Hostile Post Detection in Hindi

Group 25
Priya Mishra (170508, priyamis@iitk.ac.in)
Raunak Shah (170560, raunaks@iitk.ac.in)
Sagnik Mukherjee (170606, sagnikm@iitk.ac.in)
Yash Maheshwari (170817, yashma@iitk.ac.in)

OVERVIEW

01 PROBLEM STATEMENT & MOTIVATION

02 DATASET STATISTICS

03 DATA PREPROCESSING

04 METHODS

05 RESULTS

06 FUTURE WORK

01

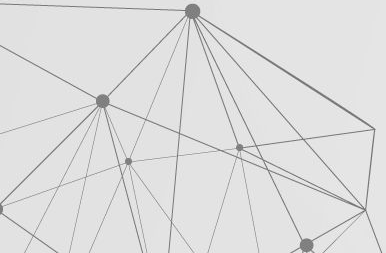
PROBLEM STATEMENT & MOTIVATION

Problem Statement

Multi-label multi-class classification of **Hindi posts** into non-hostile and hostile classes

Our goal is to maximise the weighted F1 score across all classes

Part of Shared task @ CONSTRAINT 2021, AAAI 2021



Motivation

Young people are exposed to more hate online during COVID
- The Conversation

Increase in online hate speech leads to more crimes against minorities
- Matthew L Williams et al.

Non-English tweets are now 50% of the total
- Twitter India MD

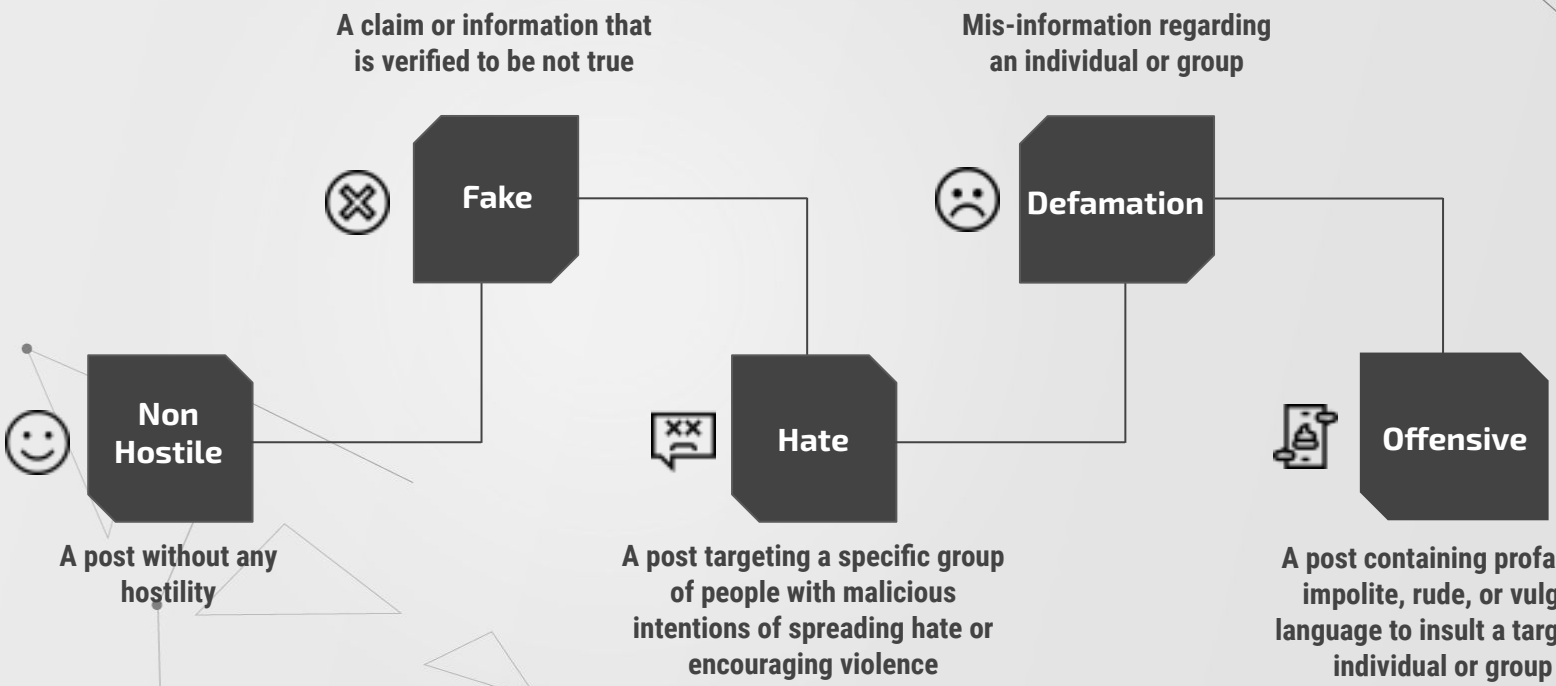
- 900% increase in hate speech on Twitter directed towards China and the Chinese
- 200% increase in traffic to hate sites and specific posts against Asians
- 70% increase in hate between kids and teens during online chats
- 40% increase in toxicity on popular gaming platforms, such as Discord

