Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7061–7070
Marseille, 20-25 June 2022

HindiMD: A Multi-domain Corpora for Low-resource Sentiment Analysis

Mamta[1], Asif Ekbal[1], Pushpak Bhattacharyya[2],
Tista Saha[3], Alka Kumar [3], Shikha Srivastava[3]
[1] Department of Computer Science and Engineering
Indian Institute of Technology Patna, India
[2] Department of Computer Science and Engineering
Indian Institute of Technology Bombay, India
[3] Centre for Development of Telematics (C-DOT, India)
{mamta_1921cs11, asif}@iitp.ac.in, pb@cse.iitb.ac.in, {shikha, alkakm, tista}@cdot.in

Abstract

Social media platforms such as Twitter have evolved into a vast information sharing platform, allowing people from a variety of backgrounds and expertise to share their opinions on numerous events such as terrorism, narcotics and many other social issues. People sometimes misuse the power of social media for their agendas, such as illegal trades and negatively influencing others. Because of this, sentiment analysis has won the interest of a lot of researchers to widely analyze public opinion for social media monitoring. Several benchmark datasets for sentiment analysis across a range of domains have been made available, especially for high-resource languages. A few datasets are available for low-resource Indian languages like Hindi, such as movie reviews and product reviews, which do not address the current need for social media monitoring. In this paper, we address the challenges of sentiment analysis in Hindi and socially relevant domains by introducing a balanced corpus annotated with the sentiment classes, *viz.* positive, negative and neutral. To show the effective usage of the dataset, we build several deep learning based models and establish them as the baselines for further research in this direction.

Keywords: Sentiment, Low-resource Language, Indian Language, Multi-domain, Deep Learning, BERT

1. Introduction

Social media has become extremely popular among users to express their opinions on various serious events, issues and trending topics. Sentiment analysis is an important field of study that seeks to investigate the polarity of such user-expressed opinions (Pang et al., 2002). It has a wide range and powerful applications in natural language processing (NLP), from customer service to shaping political campaigns (Pang and Lee, 2004; Blitzer et al., 2007; Malo et al., 2013; Bakliwal et al., 2013; Akhtar et al., 2017). However, emerging social media platforms like Twitter and Facebook have become the new channel of information dissemination for many negative groups for recruitment in order to promote terrorist acts and illegal drug trades [1], incite violence and crime by influencing others. The mining of opinions expressed in these domains provides valuable stores of information that can be used by security agencies and the government in national counter-terrorism investigations and monitoring crime, drug activities, and domestic threats to ensure public safety. This information is also helpful in the defense sector for national security strategies, which are vital to national safety and security. Initial research was focused on product and movie reviews (Kumar et al., 2020) and now broadened to the other domains (Mamta et al., 2020), including finance (Moreno-Ortiz et al., 2020; Gaillat et al., 2018), politics (Abercrombie and Batista-Navarro, 2020), medicine (Zlabinger et al., 2018; Yadav et al., 20 18), etc. However, the majority of them are limited to a resource-rich language like English (Chakraborty et al., 2020; Fredriksen et al., 2018). India is a linguistically diverse country with 22 official scheduled languages. Hindi is the most spoken language of India and the fourth wide spoken language globally, leading to a vast increase in Hindi content on the web. Despite its popularity, sentiment analysis in Hindi is challenging due to resource scarcity. Limited research efforts have been put towards sentiment analysis of Hindi (Dashtipour et al., 2016; Kulkarni and Rodd, 2021). The available datasets contain a very small number of samples, often in the range of 200 or 250 reviews for binary classification (e.g., positive and negative) (Balamurali et al., 2012; Joshi et al., 2010). In (Akhtar et al., 2016), the authors released a dataset of product reviews and movie reviews containing 5417 and 2152 instances annotated with four different classes, *viz.* positive, negative, neutral, and conflict. The dataset, however, is imbalanced and does not address the need for social media monitoring in microblog text for the previously discussed socially relevant domains.

In our knowledge, there is no publicly accessible Hindi corpus dedicated towards sentiment analysis for these domains, which is crucial to maintain law and order situations. In our work, we firstly create a balanced multi-domain tweet corpus for the low-resource Indian language, Hindi, containing sufficient samples to train deep learning based models. The corpus is annotated

[1] https://www.business-standard.com/article/technology/narcotics-sourced-through-internet-social-media-report-112022900116_1.html

for 3 classes, *viz.* positive, negative, and neutral. We then develop a Multilingual Bidirectional Encoder representation using Transformers (mBERT) based baseline model to demonstrate the effective usage of the dataset. We obtain the accuracy and F1-measure values of 70.24% and 70.00%, respectively.

