HATE SPEECH ON BILLBOARD HOT 100 SONGS

Niharika

1. Abstract

In today's era, the music industry has experienced massive growth and has gone through significant composition alterations. A significant portion of music audio consists of lyrics. Lyrics represent a vital constituent in a song's semantics; therefore, their analysis complements acoustic and appreciative metadata. As foul language alters our behavior negatively, it is crucial to keep people away from music that advocates hate. The hate and discriminative language of such songs mostly influence lower age groups. Therefore we need to expose and highlight such songs and create awareness.

This project has tried to detect hate speech using the BERT[1] Multilingual model. Since BERT is a pre-trained multilingual model, it enables zero-shot learning. Also, by changing the pre-trained BERT's outer layer, we can get the desired results and support transfer learning. The fine-tuning of the model is done on the OLID[2] dataset.

The metrics used for evaluating the model are the overall accuracy, f1-score, precision, and recall. The model achieved a 76% macro-F1 score, and therefore the analysis made can be called serviceable. Upon examining the most frequent words, one could easily conclude that our analysis produced the desired results.

2. Introduction

In the research paper titled "Attention Is All You Need"[10], transformers' novel architecture was proposed. Before transformers, LSTMs (Long-Short-Term-Memory) was a popular choice for sequence-to-sequence models. However, it faces long dependency issues and cannot be trained in parallel. Transformers applies the attention-mechanism to look at an input sequence and decide which other parts of the sequence are essential. Like LSTMs, transformers also have an encoder-decoder architecture. The Transformer architecture solves sequence-to-sequence tasks without implying Recurrent networks(GRUs, LSTMs, etc.). Transformers are more reliable than all the other architectures as it uses multi-head attention mechanisms and positional embeddings. Based on this architecture BERT was introduced. BERT need the encoder part of the Transformer as the goal is to generate a language model. In the paper, Google AI proposed "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." This model was able to obtain state-of-the-art results by utilizing attention and transformer architecture.

According to a medium article[8], by performing sentiment analysis on Billboard Hot 100 songs, it was observed that song lyrics are getting 1.3% negative every year. Billboard Hot 100 is a ranking list that is weekly refreshed with the most trending songs in the United States. Songs of the recent six years have been analyzed from this listing using the BERT multilingual model.

BERT, or `Bidirectional Encoder Representations from Transformers', is a novel natural language model that has inspired many NLP architectures by its training approaches, e.g., ROBERTa and ALBERT. Applying BERT comes from the fact that it obtained state-of-the-art results on eleven natural processing tasks.

Bert is pre-training of deep bidirectional transformers for language understanding conditioning on both left and right context using attention mechanisms. In this project, the

bert-base-multilingual-uncased model has been used. This model has 12-layer of transformers stacked upon each other with 768-hidden layers and 168M parameters. It is trained on the lower-cased text in the top 102 languages with the largest Wikipedias. Cased models have separate vocab entries for differently-cased words (e.g., in English, 'the' and 'The' will be different tokens). In this project, the uncased model has been used as letter casing was not required in the analysis.

The main purpose of using this model is to implement zero-shot learning and transfer learning. Zero-shot means that the BERT multilingual model was in one language and then evaluated on the foreign language without involving machine translation in either pre-training or fine-tuning. Table 1 shows the multilingual BERT performance on zero-shot performance where the model was trained on the English MultiNLI, and then evaluated on the foreign language XNLI test.

System	English	Chinese	Spanish	German	Arabic	Urdu
BERT-Zero Shot Uncased	81.4	63.8	74.3	70.5	62.1	58.3

Table 1 : Multilingual BERT on XNLI dataset

According to the paper, "How multilingual is Multilingual BERT? " (Telmo Pires, Eva Schlinger, Dan Garrette), it has been shown that multilingual BERT produces good results at zero-shot learning cross-lingual model transfer. In zero-shot task-specific annotations are done in one language are used to fine-tune the model for evaluation in another language. However, for monolingual data, i.e., if only the English MultiNLI dataset were for training and testing, English Bert performs better than multilingual BERT. English Bert produces an accuracy score of 84.2 and whereas multilingual achieved a score of 81.4. So for a single language, mBERT produces worse results. However, it is not feasible to train and maintain dozens of single-language models. The results of the paper conclude that M-BERT creates multilingual representations. Also, through the table above, we could see the results highlight that mBert performance is quite well; however, compared zero-shot should be avoided as monolingual performs quite better. This implies that for high-resource languages, the Multilingual model is somewhat worse than a single-language model.