- L1GHT

02

DATA STATISTICS

Dataset - Classes



Dataset - Examples

Posts	Labels
मेरे देश के हिन्दु बहुत निराले है। कुछ तो पक्के राम भक्त है और कुछ बाबर के साले है 🙏 जय श्री राम 🙏	Hate,offensive
सरकार हमेशा से किसानों की कमाई को बढ़ाने के लिए नई-नई स्कीमें लाती रहती है, ताकि उन पर ज्यादा आर्थिक बोझ न पड़े. https://t.co/8iy2MJSBAs	Non-hostile
@prabhav218 साले जेएनयू छाप कमिने लोग हिन्दुओं को यह कहते है की संविधान सबको बराबर अधिकार देता है। सच्चाई यह है कि यह बराबर अधिकार नहीं देता है।	Defamation,offensive
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

Dataset - Statistics

Label	Train data	Validation data
Non-hostile	3050	435
Fake	1144	160
Hate	792	110
Offensive	742	103
Defamation	564	77
Total	5728	811

03

DATA PREPROCESSING

Data Preprocessing

Preprocessing Methods

Manual Preprocessing for RNNs

We wrote code to preprocess data for our baseline RNN models

Removed URLs, emoticons, usernames, punctuation, etc

Tokenized (Indic-NLP)

Lemmatized (Stanford-NLP)

Inbuilt BERT Tokenizer

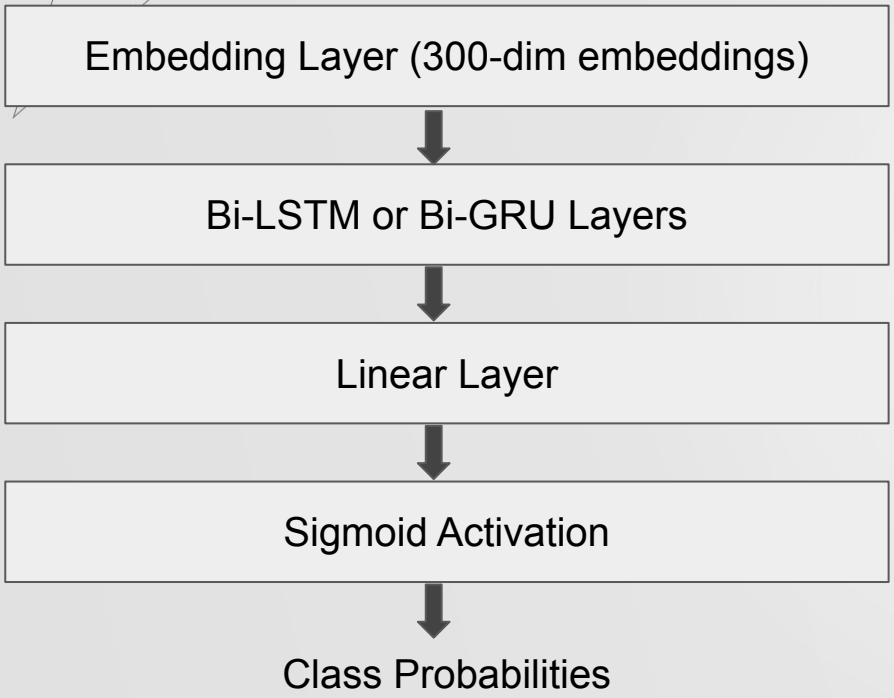
BERT based models have their own tokenizer which is tailored to the particular model used