In the remaining part of the paper, the structure is as follows. Section 2 briefs the literature. Section 3 outlines the detailed processes and challenges involved in creating our corpus. Section 4 elaborates on the experiments, and Section 5 presents the experimental results and detailed analysis. Section 6 concludes the paper and future plans for research.

2. Related Work

A survey demonstrates that the majority of research in terms of resource creation and model development is done for the resource-rich language English, and there is a scarcity of annotated standard corpora for resource-poor language Hindi (Dashtipour et al., 2016). This section briefs the various resources and models created for low-resource sentiment analysis, focusing on the Hindi language.

Joshi et al. (2010) created corpora for Hindi containing 250 movie reviews containing equal positive and negative labelled data and Hindi SentiWordNet (HSWN), a lexical resource for Hindi. Bakliwal et al. (2012) translated the English product review dataset to Hindi using google translate and developed Hindi subjectivity lexicons for sentiment classification. They explored how the synonym and antonym relations can be exploited using sample graph traversal to generate the subjectivity lexicon. Authors in (Balamurali et al., 2012) collected user-written travel destination reviews from various blogs and Sunday travel editorials. The final dataset consists of 100 positive and negative reviews each, and a cross-lingual framework is proposed utilizing Linked WordNets of two languages (English and Hindi). These datasets contain a limited number of instances, which are not sufficient to train deep learning based classifiers. Mittal et al. (2013) increase the coverage of HSWN by including more opinion words and develop a corpus consisting of 380 positive and 282 negative reviews. They devised rules to handle negations and discourse relations, which highly influence the sentiments expressed in the review. The final sentiment is assigned by aggregating the polarity scores of all the words. Patra et al. (2015) provided a tweet dataset consisting of 1688 tweets annotated for 3 classes (positive, negative, and neutral) and organized a SAIL (sentiment analysis in Indian languages) task to advance research in Indian languages.

Authors in (Singhal and Bhattacharyya, 2016) used a translation system to convert input sentences from any language to English and then proposed Convolutional neural network (CNN) based classifier built on the top of English word embeddings. Authors in (Akhtar et al., 2016) created two Hindi datasets containing 5,417 product reviews and 2,152 movie reviews annotated for 4 classes. However, the corpus is imbalanced. Further, they proposed a CNN for automatic feature extraction and Support Vector Machine (SVM) for classification. In (Can et al., 2018), authors trained a sentiment analysis model using recurrent neural networks (RNN). They made use of English Glove embedding for reviews in English. They translated the other languages to English using a machine translation system and reused the trained model to evaluate the sentiment. The work described in (Attia et al., 2018) proposed a language independent model for classifying sentiment. They used a CNN for training and randomly initialized word embedding and learned it during training. Rani and Kumar (2019) experimented with a different configuration of CNN and created a Hindi movie reviews corpus, crawled from online newspapers and websites for their experiments.

The available datasets have the following limitations: (i). the limited number of samples; (ii). poor quality because of the use of a translation system; (iii). imbalanced; (iv). do not satisfy the need for social media monitoring for socially relevant domains. To address these limitations, we create a balanced dataset crawled from the Twitter.

3. Resource Creation

With the goal of creating a multi-domain dataset to support research on relatively low-resource languages like Hindi, we consider the domains that are more relevant to the society, as we discussed in Section 1. The domains we considered are: cybercrime, politics, terrorism, technology, and other social issues like casteism, crime, human trafficking, communal disputes, and narcotics. We design a tweet crawler to collect raw data for these domains, pre-process the data to remove noisy data, and annotate the tweets for the sentiment expressed in it. The resources can be obtained from [2]. We elaborate on the detailed process in the subsequent sections.

3.1. Data Crawling

Our first step is to crawl data from the Twitter. To this end, we use the Streaming API [3] and Twitter Search API [4] complying with Twitter's terms of service. The search API crawls tweets from the last seven days while the Streaming API retrieves real-time streaming data. We filter the keywords for socially relevant domains. Some of the keywords we use for data collection are listed in Table 1.

