, making the resource shared in this work as the first and largest annotated corpus for studying emotions from Hindi text, and facilitating development of linguistic resources in low-resource languages.

According to a joint report by KPMG and Google[2] published in 2017, there are 234 million Internet users in India using one of the Indian languages as their medium of communication against 175 million users using English. This gap is predicted to increase by 2021, with users using Indian languages reaching 536 million. Among the 122 languages officially recognized by Indian constitution, Hindi is the most spoken, followed by Bengali and Telugu. This has also led to a sudden rise in interest from social media companies like Facebook, and Internet search companies like Google to increase their support for popularly used Indian languages, making our work in this domain apt and timely.

Related to the task of emotion analysis

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

in Hindi, previous attempts have been made in developing corpus for predicting emotions from Hindi-English code switched language used in social media (Vijay et al., 2018) (2,866 sentences) and from auditory speech signals (Koolagudi et al., 2011). Some work has been undertaken in a closely related task of sentiment analysis and datasets have been created for identifying sentiments expressed in movie reviews (Mittal et al., 2013) (664 reviews), Hindi blogs (Arora, 2013), and generating lexical resources like Hindi Senti-Wordnet (Joshi et al., 2010). Given the dearth of resources for analyzing emotions from Hindi text, we take this opportunity to present and publicly share 'BHAAV' - a corpus of 20,304 sentences collected from 230 different short stories spanning across 18 genres, written in Hindi. Each sentence has been annotated by three native Hindi speakers with at least ten years of formal education in Hindi, with the goal of identifying one of the five different emotions - *anger*, *joy*, *suspense*, *sad*, and *neutral*.

Stories are melting pot of different types of emotions expressed by the author through the characters and plots that he develops in his writing. Emotions in storytelling has been previously studied resulting in identification of six basic types of emotional arcs in English stories (Reagan et al., 2016). This motivated us to develop BHAAV from Hindi stories. We believe that apart from studying emotions in Hindi text, the presented corpus would also enable studies related to the analysis of Hindi literature from the perspective of identifying the inherent emotional arcs. It also has the potential to catalyze research related to human text-to-speech systems geared towards improving automated storytelling experiences. We keep the order of the sentences intact as they occur in their source story. This makes the corpus ideal for performing temporal analysis of emotions in the stories, and provides enough information for training machine learning models that takes into account temporal context.

Major contributions of our work are:

- *Publicly share the first and the largest annotated corpus of 20,304 sentences in Hindi (BHAAV), collected from 230 different short stories spanning across 18 genres, identifying one of the five different emotions - anger, joy, suspense, sad, and neutral.*
- *Describe potential applications of BHAAV corpus, the process of annotation and main challenges in creating an emotion analysis text corpus in a low-resource language like Hindi.*
- *Report performances of baseline classifiers trained for identifying emotion expressed in a sentence of a story written in Hindi.*

2 Related Work

In this section, we present works relevant to that of ours. It is necessary to mention that there has been extensive work in sentiment analysis especially in the past two decades. Although, there is significant intersection between techniques used for sentiment analysis and emotion analysis, yet the two are different in many ways. Emotion analysis is often tackled at a fine grained level and has historically proved to be more challenging due to subtleties involved in identifying and defining emotions that are orthogonal to each other. Additionally, resources for emotion analysis are scarce when compared to sentiment analysis. For a detailed survey of methods, datasets and theoretical foundations on sentiment analysis and emotion analysis, please refer (Yadollahi et al., 2017; Lei et al., 2018; Cambria et al., 2017; Poria et al., 2017)

Analyzing emotions from text has been primarily manifested through four different types of tasks - *Emotion Detection* (Gupta et al., 2013), *Emotion Polarity Classification* (Alm et al., 2005), *Emotion Classification* (Yang et al., 2007), and *Emotion Cause Detection* (Gao et al., 2015). The scope of this work is limited to the task of *Emotion Classification.* (Pang et al., 2008) mentions that emotions are expressed at four levels - *morphological*, *lexical*, *syntactic* and *figurative*, and noted that as we move from morphological to figurative, the difficulty of the emotion analysis task increases and number of resources for the same decrease. Developed from stories written in a morphologically rich language, BHAAV primarily deals with the first and the last levels.

