Implementing Sentiment Analysis on Code-Mixed Language (HINGLISH) using ULMFit

Rajveer Beerda, Manmeet Sethi, Kabir Kohli, Zubin Roy

School of Engineering and Applied Sciences,
Bennett University,
Plot Nos 8-11, TechZone 2, Greater Noida, Uttar Pradesh 201310

1. Introduction:

It is natural when a geographical space, having one prominent language as a means of communication undergoes alteration to utilize another as its official language, for a portmanteau to appear [1]. Such is the origin of 'Hinglish', a culmination of the official language of India, 'Hindi' and the second most spoken language, 'English'. It consists of a grammatical setup from Hindi and utilization of Roman script. It is frequently used online to ease the mode of communication. This is also known as Code mixing or code-switching, which is the use of words or phrases from two or more languages together in a single conversation/sentence.

Sentiment analysis, also known as opinion mining, is the study of people's attitudes, opinions, emotions, sentiments or appraisals about objects such as groups of pepole, project, services, politics, issues etc. and their attributes through means of computational techniques, tools and method-ds i.e., if a person has negative, neutral or positive views, opinions etc. about a particular object, which is a part of the field of Natural Language Processing (NLP) [2]. The establishment and swift nature of its growth can be attributed to the rise of social media texts such as micro-blogs and chats; with the flow of information being at an all-time high. The reason behind the proliferation of already existing large volumes of opinionated data can be credited to the fact that opinion or views are an integral part of most of human activities and are the source of influence behind an individual's behavioural pattern [3].

At the end of 2017 India had the highest Social hostility index (SHI)[4] and as the number of users of online platforms grows, the authorities are pushing for ways to make these platforms safe for all, safe from cyberbullying, safe from the spread of fake news or biased news and all these issues are being tackled by large scale NLP systems that companies like Facebook and Twitter deploy to analyse the posts on their platforms, but most of these systems are primarily built for English. With statistics supporting the fact that more than half of the posts on twitter are in a language other than English, there exists a need for researchers to pursue NLP research for other languages as well.

Though, such mixed languages pose immense challenges due the fact that several words from these languages have randomized spellings due to ambiguity resulting through various interpretations i.e., different spellings for different people and traditional NLP systems rely on monolingual resources to handle these mixed languages. The traditional way to approach this problem of Sentiment Analysis is to take the models trained on monolingual data and use them to classify this multilingual language statements into the underlying sentiment of Positive, Negative or Neutral.

Mathur et al. (2018) [5] presented a Multi-Input Multi-Channel Transfer learning-based model in conjunction with multiple feature inputs for detection of offensive Hinglish tweets. It functions using primary word embedding coupled with secondary extracted features as inputs and utilizing those inputs to train a multi-channel CNN-LSTM architecture previously trained on an English offensive tweet database acquired from CrowdFlower [6]. It outperformed transfer learning based LSTM and CNN models in the field of transfer learning based text classification.

However, Howard et al. (2018) [7] proposed a Universal language Model Fine Tuning (UMLFit) which outperformed state of the art CoVe [8], a transfer learning-based method and several other models on majority of the datasets.

In this paper, we implement ULMFit on the code-mixed language ‘Hinglish’ to attain the best accuracy and F1 score for Sentiment Analysis, using multilingual data to train the models from scratch , in addition to using some pre-trained models and training them further on this data to specifically address our problem. As textual data is best represented by long term dependencies, we start our experimentation with an RNN based model, and gradually move towards newer RNN architectures including LSTMs and finally a ULMfit model.

2. Related work:

Sentiment analysis is a method of text analysis through use of natural language processing techniques and attaining subjective information and provides quantification of people’s views or opinions on objects such as groups of people, project, services, politics, issues etc and classifies them as positive, neutral and negative.

Code mixing, also known as code-switching, is the use of words or phrases from two or more languages together in a single conversation/sentence. Code mixing is a common phenomenon especially in multilingual countries like India which has 22 official languages, with English and Hindi being the most popular.

Ain QT et al. (2017) [9] gives an overview of various approaches to sentiment analysis tasks and shows how different DL networks like Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTM), bi-directional LSTM etc. perform on a wide variety of data from different domains. It is shown that DL networks are better than the classical approaches like SVM and naive Bayes for sentiment analysis tasks. DL networks are capable of both supervised and unsupervised learning depending upon the nature of training data and perform automatic feature extraction without needing human intervention.

Zhou et al. (2015) [10] proposed learning bilingual sentiment word embeddings (BSWE) incorporating sentiment information of text into bilingual embeddings using labelled corpus and their translations.

Wang et al. (2016) [11] tried a different approach and proposed a Bilingual Attention network (BAN) aggregating both mono- and bi-lingual informative words resulting in attentive vectors which were integrated to predict emotion.

In the same year, Joshi et al. (2016) [12] proposed a sub word level LSTM instead of a word or character level LSTM for sentiment analysis on Hindi-English code-mixed text resulting in learning information regarding sentiment value of morphemes.

