INDRAPRASTHA INSTITUTE of
INFORMATION TECHNOLOGY DELHI

MIDAS@IIITD
Multimodal Digital Media Analysis Lab

Did you offend me? Classification of Offensive Tweets in Hinglish Language

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, Rajiv Ratn Shah

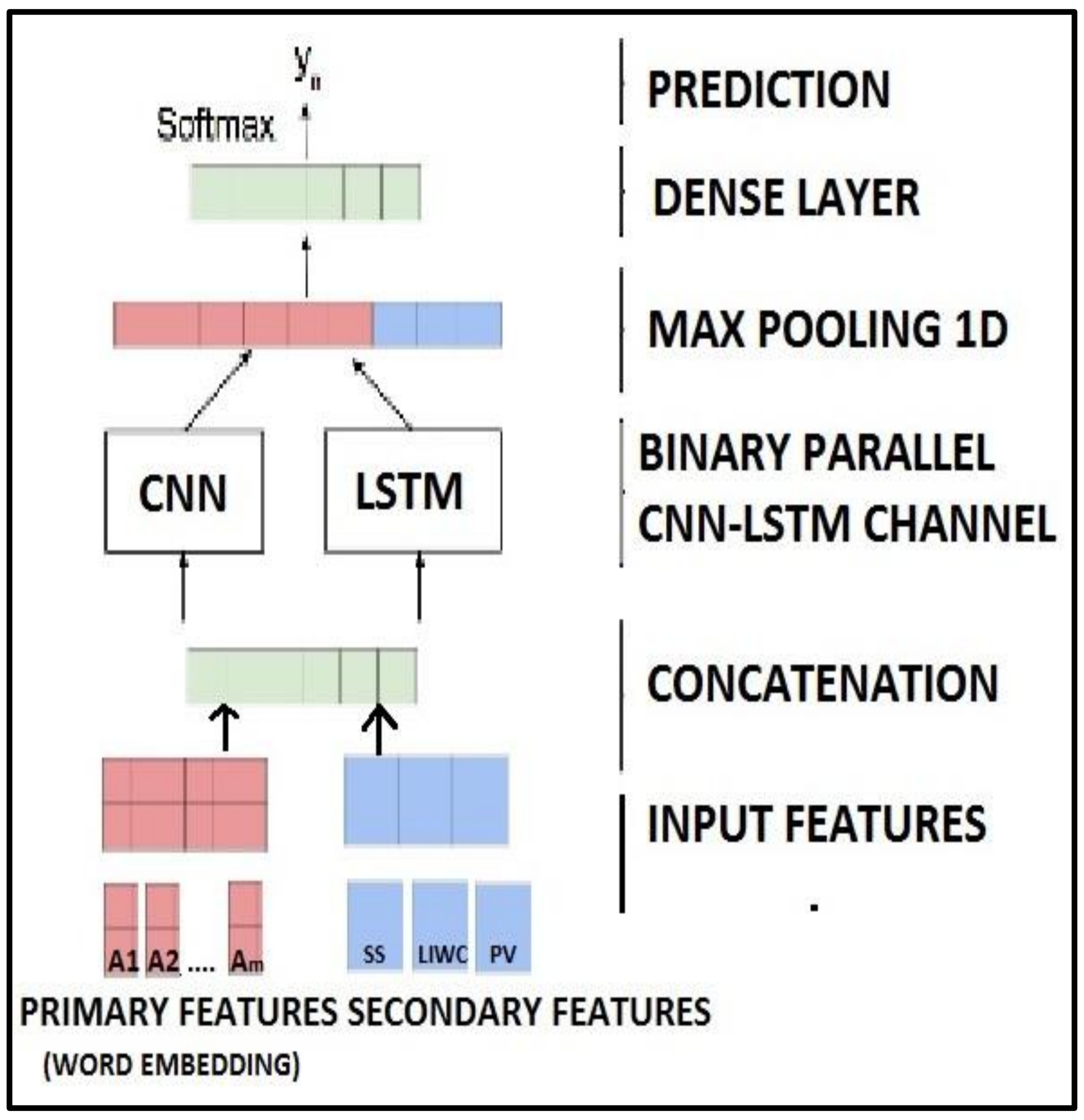
Research Goal

MOTIVATION: Due to the global reach and anonymity provided by the internet, users online have taken to using bad language and abusing people online as a license. The use of code-switched languages (e.g., Hinglish, a blend of Hindi with the English) is getting much popular on Twitter due to their ease of communication in native languages. However, spelling variations and absence of grammar rules introduce ambiguity and make it difficult to understand the text automatically.

PROBLEM STATEMENT: We aim to solve the problem of detecting offensive Hinglish tweets through the development of a transfer learning based deep learning model that analyses the input text and segregates them as: (1) *Not Offensive*, (2) *Abusive* and (3) *Hate-Inducing*.

System Architecture

- **Preprocess** data and split it as train and test
- **Primary Features** : *word embeddings* taken from-
  - Glove
  - Twitter Word2vec (Tw)
  - FastText (Ft)
- **Secondary Features** : *hierarchical contextual features* like
  - Sentiment score(SS) (SentiWordNet)
  - LIWC (Linguistic Inquiry and Word Count)
  - Profanity vector (PV)- vector denoting presence of a swear word in the tweet



MIMCT Model

- **Binary Parallel CNN-LSTM Channel**: The concatenated input features are passed through both the CNN *(3 Conv1D layers and dropout layer)* and LSTM *(Single LSTM layer and dense layer)* parallely with *Adam* optimizer and L2 regularization.
- The output is passed to a **MaxPooling 1D** layer.
- **Output:** This vector is reshaped and fed to a *softmax layer* with 3 units for each class.

System Features

- Dictionary used for Hinglish tweets had
  - Hinglish words transliterated to Devanagari Hindi, followed by translation to English
  - English translations of words from Hinglish profanity list
  - Spelling variations of various popular Hinglish words.

Label	EOT	HOT
Non-offensive	7274	1121
Abusive	4836	1765
Hate Inducing	2399	303
Total	**14509**	**3189**

Dataset Distribution

- MIMCT achieves state of the art results with the combination of using (Tw+Ft+SS+LIWC+PV) as input features outperforming baselines of SVM with TF-IDF features and transfer learning over CNN models and LSTM models.
- It uses CNN-LSTM channel pre-trained on English tweets to re-learn for the code-switched language.

Contributions

• Creation of an annotated dataset of tweets in Hindi-English code-witched language.

• Experimentation of transfer learning based neural architecture called Multi-Input Multi-Channel Transfer learning (MIMCT) model for classifying tweets in Hinglish language as abusive, hate-inducing or non-offensive.

•It suggest ways to transform code-switched Hinglish into English text for the purpose of natural language processing.

•An important contribution of the paper is to analyze informal languages on social media such as Hinglish for hate speech.