Stacked embeddings and multiple fine-tuned XLM-RoBERTa models for Enhanced hostility identification

Siva Sai, Alfred W. Jacob, Sakshi Kalra, Yashvardhan Sharma

Birla Institute of Technology and Science,
Pilani.

Introduction

Introduction

- Hostile speech detection includes identifying the content that is fake, hateful, defaming, or offensive.
- **Challenges**
- The broad range of topics covered under hostile speech detection poses a challenge for identification due to different "patterns of writing" used by the users for different categories.
- The linguistic diversity of the Hindi language among different parts of the nation further increases the problem as that also would imply diversity of expression of hostile speech.

Dataset

- Used the dataset provided by Constraint 2021.
- Constraint 2021 focuses on the identification of hostile posts in Hindi Devanagari script, collected from Twitter & Facebook.
- The set of valid hostile categories are fake news, hate speech, offensive, defamation.
- Two sub-tasks :
  - Coarse-grained hostile speech detection - hostile v/s non-hostile
  - Fine-grained multi-label multi-class hostile speech identification where *each post can belong to one or more of these hostile classes.*

Few samples from the dataset

Text	Label(s)
#Saharanpur चेकिंग के बहाने उगाही करने वाले गिरफ्तार, 2 शातिर बदमाशों को पुलिस ने किया गिरफ्तार, स्कूटी, 2 तमंचे, मोबाइल बरामद किया, देहात कोतवाली पुलिस ने किया गिरफ्तार। @saharanpurpol @digsaharanpur @Uppolice #policemedianews	**non-hostile**
जापान के नोबल पुरस्कार जीतने वाले ओर 4 साल तक चीन में काम करने वाले प्रोफेसर #टासुकू_होंजो का दावा, कोरोना प्राकृतिक नही कृत्रिम वायरस है जो चीन ने बनाया है।	**fake**
दिल्ली में हिंदुओं और सिक्खों की सामूहिक हत्याएं करने के लिए आतंकवादी जेहादी घी के डिब्बे में अवैध हथियार सप्लाई करते हुए धर दबोचे गए।	**fake,hate**
कुत्ते में भी एकता होती है विचारों में विरोध होती लेकिन फिर भी उसे में एकता होती है पत्रकार में नहीं आज तक न्यूज़ चैनल पर आशुतोष सर को गौरव भाटिया बोला दोनों कान पर थप्पड़ मारूंगा वहां बैठा पत्रकार रोहित सरदाना सुनती रही मुझे बहुत दुख हुई	**offensive**

Proposed Techniques & Models

Pre-processing

- Removal of emojis.
- Removal of user mentions, URLs, hashtags and numbers.
- Removal of all kinds of punctuation except full-stop(‘|’ for hindi language) and comma.This minimal punctuation helps in preserving semantics of the text, which is useful for Transformer networks which are context-based.
- Removal of extra white spaces and other characters, other than alphabets in Devanagari script.