[2]The corpus is publicly available at `https://www.iitp.ac.in/~ai-nlp-ml/resources.html$#$sentimentM`

[3]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

[4]https://developer.twitter.com/en/docs/tweets/search/overview

Keyword	Translation
प्रौद्योगिकी (praudyogikee)	technology
हथियार (hathiyaar)	weapons
नशीले पदार्थों (nashile padaratho)	narcotics
मानव तस्करी (manav taskari)	human trafficking
अपराध (aparaadh)	crime
सांप्रदायिक विवाद (saampradaayik vivaad)	communal dispute
साइबर अपराध (cyber aparaadh)	cyber crime
आतंक (aatank)	terror
नक्सलवाद (naksalavaad)	naxalism
कश्मीर (Kashmir)	Kashmir
चक्रवात (chakravaat)	cyclone
जातिवाद (jaativaad)	casteism
आतंकवाद मुकाबला (aatankavaad mukaabala)	counter terrorism
विश्व शांति (vishv shaanti)	world peace

Table 1: Keywords

3.2. Data Filter or Pre-processing

We design a filter to clean the data to ease the annotation process and evaluation. At first, we remove: (i). tweets containing non-Hindi words except user mentions, hashtags, and URLs. For this, we use the indicnlp library (Kunchukuttan, 2020) to detect the language of each word; (ii). tweets with fewer than ten characters; (iii). tweets containing only URLs or user mentions; (iv). duplicate tweets; (v). tweets containing multimodal data. For experiments, we replaced the URLs present in the tweet with word *url* and all the English hashtags, user mentions, and retweet symbols are removed.

3.3. Annotations Process

A manual annotation of the dataset is conducted after pre-processing. The annotation process may rely on the annotator's opinions. In order to achieve high-quality annotation, we establish strict criteria for the choice of category. The linguist team comprised three linguists who have post-graduate level experience and good knowledge of Hindi.

Our sentiment annotations follow the guidelines used in the SemEval shared task (Rosenthal et al., 2015; Mohammad, 2016), which were explained to the annotators before starting the annotation process. Annotators were also provided with gold labelled samples to gain a deep understanding of sentiment labels. For every tweet, linguists write the overall polarity of the tweet in 3 categories *viz.* negative, neutral, and positive. Annotators were advised to refrain from being biased towards either a specific demographic area, religion, or ethnicity while annotating the tweets.

3.4. Challenges

In annotation process, we encountered the following main challenges:

- There are cases where tweets have both positive and negative content. The overall polarity in these cases is determined by the volume of the negative or positive content.
- A few tweets provide readers with information about a negative or positive situation or event, but the writer does not express his or her own opinion. For example,

  पारिवारिक विवाद में पति ने पत्नी पर किया धारदार हथियार से जानलेवा हमला

  Transliteration: parivaarik vivad main pati ne patni par kiya dhaaradaar hathiyaar se jaanaleva hamala

  Translation: In a family dispute, the husband attacked the wife with a sharp weapon.

  There is no opinion expressed in this tweet, so this can be annotated as neutral due to the non-expression of opinion or negative due to the negative situation or event. In our case, we opt to annotate tweets of this nature according to the situation described by the author. Therefore, the above tweet is marked as negative.

- In some tweets, the writer asks a question to express frustration. For example,

  आखिर क्यों.... महिलाओं की सुरक्षा और सम्मान की बात करने वाली पार्टी की सरकार बनते ही राजस्थान में अपराध का ग्राफ अचानक बढ़ गया?

  Transliteration: aakhir kyon.... mahilaon ki suraksha aur sammaan ki baat karane vale paartee kee sarakaar banate hee raajasthaan mein aparaadh ka graaph achaanak badh gaya?

  Translation: After all, why... the crime graph in Rajasthan suddenly increased as soon as the government, which talked about the safety and dignity of women, was formed?

  The writer of this tweet expresses frustration in asking a question. Based on this, we decided to mark the overall polarity as negative.

3.5. Data Distribution

In the corpus we created, we have 9,090 tweets spanning various domains, comprising 2,935 positive, 3,350 negative, and 2,805 neutral tweets. Figure 1 illustrates some statistics of our dataset. We analyze the frequency of few words relevant to our considered domains. The statistics are shown in Figure 2. It can be observed from the figure that the words कश्मीर (Kashmir), आतंक (aatank) (terror), अपराध (aparaadh) (crime), हथियार (hathiyaar) (weapons), and जातिवाद (jaativaad) (casteism) are occurring more frequently in the corpus.

3.6. Quality Test

To produce a reliable dataset annotated by the multiple annotators, it is essential to find agreements among three annotators. In order to audit the quality of annotation from different annotators, we measure the inter-rater agreement. Cohen's s Kappa coefficient is a statistical measure to analyze the inter-rater agreement. The