Most of the work for creating data resources for emotion analysis has been fairly limited to building emotion lexicons (Strapparava et al., 2004; Pennebaker et al., 2001; Shahraki and Zaiane, 2017; Mohammad and Turney, 2013), or concentrated in annotating emotions of individual sentences without giving any context (Strapparava and Mihalcea, 2007). This approach, as indicated by many, is a non-holistic approach for a task such as emotion analysis (Schwarz-Friesel, 2015; Ortony et al., 1987). BHAAV not only presents annotated sentences, but also provides their context.

Lastly, when it comes to the task of analyzing emotions from text, there are no datasets available in Hindi. Although, resource-poor Indian languages have started catching up their richer counterparts in the domain of sentiment analysis (Mittal et al., 2013; Arora, 2013; Joshi et al., 2010), yet sufficient work needs to be done considering the pace at which these languages are finding their uses in modern digitally driven India. The lack of resources can be judged from the 'wide' usage of one of the very few Hindi datasets for sentiment analysis task (Balamurali et al., 2012). It consists of just 200 positive and negative sentences for each of the two major Indian languages, Hindi and Marathi. Another popular and a recent attempt is by (Patra et al., 2015). They released a dataset containing approximately 1500 tweets for each of the languages of Hindi, Bengali and Tamil for Aspect Based Sentiment Analysis. BHAAV is certainly an attempt to fill this gap and create a large, effective and high quality resource for emotion mining from text.

3 Language Specific Challenges

As already mentioned and pointed in (Yadollahi et al., 2017), the computational methods used in the tasks pertaining to sentiment analysis can readily be applied to the emotion analysis tasks. Therefore, the challenges for emotion analysis from text are very similar to that of the domain of sentiment analysis from text. For a detailed description of the challenges one can refer (Mohammad, 2017). However, our task of identifying emotions from sentences poses additional challenges due to the inherent characteristics of Hindi language. We point out some of these language specific challenges as identified by (Arora, 2013), in order to draw a complete picture of the intricacies of the task and emphasize on the fact that there is a scope of developing methods specific to Hindi, and not all methods developed for English can be directly translated to Hindi.

Word Order - The order in which words appear in a sentence plays an important role for determining polarity as well as subjectivity of the text. As opposed to English, which is a *fixed order language*, Hindi is a *free order language*. For any sentence in English to be grammatically correct the 'subject' (S) is followed by 'verb' (V), which is followed by 'object' (O) - [SVO]. For example the English sentence - *Ram ate three mangoes*, which follows the [SVO] pattern, can be expressed in three ways in Hindi that do not adhere to the [SVO] pattern - 'राम ने तीन आम खाया' [SVO], 'तीन आम खाया राम ने' [OVS], and 'खाये तीन आम राम ने' [VOS]. This lack of order can pose challenges to the machine learning algorithms that take into account the order of the words.

Morphological Variations - Hindi language is morphologically rich. This means that a lot more information can be expressed in a word in Hindi for which one might end up writing many more words in English. One of the example is that of expressing genders. For example, when using the word 'खायेगी', which means 'will eat' in English, one can not only indicate that someone will eat but also provide cues of the person's gender (in this case female - the male variant is 'खायेगा').

Handling Spelling Variations - A word with the same meaning can appear with multiple spelling variations. Occurrence of such variations can pose challenges for the machine learning models that has to take into account all the spelling variants. For example the word 'मेहेंगा', which means 'costly' has another variant महंगा that means the same.

Lack of Resources - The lack of lexicons, developed techniques and elaborate resources in Hindi also adds to the challenge, which is also one of the main motivations behind this work.