Language Models *from a word-embeddings perspective*

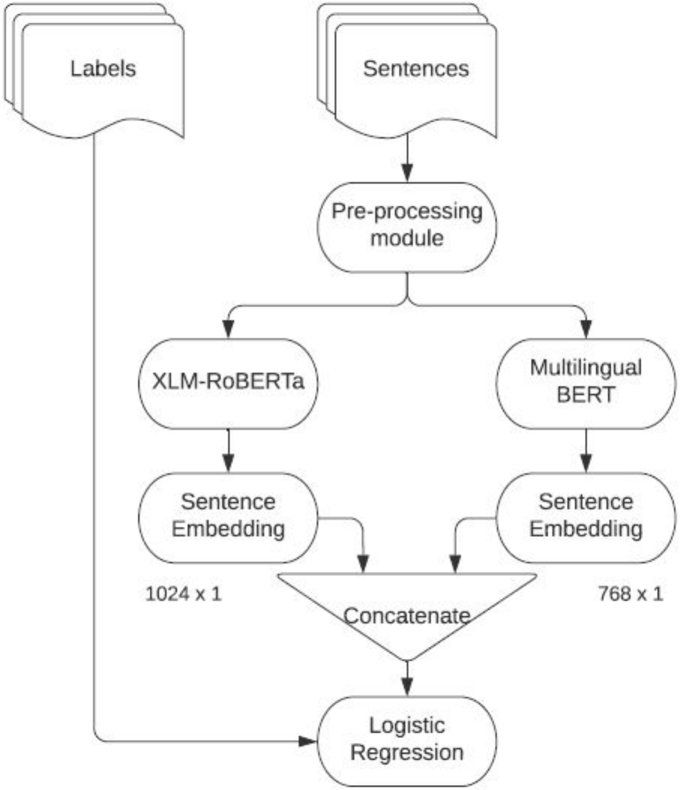
- **XLM-RoBERTa** is a multilingual model trained on 2.5 TB data from CommonCrawl.
- XLM-RoBERTa gives 1024 length vectors for each word , which are then averaged to get the embedding for the sentence.
- **ULMFit(**Universal Language Model Fine-tuning for Text Classification) - a novel method for fine-tuning of neural models for inductive transfer learning
- ULMFiT involves 3 major stages: LM pre-training, LM fine-tuning and Classifier fine-tuning.
- ULMFit embeddings are of 480 dimensions for each word.

- **mBERT** is trained on 104 languages, using the Wikipedia Corpora.
- Can be fine-tuned and applied to other monolingual and cross-lingual tasks where data is scarce.
- 768 dimensional word embeddings.
- **Flair** is a language model used to create contextual string embeddings
- Trained on JW300 corpus with more than 300 languages.
- Gives a 2048 dimensional vector for each word
- *Used Logistic regression as a classifier on top of these embeddings for baseline results.*

Stacked word embeddings

- Stacking multiple pre-trained embeddings yields better results than using a single type of word embedding.
- In effect, stacking is to **concatenate the final feature vectors** from more than one language model to create a single feature vector that is **richer in textual features**
- Used two different types of stacked embeddings: mBERT with XLMR, mBERT with Flair Backward, and Flair Forward.
- Used Logistic regression as a classifier on top of these stacked embeddings.