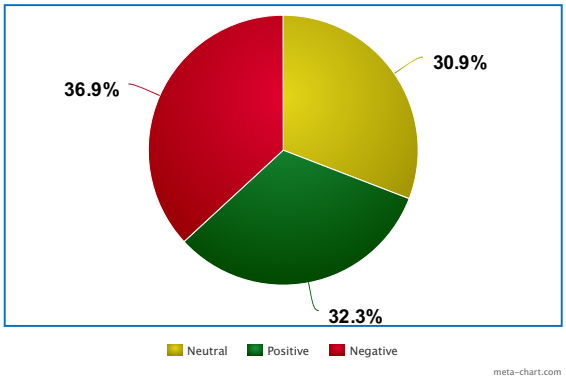

Figure 1: Data Distribution

measure is considered more robust than the simple percent agreement calculation. Kappa score can be defined as:

$$K = \frac{Prob(o) - Prob(e)}{1 - Prob(e)} \quad (1)$$

where Prob(o) and Prob(e) are the observed and by chance agreement among different annotators. Kappa score obtained for our corpus is 0.81 with a confidence percentile of 95%. As evidenced by the Kappa score, the data is of acceptable quality. We used a majority voting approach to merge three annotated versions of the corpus to obtain the final version. Some tweet examples, along with annotations, are presented in Table 2.

4. Experiments

To establish strong baselines, we implement the recent transformer based architectures because of its success in solving various Natural Language (NLP) tasks, including sentiment analysis(Devlin et al., 2018; Liu et al., 2019; Xu et al., 2019). We also compare this model with the other deep learning based models to provide a strong baseline.

4.1. Word Embedding

We first tokenize the input sentence into a sequence of words $T = t_1, t_2, t_3, ..., t_n$, with each token t_i being initialized using the pre-trained word embedding vectors provided by the fasttext (Bojanowski et al., 2017). Subword information is used by fasttext to generate the embedding, and as a result, it can handle out of vocabulary words. This word embedding vectors are given as input to the deep learning models listed below.

Convolutional Neural Network (CNN) CNN has been successfully applied to solve various NLP tasks (Kumar and Singh, 2019; Kim, 2014). This mathematical construct is composed of three types of building blocks: convolution, pooling, and layers that are fully connected. The first two layers, convolution and pooling, extract features. The convolution layer applies filters of varying size to preserve n-gram information, max pooling to extract the most relevant features, and the third layer, a fully connected layer, maps them into final output class. In our task, we use a convolutional layer, containing 128 filters of size 2, 3, and 4.

Long Short Term Memory (LSTM) The LSTMs are special kinds of recurrent neural networks (RNNs) that handle the vanishing gradient problem by gating mechanism, allowing them to learn long-term dependencies (Hochreiter and Schmidhuber, 1997). In our task, we use two LSTM layers stacked on the top of each other, consisting of 128 units each. Last hidden state of LSTM captures sentence information and is given as input to final classification layer.

Gated Recurrent Unit (GRU) GRU (Chung et al., 2014) has two gates, input and output gate, to modulate the flow of information without having a separate memory cell. In the LSTM unit, the output gate controls how much cell content is exposed to each unit, whereas in the GRU, the recurrent state is exposed without any control. There are a fewer parameters to learn in GRU, consequently it takes less time to train than LSTMs. For our task, we use two layers of GRU on top of each other, with 128 units in each GRU layer. Similar to LSTM, the last hidden state of GRU is given as input to the output layer.

Attentive LSTM and GRU LSTM and GRU treat all words in the input sentence equally while generating the final sentence representation. However, some words like adjective words are more important for sentiment classification. Attention mechanism highlights the important words responsible for the model's predictions, rather than focusing on all the words. Let h_t and $h_{t'}$ be the hidden representations of tokens w_t and $w_{t'}$ at time steps t and $t^{'}$, respectively, for the LSTM or GRU models. Attention mechanism is implemented as follows:

$$f_{t,t'} = tanh(W_f h_t + W_{f'} h_{t'} + b_f) \quad (2)$$

$$\alpha_{t,t'} = \sigma(W_a f_{t,t'} + b_a) \quad (3)$$

where W_f and $W_f^{'}$ are weight matrices associated with hidden states h_t and $h_t^{'}$ respectively; σ is the element-wise sigmoid function; W_a is the weight matrix corresponding to their non-linear combination; b_a and b_g are the bias vectors. The final token representation of t is obtained by the weighted sum of hidden representations $h_{t'}$ of all other tokens at time step t′, i.e.,

$$\sum_{t=1}^{n} \alpha_{t,t'} . h_{t'} \quad (4)$$

4.2. Bidirectional Encoder Representation from Transformer (BERT)

BERT (Devlin et al., 2018) provides contextual representations and proves to be beneficial in several NLP tasks. It is pre-trained on two unsupervised tasks, *viz.* masked language model and next sentence prediction task, in order to bring awareness of both previous and