Zhou et al. (2016) [13] argued that s two attention-based LSTM, each hierarchically structured, could efficiently adapt sentiment information from a resource-rich language to a resource-poor language for improvement in sentiment classification at a document level.

Kapoor et al. (2018) [14] endeavoured in transfer learning, applying it to sentiment analysis and presented a hate speech classification LSTM based model classifying Hinglish text into three categories: benign, hate-inducing and abusive.

However, Ravi et al. (2016) [15] demonstrated that merging of Radial Basis Function Neural Network, term frequency-inverse document frequency-based feature representation and gain ratio-based feature selection worked best in case sentiment classification of Hinglish text.

3. Research Methodology:

The aim of this study is to efficiently predict the underlying sentiment of a given code-mixed tweet. To achieve this, we will first use an RNN with LSTM layers to set a baseline score and then use transfer learning with Fine-Tuning using ULMFit to observe the improvements.

The reason for selecting ULMFit for this task is its ability to be Fine-Tuned on a small dataset while starting out with a pre-trained model. The model is pre-trained on Wikitext-103 dataset[reference], that contains 103 million English words.

ULMFit has outperformed many other State-of-the-art methods in multiple NLP tasks but hasn't been studied in code-mixed languages. Since in Hinglish data, there are multiple English words occurring along with Hindi words, we expect ULMFit to perform well on this data.

3.1. Data Collection:

The dataset used for this research was provided by CodaLab Competition -SentiMix Hindi-English[reference]. The dataset consisted of 17000 tweets in Hinglish (Mixture of Hindi and English) with their sentiment labels[positive/negative/neutral] and word-level language tag. Table 1.1 contains a sample of raw data.

Table 1.1 Samples of raw data.

	Tweet ID	Label
meta	10466	neutral
@	O	
DelhiPolice	Hin	
sir	Eng	
local	Eng	
police	Eng	
station	Hin	
pe	Eng	
complaint	Eng	
krne	Hin	
par	Hin	
bi	Hin	
sunwai	Hin	
nhi	Hin	
hai	Hin	
..	O	
mene	Hin	
5	O	
may	Hin	
2019	O	
Ko	Hin	
complaint	Hin	
karwai	Hin	
…	O	
https	Eng	
//	O	
t	Eng	
.	O	
co	Hin	
/	O	
YUFZvNNfUz	Hin	

3.2. Data pre-processing:

The first step was to combine given tweets as they were separated at word level. Next, the word-level language tags were discarded as for this study we are going to treat Hinglish as a single language.

The tweets were then pre-processed to remove URLs, Hashtags, Mentions and were lowercased and cleaned. Table 1.2 depicts the data after being pre-processed.

3.3. Training the Models:

To set the baseline score we start by implementing an RNN model with LSTMs. The tweets were first converted into vectors by tokenizing them, which gave each unique word in the dataset a label for representation in vector space. Each tweet(vector) was padded to provide uniformity in length and the categorical labels were converted into One Hot Vector.

Table 1.2 Pre-processed data.

ID	Text	Label
10466	@DelhiPolice sir local police station pe complaint krne par bi sunwai nhi hai .. mene 5 may 2019 Ko complaint karwai	neutral
19266	Ve Maahi song from #Kesari is current favourite ! #Music #Melody #ArijitSingh @AseesKaur you are becoming my favouri	positive
33731	@ZeeNewsHindi Muslimo ka vote hasil karne wala chamcha .. Jb v bolega bura hi bolega .. Aur bolega bjp virodhi h muslim k	negative
19296	RT @YkabirYusuf #paniwala @UNICEFIndia A great move by GoI bringing water resource and drinking water together . Hope effort to sustain gai	positive
2304	Amethi me bjp ka ek pilla mara to bjp k neta aur kutti media esko aise dhika rahi thi jaise koi cm & pm mer gaya hoL	negative
21390	Chand Tare Tod Full Video Song | Yes Boss | Shahrukh Khan Juhi Chawla |... https // t . co / Q3AsQmtK9j via @YouTube	neutral

The model consisted of an embedding layer, followed by 1 D Convolution and 1D MaxPooling layers, after which LSTMs were added followed by Batch Normalization and Dense Layers. A variety of hyperparameter values for these layers were used for experimentation. The model was configured for training with binary_crossentropy as loss function and adam as an optimizer. The model's training accuracy reached 95%+ within 5 epochs, while the testing accuracy remained in the range of 48-52% after testing several times.

4. Analysis and Discussion:

Though ULMFit has been implemented in text classification in the past, through this research we have utilized it in sentiment analysis which, to the best our knowledge, is a novel approach to opinion mining in the domain of code-mixed languages.

Our hypothesis for using ULMFit was based on fact that even though it had been trained on 103 million English words, it could perform on a 'Hinglish" language set quite well due its inherent characteristic to be fine-tuned.

We trained from the scratch to create a baseline and realised that training from scratch would lead to overfitting due to lack of substantial data. The LSTM architecture was utilized to set a baseline.