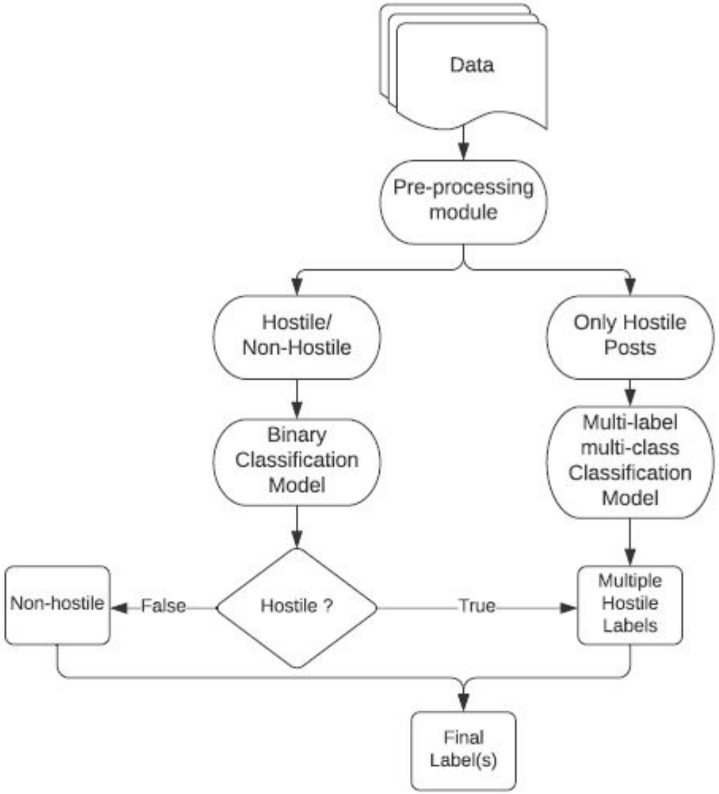
Stacked word embeddings



Multiple Fine-tuned Transformers Neural Networks

- A logical ensemble of two fine-tuned XLM-RoBERTa models.
- **Fine-tuned model-1**: *Binary classification model*: specializes in classifying a post as hostile or non-hostile
- Training data is divided into two classes - hostile(presence of any hostile label) and non-hostile.
- **Fine-tuned model-2:** *Multi-label multi-class classification model*: specializes in identifying multiple hostile classes.
- All the non-hostile posts are removed from the training data.
- After this step, 2678 samples are present in training dataset and 376 in the validation dataset.