4 Corpus Generation and Annotation

One of our primary aims was to create a manually annotated large corpus for performing emotion analysis from text in Hindi. We also wanted to capture the context in which a given piece of text occurs. Therefore, we decided to

Dataset/Genre	Fleiss's Kappa	Krippendorff's alpha
BHAAV dataset	0.80241357	0.802416004636
आदर्शवादी (Idealist)	0.741594061693	0.740462670362
प्रेमपरक (Romantic)	0.906633895329	0.90659267822
शहरी जीवन (Urban Life)	0.825558745183	0.824972237422
शोषक और शोषित वर्ग (Exploiter and Exploited Class)	0.798745458224	0.797991236844
नीतिपरक (Moral Stories)	0.865927372214	0.865857301713
किसान जीवन (Life of a Farmer)	0.885919097027	0.886009070648
ऐतिहासिक (Historical)	0.921326884658	0.921310220449
प्रेरणादायक (Inspirational)	0.897278304968	0.897266011499
देश भक्ति संभंधित (Patriotic)	0.991438728869	0.991446652425
व्यक्तिगत जीवन की समस्या (Personal Issues/Problems)	0.879354980254	0.879386098588
रूढ़ि और अंधविश्वास (Dogmatic and Superstitious)	1.0	1.0
संयुक्त परिवार की समस्या (Joint Family Problems)	0.887607633305	0.887631947099
रहस्यमयी (Mystery)	0.908185638589	0.908330457028
यथार्थवादी (Realistic and Pragmatic)	0.912201981538	0.912237344469
ग्रामीण जीवन (Village Life)	0.967427620692	0.967434538508
उपदेशपरक (Instructive)	0.919704200254	0.919700170635
भोगे हुए यथार्थ की कहानी (Real Stories)	0.911361226137	0.911352174976
समाज सुधारक (Society and its Reformation)	0.878218452096	0.878149647043

Table 1: Inter-annotator Agreements as measured using Fleiss‘s Kappa (Fleiss and Cohen, 1973) and Krippendorff‘s alpha (Krippendorff, 2011) for the entire BHAAV dataset and for each genre.

extract all the sentences from short stories belonging to genres popular in Hindi. We started with 30 genres, but narrowed down to only 18, depending on the availability of online content. Throughout the process of deciding on genres and finding online content relevant to them, we took help from an expert in Hindi literature, who provided 500 online URLs containing a popular short story belonging to one of the genres. Table 1, provides a list of genres available in our final dataset. Whenever possible we also searched for an audio book[3] where the same story has been narrated by a narrator. This was done in order to help the annotators during the annotation process, in case they have to refer to examples of how a narrator/reader would express the emotion of a sentence in the context of the story. All our annotators were native Hindi speaking volunteers who had a minimum of 10 years of formal education in Hindi, and showed great interest in reading the stories.

Emotion	Sample Sentences
joy	बादशाह ने कहा तुम्हारी कहानी पहली दोनों से अधिक मनोरंजक है (The king said that your story is more entertaining than the previous two stories)
anger	रुपिया नई देगा तो उसका खाल उतारकर बाजार में बेच देगा (If he doesn't gives the money then I will take out his skin and sell it in the market)
suspense	मजदूर ने अब तक तो झलक भर देखी थी अब तो उसे पूरी नजर भर देखा तो ठगा सा खड़ा रह गया (Till now the worker had only seen his glimpses, but when he saw him fully he was just stunned)
sad	उसने रुँआसे होते हुए मम्मी की ओर देखा (With teary eyes he saw towards his mother)
neutral	मैं इसकी मां हूं (I am his mother)

Table 2: Sample sentences from BHAAV dataset for each emotion label.

All the URLs were scraped and the text was extracted from them. Not all of them could be retrieved. We ended up retrieving and extracting text from 230 stories. The extracted text was split into sentences in an automated way and contained many unnecessary text that were not a part of the story. During the annotation process the annotators filtered the unwanted text and only annotated the relevant portion. Whenever the sentences were not correctly split, the annotators also corrected them. A total of 5 annotators were used for annotating the entire corpus, such that each sentence gets at-least 3 annotations. During the annotation process the annotators had access to the main URL of the story and the list of audio books. Each story was annotated in one sitting and it took 9 months for finishing the process.

