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

Mamta, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, Shikha Srivastava
AI-NLP-ML Group, Department of Computer Science and Engineering, Indian Institute of Technology Patna, India,
Indian Institute of Technology Bombay, India.
Centre for Development of Telematics (C-DOT, India)

Motivation and Contribution

- 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, etc.
- Mining opinions in these domains can help security agencies and the government.
- 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.
- Our study aims to create a balanced multi-domain tweet corpus for the low-resource Indian language, Hindi.

Data Collection

- We crawl data from Twitter using Twitter's streaming API and Twitter Search API. .

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

Data Preprocessing

- Tweets containing non-Hindi words except user mentions, hashtags, and URLs.
- Tweets with fewer than ten characters.
- tweets containing only URLs or user mentions
- duplicate tweets
- tweets containing multimodal data.

Annotations

- 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.
- For every tweet, linguists write the overall polarity of the tweet in 3 categories viz. negative, neutral, and positive.
- If tweets have both positive and negative content. The overall polarity is determined by the volume of the negative or positive content.
- If a tweet provide readers with information about a negative or positive situation or event, but the writer does not express his or her own opinion. These cases are marked according to the situation described.
- If the writer asks a question to express frustration.
- Kappa score obtained is 0.81 with confidence percentile of 95%.

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

Table 2: Data Distribution

Experiments

- We experiment with Convoutional Neural Network (CNN), Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), Multilingual BERT (mBERT) models.
- We use Keras and Pytorch, Python based libraries to develop our models.
- For CNN, LSTM, GRU models, we use the Hindi word embeddings provided by fasttext.

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 3: Experimental Results on different models

Detailed analysis

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 maha-booba
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

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

Figure 1: Output of different models

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 4: Qualitative error Analysis

Conclusion

- Proposed a multi-domain corpus to push forward the research for sentiment analysis in low-resource language for the socially relevant domains.
- 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.

References

- Akhtar et al. 2016. A hybrid deep learning architecture for sentiment analysis. In COLING-2016
- Devlin et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiV-2018
- Dashtipour et al. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. In Cognitive Computation 2018