04
METHODS

BASELINE

RNN-based Models

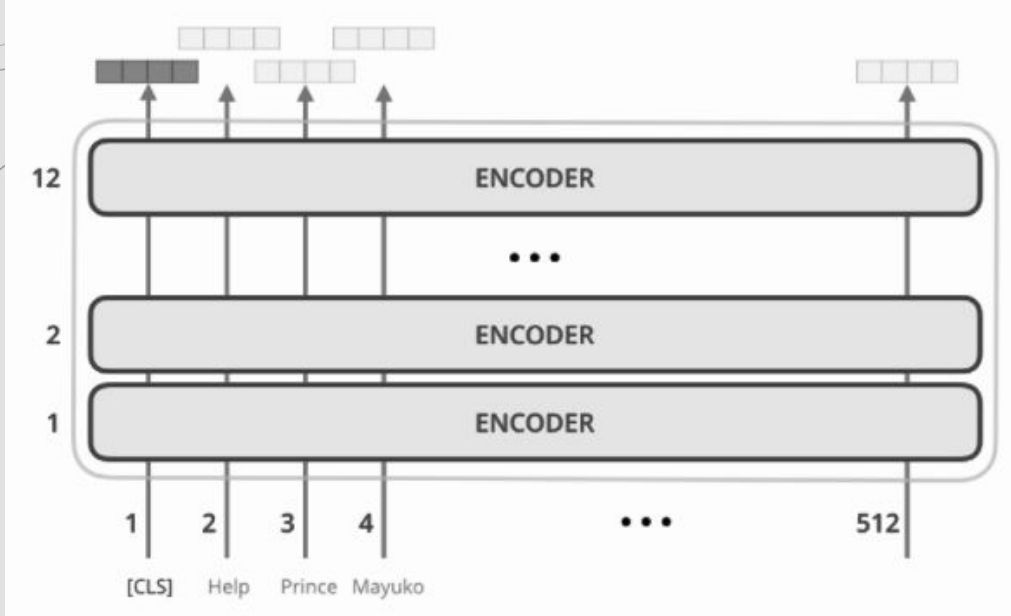
We use LSTM and GRU models as baseline results



- We use 300 dimensional word2vec embeddings
- The general architecture for the RNN-based models is shown
- We experiment with 1/2/4 Bi-LSTM and Bi-GRU layers
- The class probabilities are used to predict the classes using uniform threshold of 0.5 or varying optimum thresholds found for each class using precision-recall curves