The guidelines for annotating emotions was designed to be very short and concise with regards to the definitions of the categories to be assigned. All the definitions were kept brief and aligned with (Plutchik, 1984), along with sample annotated sentences from the domain. We asked the annotators to identify only one of the five emotions expressed in a sentence of a story - *anger*, *joy*, *suspense*, *sad*, and *neutral*. We went with the above categories of emotions mainly due to their extensive use in other works and also included *suspense* as we were dealing with the domain of stories in which *suspense* is often a popular emotion infused by the authors in creating interesting plots. The annotators were instructed not to be biased by their own emotions towards a statement in the story while labeling them, and was asked to identify only the emotion that an unbiased narrator/reader of that story would like to express while reading it to someone. Whenever confused, they were asked to refer the audio book of the story if available, or one of the authors, or mark it as *neutral* if that doesn't clear the confusion.

Emotion	No. of Sentences	No. of Sentences (Train data)	No. of Sentences (Test data)
joy	2,463	2,242	221
anger	1,464	1,321	143
suspense	1,512	1,389	123
sad	3,168	2,843	325
neutral	11,697	10,478	1,219

Table 3: Distribution of sentences in different categories of emotions in the BHAAV dataset.

General statistics of the dataset is presented in Table 3. The overall inter-annotator agreements and the agreements for individual genres are presented in Table 1. Some samples

[3] Example of audio books for some of the stories - https://www.youtube.com/user/sameergoswami/playlists

sentences as annotated by the annotators are shown in Table 2. Next, we present some of the challenges that we faced during the annotation process that we think should be explicitly pointed out in order to provide a true picture of the corpus as well as to give an idea of the difficulties in carrying out such a process.

4.1 Challenges in Annotation

Apart from the challenge of annotating a low-resource language for which one can seldom get high quality crowd workers, there were certain challenges that were both specific to the domain of stories as well as generic ones peculiar to the tasks of sentiment and emotion analysis. Some of the prominent ones as identified from the feedbacks of the annotators are presented below with examples.

Identifying Implicit Emotions - The annotators were asked to identify the emotions whenever it was both explicitly and implicitly expressed. Identifying implicit emotions were sometimes confusing for the annotators and on taking a closer look we did find some of them being marked as neutral. An example of explicitly expressed emotion would be - Example 1, in which the speaker by using the words such as सुहावना (refreshing), मनोहर (beautiful) clearly indicates that he is happy with the nature, thus expressing his joy in the statements.

. *Example 1* - कितना मनोहर, कितना सुहावना प्रभाव है| वृक्षों पर अजीब हरियाली है, खेतों में कुछ अजीब रौनक है, आसमान पर कुछ अजीब लालिमा है| *(It is such a beautiful and enjoyable feeling. There is a strange greenery on the trees, some strange liveliness in the fields, there is some weird but enjoyable redness in the sky)*

An example of implicitly expressed emotion would be - Example 2, in which a child’ s grandmother is complaining about her son being too hasty of going to the mosque. She complains of his ignorance of knowing anything about driving a household and its inherent difficulties. Although there aren’t any explicit word indicating her state of the mind, there is an implicit pointer that she is feeling irritated due to the haste and hence is angry over him. These types of emotions are totally contextual and could be identified only while reading the story. We believe that capturing these emotions are also necessary in order to make our annotation process holistic. Although, we don’t train any classification model in this work that can take these types of context in order to predict the final emotion of a sentence, yet we think that BHAAV as a dataset provides an opportunity to build such contextual models making it a rich corpus unlike many other previous ones as already pointed out in Section 2. We would certainly like to take it up as a future work.

. *Example 2* - अब जल्दी पड़ी है कि लोग ईदगाह क्यों नहीं चलते| इन्हें गृहस्थी की चिंताओं से क्या प्रयोजन| *(Now he is feeling why don’ t people go to the mosque a little faster. What do they (the children) know about household chores)*

Primary Target of Opinion - Another challenge comes when there is not even an implicit clue in the immediate context of a sentence. For instance, in a story, sometimes a character is developed as an adversary to a particular prop (or, PTO (Primary Target of Opinion). The prop can be another character or some inanimate object or phenomena. From the start of the story, the character expresses his emotions in a characteristic manner towards that PTO. Thus if a sentence or a context does not have any explicit clues to know the state of the mind of the character, identifying the PTO and the character‘s emotions towards PTO gives some connotation to that sentence. This is in line to what was suggested in the work (Mohammad, 2016). An example of such an instance as presented in Example 3, can be derived from the famous story by Premchand, “Eidgah” . The following sentence when read in isolation could potentially trick someone into thinking whether the boy speaking these dialogues is expressing mercy or even neutrality, when he is actually expressing joy.