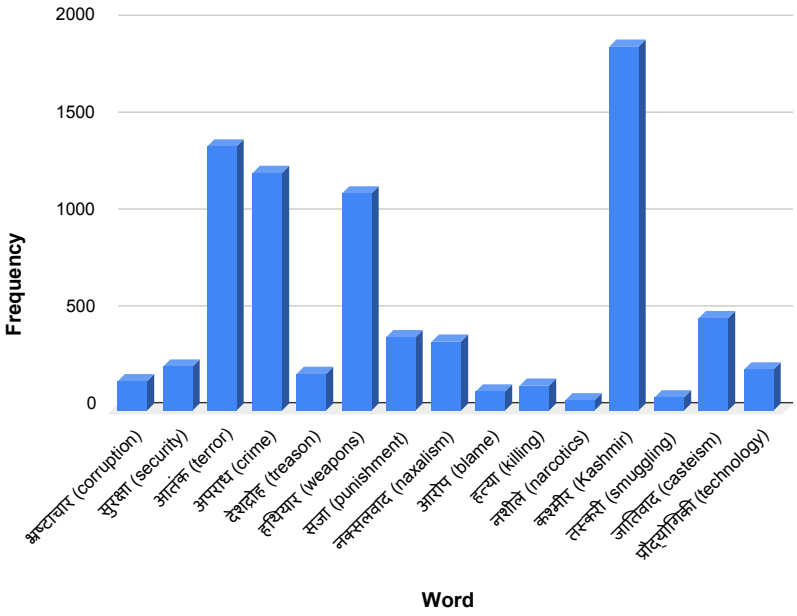

Figure 2: Frequency of few domains related words in the corpus

future contexts. In the masked language model, 15% of the tokens are replaced with $[MASK]$ tokens, and BERT is trained to predict those masked words. Hence, the BERT model captures the relationship between the words and can assign representation to each token based on the context in which it appears. Similarly, with the help of the next sentence prediction task, BERT captures the relationship between the sentences. For our case, we use multilingual BERT (mBERT) pre-trained on 104 languages, including Hindi on the largest Wikipedia corpora.

Input Representation Input to mBERT model is the sentence T, tokenized into WordPiece tokens (Wu et al., 2016) having length l. BERT adds two special tokens, at the beginning and end of the tokenized sentence. The input sequence is represented by $S = (CLS, t_1, t_2, t_3...t_l, SEP)$. Token embedding, segment embedding, and positional encoding vectors are assigned to each WordPiece token t_i, where the positional embeddings help capture word order, and segment embeddings preserve the information about the number of inputs to the BERT. The final input representation for a given input is calculated by summing over all the three embedding vectors, which are further fed to the BERT encoder for contextual representations.

Multi-lingual BERT Encoder BERT encoder is constructed by the transformers blocks, used in 12 different layers and multi-head attention is used in each layer of the BERT encoder. The output representation from the previous transformer layer is fed as input into the next layer. The multilingual BERT encoder maps the input representation into a sequence of contextual embedding vectors. The output of the $[CLS]$ token is considered our final input sequence representation, which inherently captures the information of the whole sequence.

4.3. Experimental Setup

We split the dataset into three parts: training, validation, and test set. We distribute 70%, 20% and 10% of the data as training, test and development, respectively. There are 6,544 samples in the training data set, 728 samples in the validation set, and 1818 samples in the test set. Table 3 shows the detailed class-wise distribution of train, validation, and test set. We use Keras [5] and PyTorch [6], Python-based deep learning libraries, to develop our models. BERT model is implemented in PyTorch and other models are implemented in Keras. The fasttext (Bojanowski et al., 2017) Hindi provides a 300-dimensional embedding vector for each word. Each model runs in batches of 16 sentences with a cross-entropy loss function. Model weights are updated with Adam optimizer. The activation function used in hidden layers is ReLU. The epochs are set to 3 for the mBERT model and 40 for the remaining models.

5. Experimental Results and Analysis

Experimental results on our proposed corpora are summarized in Table 4. CNN classifier reports the 67.25% accuracy with precision, recall, and F1 values of 67.29%, 67.54%, and 67.23%, respectively. LSTM and GRU classifiers report better performance than the