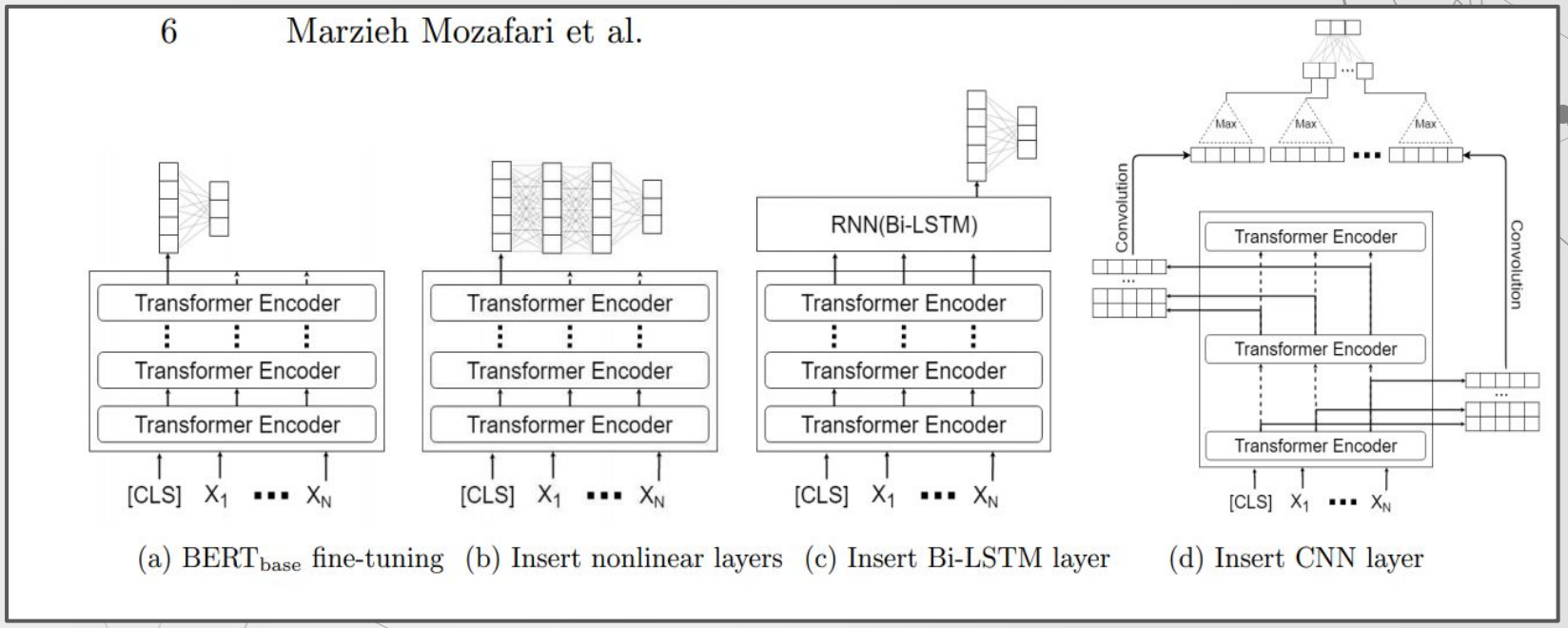
BERT based models

BERT



General Idea:

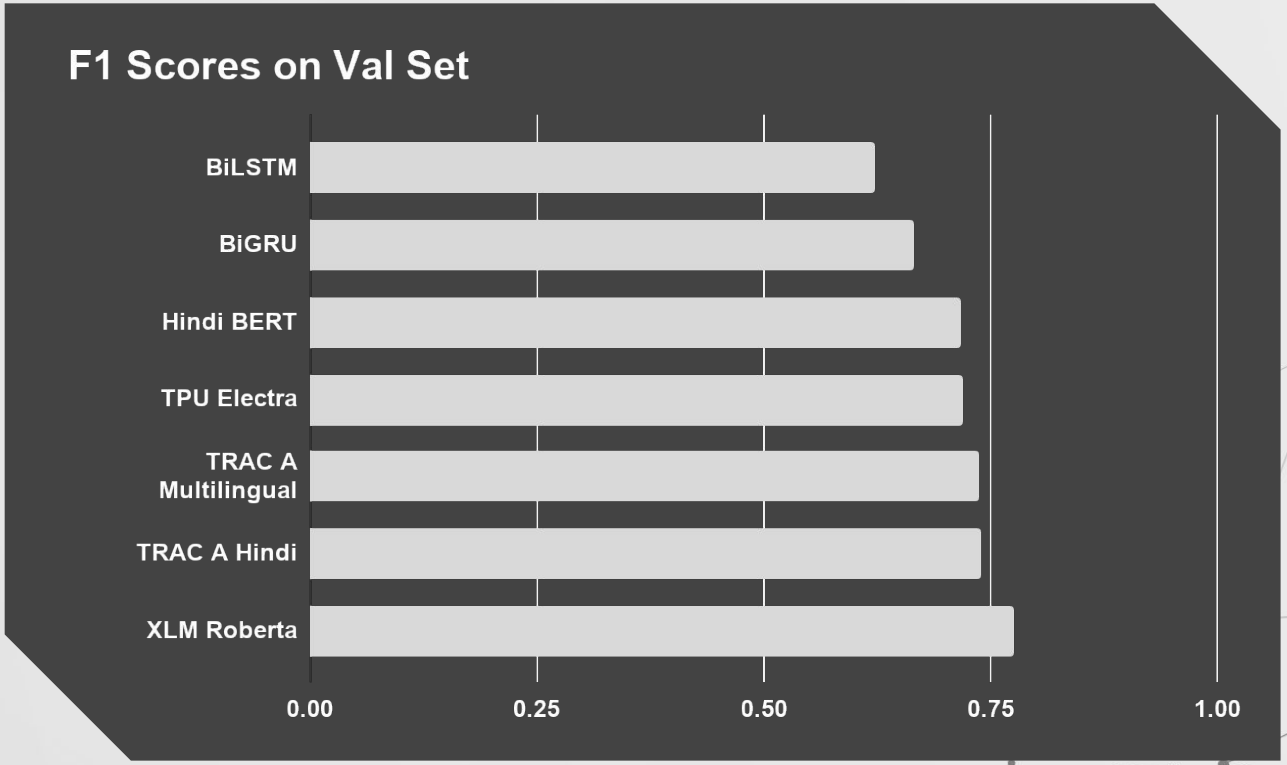
- Transformers Based Language Model
- Easy to finetune for downstream NLP tasks
- Trained for the MLM and sentence entailment tasks
- Recently some models in Hindi language have been made open sourced