. *Example 3* - मोहसिन- लेकिन दिल में कह रहे होंगे कि मिले तो खा लें| *(Mohsin- But in the hearts, they must be thinking that if they could get it, they would eat it)*

Sarcasm - A common challenge which annotators faced while annotating the dataset is the case of sarcasm, which is again prevalent in most of the previous works in sentiment and emotion analysis. Sarcasm, as it occurs, is generally accompanied by either anger or delight of the speaker at the dismay of the PTO. Thus, in most cases, the speaker of sarcastic comments were either angry with the PTO or rejoicing at its expense. Annotators were asked to differentiate between these two instances clearly using the context provided, which was sometimes challenging. An example of sarcasm is presented in Example 4, in which the actual emotion expressed is *anger*, when it could be easily misunderstood to be *joy*.

- *Example 4* - हा हा हा! अब तुम बताओगे हम क्या बोलें? *(Ha Ha Ha ! Now you would tell me what I should speak?)*

Annotating Suspense - Suspense was the toughest category for the annotators. Sometimes, it proved very difficult for the annotators to know exactly when a sentence is of the category *suspense*. The annotators were asked to mark a sentence as suspense when there is some element in it which evokes a sense of anticipation or worry. Suspense is a unique feature of stories which does not get fully expressed in other types of written materials such as news articles, formal reports, etc. Examples of such sentences are given in Example 5.

- *Example 5* - पिछले पहर को महफिल में सन्नाटा हो गया| हू-हा की आवाजें बन्द हो गयीं| लीला ने सोचा, क्या लोग कहीं चले गए, या सो गये? एकाएक सन्नाटा क्यों छा गया? *(Last afternoon, the silence was over the entire place. There were no voices around. The sounds of Hu-Ha completely stopped. Leela thought, did people go somewhere, or perhaps they slept? Why all of a sudden there is silence everywhere?)*

Next, we present the experiments performed for training the baseline models.

5 Baseline Models

In this section, we describe the baseline models that we train for the task of identifying one of the emotions - *anger*, *joy*, *suspense*, *sad*, and *neutral*, from a given sentence taken from a Hindi story. Both classic machine learning and modern deep learning models are trained. We report their performances and analyze the results. We extensively use Sklearn (Pedregosa et al., 2011) and Keras (Chollet et al., 2018) as our machine learning toolkits.

5.1 Dataset

The BHAAV dataset was randomly shuffled and split into train and test datasets with a ratio of 10:1. The distribution of labels in the two datasets are shown in Table 3. The proportion of distribution of labels in the test dataset is kept similar to the training dataset. We train our models on the training dataset and test the final predictions on the test dataset. We do not create a separate validation dataset. However, we do use validation data extracted from the training data, whenever necessary for tuning the hyperparameters of the models.

5.2 Text Preprocessing

Before training the classification models one needs to preprocess the text and represent each sentence as a feature vector. We tokenize each sentence into words and remove punctuations. We do not remove the stopwords. Since we deal with Hindi, the standard word tokenizers that are suitable for English language could not be used. Therefore, we used the tokenizer shipped with Classical Language Toolkit[4]. Each sentence is vectorized after a feature extraction step for the classic machine learning models such as Support Vector Machines. Unigrams, Bigrams and Trigrams were generated as features for each sentence and their TF-IDF (Aizawa, 2003) scores were considered as the feature values.

One of the key components of the input fed to the deep learning models are pre-trained word embeddings (Kusner et al., 2015), that are used for representing each word of the input sentences by a dense real valued vector. Since the dataset on which we train our models is relatively small, we use the pretrained word embeddings in order to prevent overfitting. This practice is commonly known as transfer learning[5]. We choose the Fasttext[6] word embeddings (Bojanowski et al., 2016), trained on the Hindi Wikipedia corpus. This was a natural choice due to its easy availability. Additionally, Fasttext is possibly a better choice than other popular word embedding methods as it is more suitable for representing words belonging to morphologically rich languages such as Hindi as described in Section 3.