[5]https://keras.io/
[6]https://pytorch.org/

Tweet	Annotation
यह आपके परम मित्र देश है जो आपके दुश्मन को हथियार देते है । **Transliteration:** yeh aapake param mitar desh hai jo aapake dushman ko hathiyaar dete hai . **Translation:** It is your best friend country which gives weapons to your enemy.	Negative
किसी विवाद और झगड़े को सांप्रदायिक रंग देना बीमार मानसिकता का प्रमाण है। यहां कोई हिन्दू-मुस्लिम विवाद नहीं है। **Transliteration:** Kisi vivaad aur jhagade ko saampradaayik rang dena bimaar maanasikata ka pramaan hai. yahaan koi Hindu-muslim vivaad nahin hai. **Translation:** Giving communal color to any dispute and quarrel is a proof of ill mentality. There is no Hindu-Muslim dispute here.	Negative
मानव तस्करी सिर्फ भारत तक सीमित नहीं है. ये पूरे विश्व का सबसे बड़ा मुद्दा है. रोक पाना भी इतना आसान नहीं है **Transliteration:** Maanav taskari sirf Bhaarat tak seemit nahin hai. Ye poore vishv ka sabase bada mudda hai. Rok paana bhi itna aasaan nahin hai. **Translation:** Human trafficking is not limited to India only. This is the biggest issue in the whole world. It is also not that easy to stop.	Negative
सर , हमें गर्व इस बात का है कि आजादी के बाद पहली बार आतंक , भ्रष्टाचार , गरीबी , बेरोजगारी से लड़ने वाला योद्धा भारत को मिला । भारत के सही मायने में अनमोल रत्न है आप । **Transliteration:** Sir, hamein garv is baat ka hai ki aajaadi ke baad pehli baar aatank , bhrashtaachaar , gareebee , berojagaaree se ladane vaala yoddha bhaarat ko mila. Bharat ke sahi maayane main anmol ratan hai aap . **Translation:** Sir, we are proud that for the first time after independence, India got a warrior fighting against terror, corruption, poverty, unemployment. You are a truly priceless gem of India.	Positive
होली मे नशीले रंगों व नशीले पदार्थों से दुर रहे **Transliteration:** Holi me nasheele rangon va nashile padaarthon se dur rahe **Translation:** Stay away from intoxicants and intoxicants on Holi	Neutral

Table 2: Annotated samples for each class

Type	Train	Validation	Test
Negative	2405	268	677
Positive	2098	253	584
Neutral	2041	207	557
Total	6544	728	1818

Table 3: Data distribution for experiments

CNN classifier due to the capability to learn sequential features better. LSTM and GRU outperform CNN by 1.28% and 1.11%, respectively. Further, we observe that the attention mechanism improves the performance of both the GRU and LSTM models by focusing more on important words. Attention enhances the accuracy values of LSTM and GRU by 0.55 and 1.08 points, respectively. Finally, the mBERT classifier outperforms all the other classifiers, illustrating the importance of contextual representations. The mBERT classifier reports an accuracy value of 70.24% with precision, recall, and F1 scores of 69.91%, 70%, and 69.87%, respectively. The mBERT reports an increase in accuracy value of 3, 1.71, 1.84, 1.16 and 0.77 points by LSTM, GRU, attentive LSTM, and attentive GRU, respectively.

5.1. Detailed Analysis

We perform a detailed analysis to analyze the output of all the models. We discuss the cases where i). the attention mechanism helps the deep models to perform correct classification, ii). all models except mBERT perform misclassification.

1) Actual Example: और उसके साथ ही वहाँ के हर आतंक के प्रेमी की सांसो का रिश्ता भी उनसे टूट जायेगा इसे भी अच्छे से ध्यान मे रखना महबूबा
Transliteration: aur uske saath hi vahaan ke har aatank ke premi ki saanso ka rishta bhi unse toot jaayega ise bhi achchhe se dhyaan me rakhana mahabooba
Translation: And at the same time, the relationship of every terrorist's lover's breath will also be broken with them, keep this in mind very well, Mehbooba.
Actual Label: Negative
Predictions: CNN: Positive; **LSTM:** Neutral ; **GRU:** Neutral; **Attentive LSTM:** Negative; **Attentive GRU:** Negative ; **BERT:** Negative

Model	Precision	Recall	F1-measure	Accuracy
CNN	67.29	67.54	67.23	67.25
LSTM	68.04	68.39	68.14	68.53
GRU	68.22	68.14	68.15	68.39
Attentive LSTM	68.48	69.12	68.60	69.08
Attentive GRU	69.38	69.40	69.39	69.47
mBERT	**69.87**	**70.00**	**70.00**	**70.24**

Table 4: Experimental results

class	**negative**	**neutral**	**positive**
negative	489	105	83
neutral	110	340	107
positive	65	71	448

Table 5: Confusion matrix for mBERT classifier

2. Actual example: इस युग में देश की राजनीति का स्तर जितना गिरा है, उसे उठने में कितना वक्त लगेगा?
Transliteration: iss yug main desh ki raajneeti ka star jitana gira hai, use uthane main kitna vakt lagega?
Translation: How much time will it take to rise to the level of politics of the country has fallen in this era?
Actual: Negative
Predictions: CNN: Negative; **LSTM:** Neutral; **GRU:** Neutral; **Attentive LSTM:** negative; **Attentive GRU:** Negative; **BERT:** Negative