Our model had a F1 score of 79.1298 on the training set and 62.60 on the testing set. Though the accuracy is not state of the art, one must remember that code-mixed languages like ‘Hinglish’ have various interpretations for different people and practitioners of opinion mining can use this model’s unique characteristic with different languages. Therefore, the applicability of ULMFit is itself quite suited to code-mixed languages.

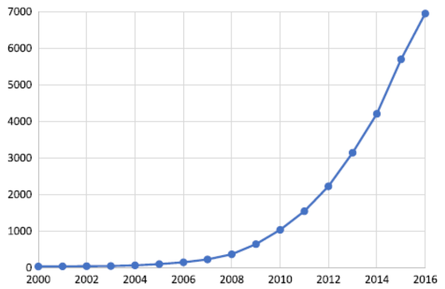

Figure 1

Figure 1. [2] shows the number of papers related to sentiment analysis and displays the rise of literature related to this field with the increase of social media platforms as a means to express opinions. Hence with the increasing need for better sentiment analysis models, companies with their R&D departments and researchers interested in this field can innovate and improve further upon our work to result in a better performing and versatile model. Our model can be utilised by social media, marketing and other sentiment analysis requiring companies or organizations to know the views of their target audience belonging to different areas with different languages and restrict illegal, abusive insulting posts or conversations.

In recent times, microblogging sites such as twitter, Tumblr etc. have become a convenient means of communication among people to express views but can also be used to spread hatred and influence the youth as done by terrorist and extremist groups. [16,17]

Additionally, it can be used by government to monitor illegal activities, hate crimes, cyberbullying and help them in coming up with the appropriate policy changes to deal with such instances.

This research can also be useful when attempting to translate code-mixed languages but would require a bigger and more detailed dataset of the languages involved.

5. Conclusion:

In this present age of ‘Internet access for all’ in conjunction with people’s need to socialize online produces huge amounts of data every single day. This data has been used by many organizations to know their target audience and reach them through personalized ads. This is what brought about the rise of sentiment analysis, but a person’s views, opinions or agendas are not always positive and this access to internet can be misused.

With this phenomenon of digitization across the world, into the hands people ranging from everyday citizens to criminals and the spread of data in different languages it has become imperative to decipher, classify, block negative and damaging aspects of it. As long as there are opinions available in the form of digitalized data, sentiment analysis would continue classifying it.

The model ULMFit provides a novel approach in sentiment analysis and fine tuning it would pave the way for its versatility in different code-mixed languages.

References:

[1]- ‘**Chutnefying English: The Phenomenon of Hinglish**’, edited by Rita Kothari, Rupert Snell.

[2]- ‘**The evolution of sentiment analysis—A review of research topics, venues, and top cited papers**’-Mika V. Mäntylä , Daniel Graziotin , Miikka Kuutila. 2018.

[3]- ‘**Deep learning for sentiment analysis: A survey**’- Lei Zhang1 | Shuai Wang2 | Bing Liu2. 2018.

[4]- https://www.pewforum.org/essay/a-closer-look-at-changing-restrictions-on-religion/

[5]- ‘**Did you offend me? Classification of Offensive Tweets in Hinglish Language**’- Puneet Mathur, Ramit Sawhney, Meghna Ayyar, Rajiv Ratn Shah. 2018.

[6]- https://www.crowdflower.com/

[7]- ‘**Universal Language Model Fine-tuning for Text Classification**’- Jeremy Howard, Sebastian Ruder. 2018.

[8]- ‘**Learned in Translation: Contextualized Word Vectors**’- Bryan McCann, James Bradbury, Caiming Xiong, Richard Socher. 2017.

[9]- ‘**Sentiment analysis using deep learning techniques: a review**’- Ain QT, Ali M, Riaz A, Noureen A, Kamran M, Hayat B, Rehman A. 2017.

[10]- ‘**Learning Bilingual Sentiment Word Embeddings for Cross-language Sentiment Classification**’- Zhou, H., Chen, L., Shi, F., & Huang, D. 2015.

[11]- ‘**A Bilingual Attention Network for Code-switched Emotion Prediction**’- Zhongqing Wang, Yue Zhang, Sophia Yat Mei Lee, Shoushan Li and Guodong Zhou. 2016.

[12]- ‘**Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text.**’- Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016.

[13]- ‘**Attention-based LSTM Network for Cross-Lingual Sentiment Classification**’- Xinjie Zhou, Xiaojun Wan and Jianguo Xiao. 2016

[14]- ‘**Mind your language: Abuse and offence detection for code-switched languages.**’- Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2018.

[15]- ‘**Sentiment classification of Hinglish text**’- Kumar Ravi and Vadlamani Ravi. 2016.

[16]- ‘**Twitter as a Corpus for Sentiment Analysis and Opinion Mining**’- Alexander Pak, Patrick Paroubek. 2010

[17]- https://www.washingtonpost.com/opinions/2019/03/16/why-social-media-terrorism-make-perfect-fit/