While training the deep learning models, each sentence in the training and test dataset is converted to a fixed size document of 126 words (maximum length of a sentence in the dataset). Padding[7] is used for sentences of length lesser than 126 words. Each word is

[4] http://docs.cltk.org/en/latest/hindi.html

[5] ftp://ftp.cs.wisc.edu/machine-learning/shavlik-group/torrey.handbook09.pdf

[6] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

[7] https://keras.io/preprocessing/sequence/

represented as a 300 dimensional vector by the word embedding model. All the words in the dataset are mapped to their corresponding word embedding vector. Whenever a word is not found in the vocabulary of the word embedding model we assign it a 300 dimensional zero vector. Each sentence is then represented as a matrix of its constituent words and their corresponding embedding vector, which is then fed as an input to the deep learning algorithms.

Hyperparameter	Range
No. of Filters for CNN	100, 200, 300, 400
Filter sizes for the CNN model	1, 2, 3, 4, 5, 6
Dense Output Layer Size	100, 200, 300, 400
Dropout Probability	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Learning Rate	0.0001, 0.001
Batch Sizes	8, 16, 32, 64, 128
Epochs	10, 50, 100, 150
LSTM units	8, 16, 32, 64, 128, 256

Table 4: Hyperparameter ranges used for random search during training deep learning models (CNN and Bidirectional LSTM).

Method	Macro Avg Precision	Macro Avg Recall	Macro Avg F1	Accuracy
Logistic Regression	0.58	0.62	0.58	0.62
SVM	0.48	0.52	0.49	0.52
Random Forests	0.44	0.59	0.45	0.59
CNN	0.50	0.55	0.51	0.55
BLSTM	0.43	0.60	0.47	0.60
Random Classifier	0.40	0.40	0.40	0.40

Table 5: Performance of the baseline supervised classification models on BHAAV dataset.

Emotion	Top 10 Important Unigram Features
joy	प्रसन्न (glad), सुंदर (beautiful), खुश (happy), हँस, (laugh), संगीत (music), खिलौने (toys), मजा (fun), आनंद (joy), हँसकर (smilingly), उछल (jump)
anger	अपमान (insult), गुस्सा (anger), क्रोध (anger), बदला (revenge), मूर्ख (idiot), सजा (punishment), जहन्नुम (hell), आग (fire), दुष्ट (evil), चिल्लाया (screamed)
suspense	आवाज़ (sound), आश्चर्य (astonishment), जिन्न (Genie), देखा (saw), युद्ध (war), छन··· (sound of anklets), कहाँ (where), जादू (magic), अचानक (suddenly), जहाज (ship),
sad	रो (cry), मर (die), रोने (crying), दुख (sadness), हृदय (heart), दुखी (sad), जीवन (life), आँसू (tears), रोते (cry), भगवान् (God)
neutral	किसान (farmer), उसने (he), बिन्नी (Binny), पूछा (asked), दादाजी (grandfather), कल (tomorrow), पंडित (pundit), मेहता (mehta), मां (mother), आना (come)

Table 6: Top 10 most important features for each emotion category as identified by the Logistic Regression model during training.