3). Actual Example: बहुत अच्छा लगा आज भारत द्वारा स्वयं बनाए गए "तेजस" का विश्लेषण देखकर। अब दिल को सुकून की अनुभूति मिली कि अब भारत विदेशों पर ही निर्भर नहीं है। यह नया भारत है। ये घर में घुस के मारता है। और अपने हथियार से मारेगा।
Transliteration: bahut achchha laga aaj Bhaarat dvaara svayan banaye gaye "Tejas" ka vishleshan dekhakar. Ab dil ko sukoon ki anubhooti mili ki ab bhaarat videshon par hi nirbhar nahin hai. Yeh naya Bhaarat hai. Ye ghar mein ghus ke maarata hai. aur apane hathiyaar se marega.
Translation: It was great to see the analysis of Tejas made by India itself today. Now the heart got a feeling of relief that now India is no longer dependent on foreign countries. This is new India. It enters the house and kills him and with its own weapon.
Actual: Positive
CNN: Negative; **LSTM:** Negative; **GRU:** Negative; **attentive LSTM:** Positive; **attentive GRU:** Positive; **BERT:** Positive

In example 1, the actual sentiment is negative. However, the presence of both positive and negative words in the tweet confuses the CNN, LSTM, and GRU models to predict sentiment as positive, neutral, and neutral, respectively. The attention mechanism helps both the LSTM and GRU models to focus more on negative words for correct classification; hence, both the models predict the sentiment correctly with attention. Similarly, in examples 2 and 3, the attention mechanism is helping both the LSTM and GRU models to perform correct classifications.
Examples 4 and 5 illustrate the cases where the mBERT model performs better than the other models.

4) Actual Example: श्रीमान जी, सालों पहले आपने शिक्षा-मित्रों की जिम्मेदारी ली, पर आजतक नहीं निभाई।
Transliteration: Shrimaan ji, saalon pehle aapane shikshaamitron kee jimmedaaree lee, par aajatak nahin nibhaee.
Translation: Sir ji, you took the responsibility of shikshaamitron years ago, but till date you have not done it.
Actual: Negative
Predictions: CNN: Neutral; LSTM: Neutral; GRU: Neutral; **attentive LSTM:** Neutral; **attentive GRU:** Neutral; **BERT:** Negative

5) Actual Example: श्रीमान जी आप ने 4 साल में कितने वादे किये एक बार आप भी सोचो 15 लाख आज तक नहीं आये।
Transliteration: shrimaan ji aapne 4 saal main kitne vaade kiye ek baar aap bhi socho 15 laakh aaj tak nahin aaye.
Translation: Sir ji, how many promises did you make in 4 years, once you also think that 15 lakhs have not come till today.
Actual: Negative
Predictions: CNN: Neutral; LSTM: Neutral; GRU: Neutral; **Attentive LSTM:** Neutral; **Attentive GRU:** Neutral; **BERT:** Negative

In examples 4 and 5, presence of word नहीं (*nahi*) negates the overall sentiment to negative. BERT model is able to capture it, but other models are not able to focus on the negation word; hence perform misclassification.

Error Analysis Further, we analyze the performance of our proposed model quantitatively through the confusion matrix, as shown in Table 5. It can be observed that most of the misclassifications of positive and negative classes are into neutral class. In the case of the neutral class, it is confused with the negative as well as the positive class.

Tweet	Actual	Prediction
सर आप धर्म और जातिवाद से काफ़ी ऊपर उठ चुके है । **Transliteration:** Sir aap dharam aur jaativaad se kaai uppar uth chuke hai. **Translation:** Sir you have risen above religion and casteism.	Positive	Neutral
मतलब है कि शान्ति चाहते हो तो मुझे विजयी घोषित करो । **Transliteration:** Matlab hai ki shaanti chahte ho to mujhe vijayi ghoshit karo. **Translation:** Means if you want peace then declare me victorious	Negative	Positive
किसी को चारा घोटाला करना हो तो कैसे करेगा ? **Transliteration:** Kisi ko chaara ghotala karna hoga to kaise karega ? **Translation:** How to do if someone wants to scam fodder?	Neutral	Negative

Table 6: Qualitative analysis of mBERT model

Table 6 shows some examples where the mBERT performs misclassification. Tweet 1 shows the positive sentiment, but the system misclassifies it as negative. The possible reason could be the absence of an explicit positive polarity marker in the tweet. Because of this, the classifier got confused and performed misclassification due to the presence of negative words. Similarly, in tweet 2, there is implicit negative sentiment due to the word तो (then), which mBERT cannot capture. In tweet 3, sentiment is neutral, but the classifier misclassifies it to the negative class. The possible reason could be the presence of negative word घोटाला (*scam*) due to which classifier got confused.