Multiple Fine-tuned Transformers Neural Networks

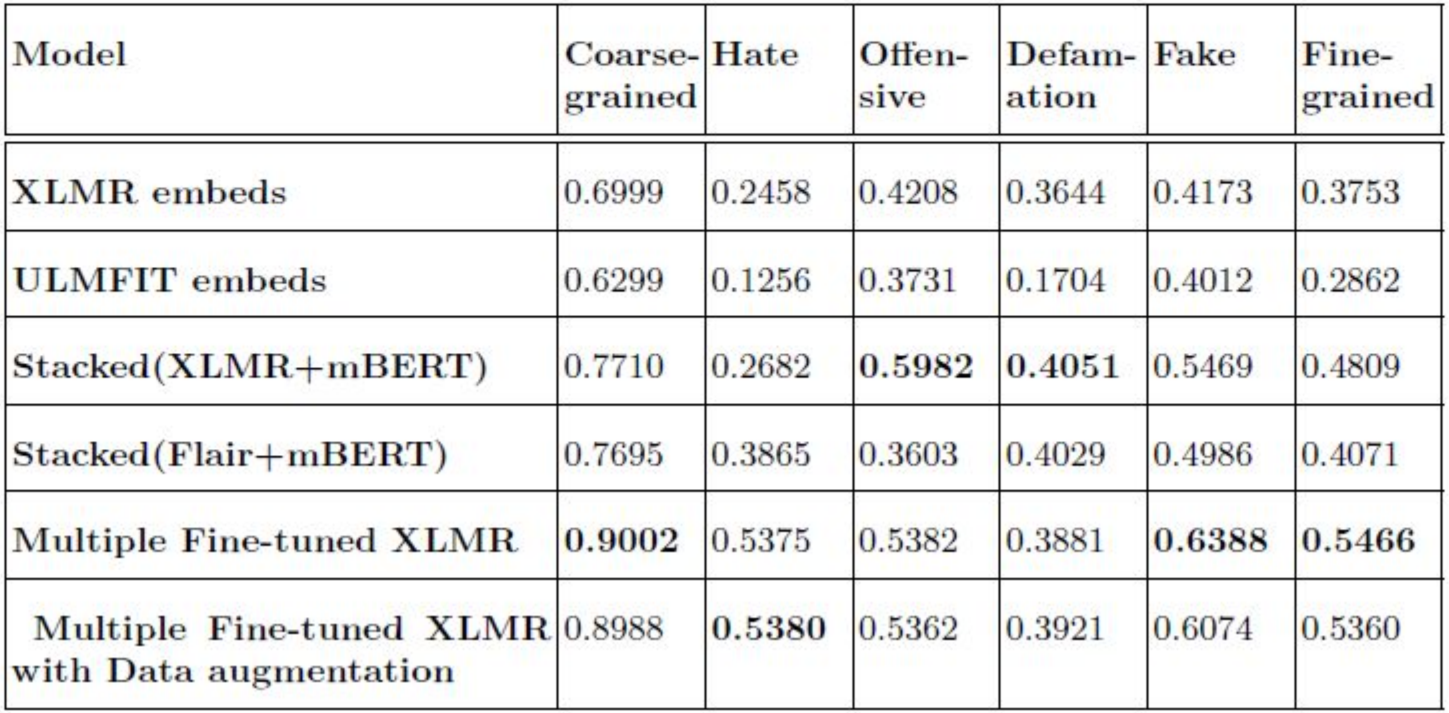


Back translation for Data Augmentation

- **Back translation** : A procedure of translating a document *that was previously translated into another language,* back to the original language.
- **HINDI => ENGLISH => HINDI**
- **Exploiting factor :** A back translation will never be 100% exactly the same as the original source text
- Back translation of translated Hindi post(now in English) to Hindi will produce the text that is different from the source text.(In case, it is same, don't consider it) So we can add this new sample to the dataset with the same labels as that of the original text.
- 4K additional samples, nearly doubling the dataset.

Results and Discussion

Results and Discussion



Model	Coarse-grained	Hate	Offensive	Defamation	Fake	Fine-grained
XLMR embeds	0.6999	0.2458	0.4208	0.3644	0.4173	0.3753
ULMFIT embeds	0.6299	0.1256	0.3731	0.1704	0.4012	0.2862
Stacked(XLMR+mBERT)	0.7710	0.2682	**0.5982**	**0.4051**	0.5469	0.4809
Stacked(Flair+mBERT)	0.7695	0.3865	0.3603	0.4029	0.4986	0.4071
Multiple Fine-tuned XLMR	**0.9002**	0.5375	0.5382	0.3881	**0.6388**	**0.5466**
Multiple Fine-tuned XLMR with Data augmentation	0.8988	**0.5380**	0.5362	0.3921	0.6074	0.5360

- On an overall basis, the fine-tuned XLM-RoBERTa model(without data augmentation) performed well, achieving the best coarse-grained F1-weighted score(0.90) and the best fine-grained F1-weighted score(0.54).
- The results show a significant improvement with stacking multiple word embeddings compared to the use of a single type of word embeddings.

Error Analysis*

- The model is not able to learn all of the multiple labels for a post correctly.
- The model predicts all the three classes - hate, defamation, offensive often when the actual labels are either one or two of the classes mentioned above.
  - This behavior can be due to the similarity among the classes in terms of their semantics. For example, users may use profane words in hateful, defaming, and offensive posts.
  - We guess that results can be improved with additional data, making the model learn meaningful differences among these classes.

* Examples in the paper.

- In many cases, the model fails to identify fake posts where fact-checking and worldly knowledge are required.
  - *BMC की अपील: अगलेसात िदनों तक गरम पानी का सेवन करें*
- Interestingly, the model predicts the fake category correctly when the text is long enough, which shows that the model can identify a subtle feature of fake posts - users try to put more content and details to make the false news appear authentic.
  - *गृह मंत्री अिमत शाह को गले के िपछलिहस्सेमें बोन कैं सर हो गया हैऔर यह मुसलमानों सेदुआ करनेके िलए कह रहेहैं जब इन को सत्ता िमली तो यह अपनेआप को खुदा समझने लगे िडटेंशन सेंटर में जो लोग हैं उनकी बद्दआु ऊपर वाला कभी रद नहीं करेगा जरूर कबूल करेगा।*
- The model is not able to identify sarcasm in posts, thus failing to learn one of the important aspects of hate speech.
  - *उद्धव ठाकरे िशवसेना को उसी ऊंचाई पर लेजाना चाहता हैं! जहाँराहुल गांधी कांग्रेस को पहुंचा चुका हैं*

Conclusions

- Presented a range of techniques for hostile speech identification in Hindi.
- Experimented with pre-trained word embeddings, stacked word embeddings, and fine-tuned XLM-RoBERTa model
- Used Back-translation for data augmentation.
- Our results demonstrate that stacking multiple word embeddings gives better results than using a single type of word embedding.
- Fine-tuned XLM-RoBERTa shows the best performance among all proposed models

- **Analysis**
- Why didn’t data augmentation improve the performance?
- Does the use of English as an intermediate language in back translation affect the quality of augmented samples?

- Each of these categories of speech has severe consequences on social media users.