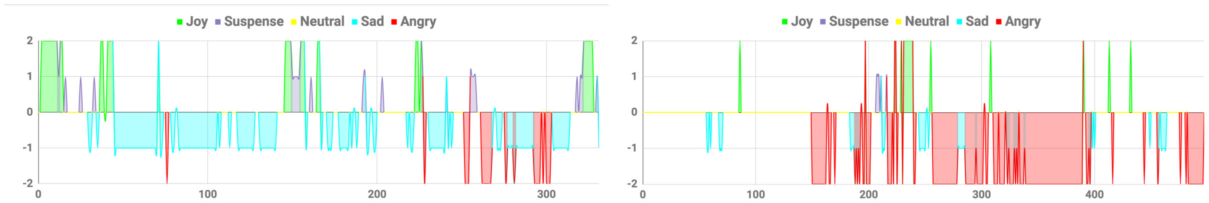
5.3 Training

All the machine learning models were trained after selecting the hyperparameters on a validation data. 10-fold Cross Validation was used for the classic techniques. For the deep learning models, random search (Bergstra and Bengio, 2012) was used for selecting the best hyperparameters among the ones shown in Table 4, that best fitted a fixed randomly selected validation data comprising of 20% of the training data. Only 100 iterations of random search was performed. Once the hyperparameter tuning was done the final model was trained on the entire training data using the selected hyperparameters. Adam (Kingma and Ba, 2014) with two annealing restarts has been shown to work faster and perform better than SGD in other NLP tasks (Denkowski and Neubig, 2017). Therefore, we use the same as our optimization algorithm for the deep learning models. As the task is a multi-class classification problem, categorical cross entropy was used as the loss function, and the final layer of both the deep learning models consisted of a fully-connected dense neural network with the extracted features as the input and a softmax output giving the prediction probability for each of the five emotion categories.

Among the classic machine learning techniques, *Support Vector Machine* (SVM) with a linear kernel (Hsu et al., 2003), *Logistic Regression* (Yu et al., 2011) and *Random Forests* (Breiman, 2001) were trained. A shallow Convolutional Neural Network with a single input channel similar to (Severyn and Moschitti, 2015), and Bidirectional Long Short Term Memory networks with an architecture similar to (Mahata et al., 2018), are the deep learning models that were trained. A random classifier that randomly generated predictions from a label distribution similar to that of the training dataset was also implemented. Table 5 summarizes the performances of the classifiers on the test dataset for the following metrics - *macro average precision*, *macro average recall*, *macro average F1-score*, and *accuracy* (Sokolova and Lapalme, 2009). We chose macro-average measures as the data is imbalanced and macro-averaging will assign equal weights to all the categories, which gives a better generic performance of any classifier.

6 Discussion

In order to analyze the possible features chosen by a machine learning classification algorithm for discriminating between different categories of emotions and to validate the ability of the BHAAV dataset in providing such features to



(a) Idealist

(b) Exploiter and Exploited

Figure 1: Flow of emotions in randomly selected stories from two different genres.

any classifier, we looked at the most important features chosen by the Logistic Regression model. Table 6 shows the top 10 most informative unigram features for each category of emotion chosen by the model in order to make the final predictions. As evident from the choices, words like प्रसन्न (glad), सुंदर (beautiful), खुश (happy), हँस, (laugh), are sensible indicators of *joy*, and so are the words like अपमान (insult), गुस्सा (anger), क्रोध (anger), बदला (revenge), for *anger*. The other categories also show a similar pattern.

We also looked at the performance of the classifiers for individual categories. The category of *neutral* had the best performance consistently, which is quiet easy to guess from the the data distribution (Table 3) and it being the majority class. The performance of the *suspense* category was consistently low. Although, the category of *anger* had a similar presence in the dataset, yet it had better performance than *suspense*. This might be due to the presence of better discriminative features for *anger* than *suspense*. The other reason could be related to challenges associated with annotating the *suspense* category (Section 4.1).

Our analysis provides a brief insight into the BHAAV dataset from which we can conclude that it is an appropriate dataset for emotion identification and classification tasks. Although, the dataset is created from stories, it can possibly be used for many other domains as it is rich in features indicating the five different emotions as presented in this work. The annotations were done from the perspective of a reader/narrator trying to express the emotion of a sentence, given the existing scenario in the story and whenever applicable trying to express the emotion of a character in the story. This also makes this dataset suitable for training automated text-to-speech interfaces (e.g., audio books) for story narration and improving them by infusing emotions in them.

The dataset is also appropriate for analyzing the flow of emotions in individual stories and study them for different genres. We plotted the flow of emotions in a randomly picked story from two different genres as shown in Figure 1. It is observable from the figures that each story has its own distinct emotion footprint. It would be interesting to study them and draw interesting linguistic insights from the Hindi literature. BHAAV can facilitate such experiments.