6. Conclusion

In this paper, we have presented a multi-domain corpus to push forward the research for sentiment analysis in low-resource language for the socially relevant domains. We have collected the dataset from Twitter, designed the annotation guidelines, and got these annotated by expert linguists. The dataset contains 2,935 positive, 3,350 negative, and 2,805 neutral instances. To demonstrate the use and quality of data, We trained deep learning based and recent transformers based classifiers for sentiment classification. Evaluation results show that the mBERT classifier outperforms all the other models and achieves an accuracy of 70.24%; hence can serve as a strong baseline for future works in this direction.

In future work, we would extend our current research by adding intensity values to each tweet, which can measure the magnitude of sentiment and further build a multitask model for both the tasks.

Acknowledgements

Authors would like to thank Centre for Development of Telematics, India (C-DOT) for funding this research. We would also like to extend special thanks to the linguists Saroj Jha (IIT Patna), Swati Srivastava (IIT Patna), and Akash Bhagat (IIT Patna), and for their support in annotation of tweets.

Bibliographical References

Abercrombie, G. and Batista-Navarro, R. T. (2020). Parlvote: a corpus for sentiment analysis of political debates. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5073–5078.

Akhtar, M. S., Kumar, A., Ekbal, A., and Bhattacharyya, P. (2016). A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493.

Akhtar, M. S., Kumar, A., Ghosal, D., Ekbal, A., and Bhattacharyya, P. (2017). A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 540–546.

Attia, M., Samih, Y., Elkahky, A., and Kallmeyer, L. (2018). Multilingual multi-class sentiment classification using convolutional neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bakliwal, A., Arora, P., and Varma, V. (2012). Hindi subjective lexicon: A lexical resource for hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 1189–1196.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., and Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.

Balamurali, A., Joshi, A., and Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73–82.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword infor-

mation. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Can, E. F., Ezen-Can, A., and Can, F. (2018). Multilingual sentiment analysis: An rnn-based framework for limited data. *arXiv preprint arXiv:1806.04511*.

Chakraborty, K., Bhattacharyya, S., and Bag, R. (2020). A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems*, 7(2):450–464.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fredriksen, V., Jahren, B., and Gambäck, B. (2018). Utilizing Large Twitter Corpora to Create Sentiment Lexica. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Gaillat, T., Zarrouk, M., Freitas, A., and Davis, B. (2018). The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Joshi, A., Balamurali, A., Bhattacharyya, P., et al. (2010). A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kulkarni, D. S. and Rodd, S. S. (2021). Sentiment analysis in hindi—a survey on the state-of-the-art techniques. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–46.

Kumar, A. and Singh, J. P. (2019). Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction*, 33:365–375.

Kumar, S., De, K., and Roy, P. P. (2020). Movie recommendation system using sentiment analysis from microblogging data. *IEEE Transactions on Computational Social Systems*, 7(4):915–923.

Kunchukuttan, A. (2020). The IndicNLP Library. `https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf`.

Liu, X., He, P., Chen, W., and Gao, J. (2019). Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Malo, P., Sinha, A., Takala, P., Ahlgren, O., and Lappalainen, I. (2013). Learning the roles of directional expressions and domain concepts in financial news analysis. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 945–954. IEEE.

Mamta, M., Ekbal, A., Bhattacharyya, P., Srivastava, S., Kumar, A., and Saha, T. (2020). Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5046–5054, Marseille, France, May. European Language Resources Association.

Mittal, N., Agarwal, B., Chouhan, G., Bania, N., and Pareek, P. (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.

Mohammad, S. (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.

Moreno-Ortiz, A., Fernández-Cruz, J., and Hernández, C. P. C. (2020). Design and evaluation of sentiecon: A fine-grained economic/financial sentiment lexicon from a corpus of business news. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5065–5072.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Patra, B. G., Das, D., Das, A., and Prasath, R. (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.

Rani, S. and Kumar, P. (2019). Deep learning based sentiment analysis using convolution neural net-

work. *Arabian Journal for Science and Engineering*, 44(4):3305–3314.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Singhal, P. and Bhattacharyya, P. (2016). Borrow a little from your rich cousin: Using embeddings and polarities of english words for multilingual sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3053–3062.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Xu, H., Liu, B., Shu, L., and Yu, P. S. (2019). Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Yadav, S., Ekbal, A., Saha, S., and Bhattacharyya, P. (20 18). Medical sentiment analysis using social media: towards building a patient assisted system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zlabinger, M., Andersson, L., Hanbury, A., Andersson, M., Quasnik, V., and Brassey, J. (2018). Medical Entity Corpus with PICO elements and Sentiment Analysis. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).