- We fine-tuned BERT as per the schemes shown
- Even most basic models surpassed baselines by huge margin
- Trained till the validation loss started to increase, and then recorded results

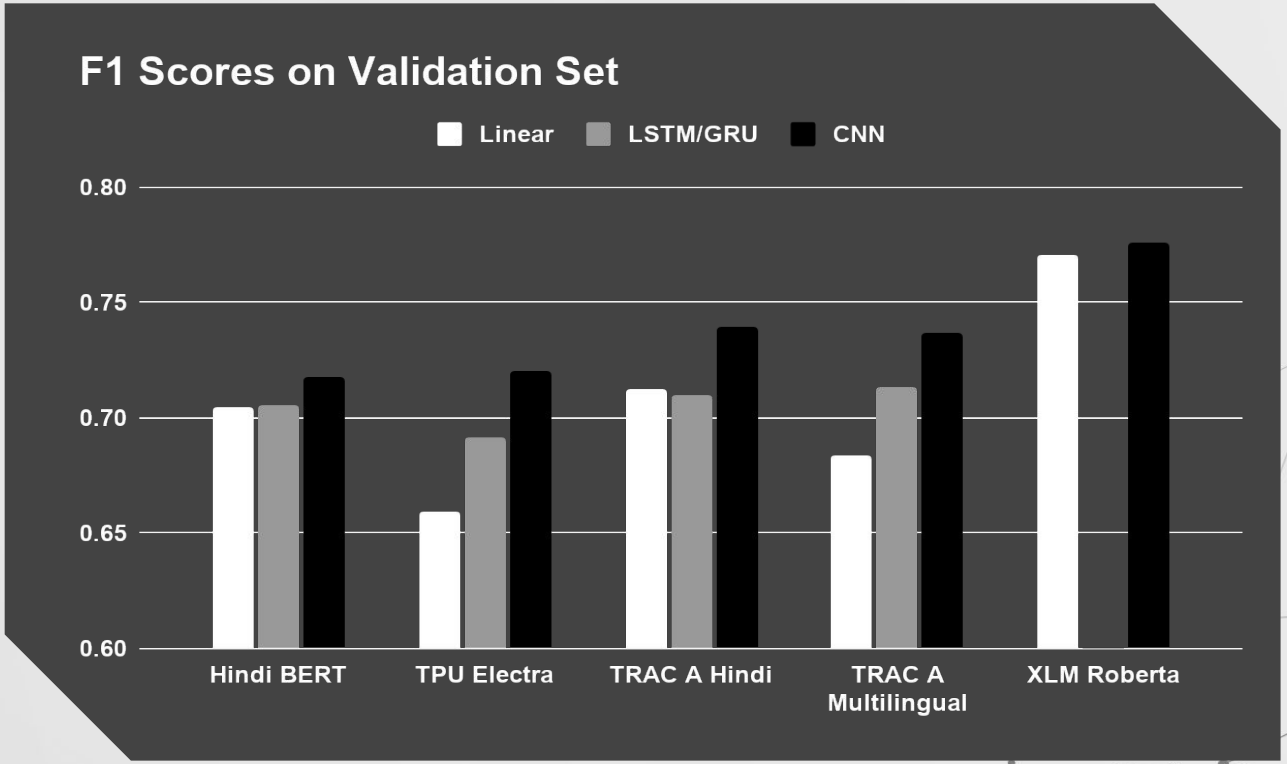
05
RESULTS

Overall Results



Note that the above results for BERT-based models are reported with CNN classifiers

Classifier Comparison

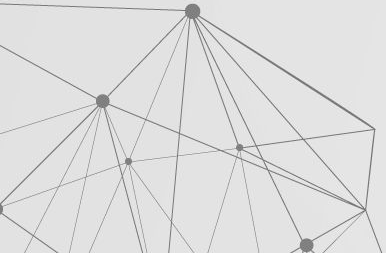


06

FUTURE WORK

Future Directions

- Expanding the Dataset
- Resolving the class imbalance
- Real Life Deployment



THANK YOU!