7 Future Work and Conclusion

In this work we publicly shared the first and the largest annotated corpus - BHAAV with 20,304 sentences in Hindi, categorized as expressing one of the five different emotions - *anger*, *joy*, *suspense*, *sad*, and *neutral*. The sentences are collected from 230 different short stories spanning across 10 genres. We provided a detailed description of the dataset, language specific challenges, annotation process, challenges associated with annotations and reported performances of the baseline classification models trained on the dataset for identifying emotions expressed in a sentence. Through different observations we confirm the dataset to be rich with emotion cues and point to the potential applications of the dataset. In the future, we plan to work on enriching the dataset with more annotations related to sentiment and discourse analysis. We believe that BHAAV will prove to be a valuable resource in Hindi and encourage further experiments in the domain of emotion analysis from text.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics.

Piyush Arora. 2013. Sentiment analysis for hindi language. *MS by Research in Computer Science*.

AR Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. *Proceedings of COLING 2012: Posters*, pages 73–82.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Roger Bougie, Rik Pieters, and Marcel Zeelenberg. 2003. Angry customers don't come back, they get back: The experience and behavioral implications of anger and dissatisfaction in services. *Journal of the Academy of Marketing Science*, 31(4):377–393.

Cynthia Breazeal and Rodney Brooks. 2005. Robot emotion: A functional perspective. *Who needs emotions*, pages 271–310.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

François Chollet et al. 2018. Keras: The python deep learning library. *Astrophysics Source Code Library*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. *arXiv preprint arXiv:1706.09733*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Virginia Francisco and Pablo Gervás. 2006. Automated mark up of affective information in english texts. In *International Conference on Text, Speech and Dialogue*, pages 375–382. Springer.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517–4528.

Narendra Gupta, Mazin Gilbert, and Giuseppe Di Fabbrizio. 2013. Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification.

Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kathrin Knautz, Tobias Siebenlist, and Wolfgang G Stock. 2010. Memose: search engine for emotions in multimedia documents. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and development in information retrieval*, pages 791–792. ACM.

Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao. 2011. Iitkgp-sehsc: Hindi speech corpus for emotion analysis. In *Devices and Communications (ICDeCom), 2011 International Conference on*, pages 1–5. IEEE.

Zoltán Kövecses. 2003. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

Zhang Lei, Wang Shuai, and Liu Bing. 2018. Deep learning for sentiment analysis: A survey. *Cornell Science Library*.

Diane J Litman and Kate Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 351. Association for Computational Linguistics.

Debanjan Mahata, Jasper Friedrichs, Rajiv Ratn Shah, et al. 2018. # phramacovigilance-exploring deep learning techniques for identifying mentions of medication intake from twitter. *arXiv preprint arXiv:1805.06375*.

Namita Mittal, Basant Agarwal, Garvit Chouhan, Nitin Bania, and Prateek Pareek. 2013. Sentiment analysis of hindi reviews based on negation and discourse relation. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 45–50.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179.

Saif M Mohammad. 2017. Challenges in sentiment analysis. In *A Practical Guide to Sentiment Analysis*, pages 61–83. Springer.

Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*.

Andrew Ortony, Gerald L Clore, and Mark A Foss. 1987. The referential structure of the affective lexicon. *Cognitive science*, 11(3):341–364.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. 2015. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer.

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. *Journal of machine learning research*, 12(Oct):2825–2830.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Niklas Ravaja, Timo Saari, Marko Turpeinen, Jari Laarni, Mikko Salminen, and Matias Kivikangas. 2006. Spatial presence and emotions during video game playing: Does it matter with whom you play? *Presence: Teleoperators and Virtual Environments*, 15(4):381–392.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.

Monika Schwarz-Friesel. 2015. Language and emotion. *The Cognitive Linguistic Perspective, in: Ulrike Lüdtke (Hg.), Emotion in Language. Theory–Research–Application, Amsterdam*, pages 157–173.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM.

Ameneh Gholipour Shahraki and Osmar R Zaiane. 2017. Lexical and learning-based emotion mining from text. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 70–74. Association for Computational Linguistics.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, pages 1083–1086. Citeseer.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for hindi-english code-mixed social media text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 128–135.

Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):25.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE.

Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75.