Transfer learning has become one of the key drivers of Machine Learning success in the industry. Ever since the transformers were introduced, NLP helps solve many tasks with state of the art performance. The combination of transfer learning methods with large-scale transformer language models is becoming a standard in modern NLP. For transfer learning, the model is pre-trained on one dataset, and then knowledge of this pre-trained dataset is used to transfer knowledge to a different task, language, or domain. BERT is perhaps the most popular NLP approach to transfer and one of the representative deep transfer learning modeling architectures for NLP.

3. Related work

Although several works are done on hate speech detection, in this section, it has been tried to describe all the previous work done for hate speech detection using the novel transfer learning on BERT model.

Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis (João A. Leite, Diego F. Silva, Kalina Bontcheva, Carolina Scarton)[3]

The work done through this project shows many similarities with this paper. A new dataset has been presented in this paper called ToLD-Br (Toxic Language Dataset for Brazilian

total of 21K tweets were manually annotated by t candidates covering multiple demographic groups. BERT models were able to achieve a 76% macro-F1 score using monolingual data in the binary case and 56% in case of zero-shot learning. In this paper it has also been highlighted that monolingual approachres outperforms multilingual approaches and we require large-scale datasets for building reliable models.

Hate Speech Detection on Hostility Detection Dataset in Hindi (Ojasv Kamal, Adarsh Kumar, and Tejas Vaidhya)[9]

This paper presents a transfer learning based approach to detect hateful content in Hindi using pre-trained models. The Auxiliary IndicBert was used for the detection. The dataset used for the task of hate detection is "Hostility Detection Dataset in Hindi." However, using the monolingual model does not improve the result as only an F1-score of 0.5725 was achieved.

HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection(Binny Mathew1, Punyajoy Saha1, Seid Muhie Yimam, Chris Biemann , Pawan Goyal, Animesh Mukherjee)A[4]

A new benchmark dataset for hate speech detection has been introduced. The dataset consists of 20K posts from Gab and Twitter. Vrious state-of-the-art models are utilized and observed. It was observed that model BERT with LIME & Attn ,attained the best scores in terms of performance metrics and bias.

BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi[8]

Novel transfer learning approach was proposed based on an existing pre-trained BERT model. More specifically, investigate BERT's ability to capture hateful context within social media content by using new fine-tuning methods based on transfer learning. To evaluate proposed procedure, two publicly available datasets that have been annotated. The results show that our solution obtains considerable performance on these datasets in terms of precision and recall in comparison to existing approaches outperforms previous baselines yielding F1-score of 81% and 91%.

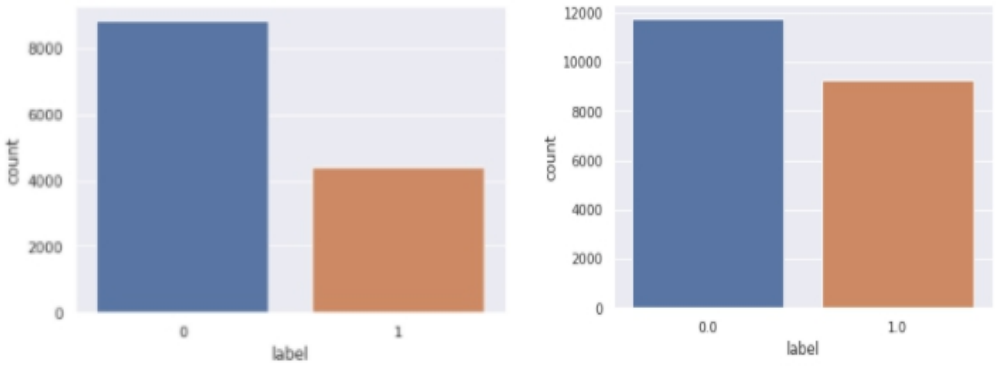
4. Dataset description

Billboard Hot 100 songs dataset has been taken from the data world. The data world is the world's largest collaborative data community, free and open to the public. The lyrics of each song has been scraped using genius.com using the LyricsGenius API Client. The model has first been trained on the OLID dataset, and the analysis has been performed on English songs. This model has also experimented on the ToLD-Br[3] dataset for evaluation of the performance on other languages.

The OLID dataset was used to train the model. The dataset comprises 14,100 annotate tweets that are classified as offensive and non-offensive. The dataset was annotated using crowdsourcing without any correction to the annotations. The gold labels were assigned, considering the agreement of three annotators. This dataset's random split has been done where 10% of this dataset has been removed for testing purposes. The dataset is not biased and therefore serves the purpose of training. Further, the dataset has been classified in more categories beyond the need of the current predictive model required. Figure 1 shows the distribution of offensive and non-offensive tweet datasets. The dataset has around 60% of the offensive tweets.

The model has been tested on the ToLD-Br[3] dataset, which is in the Brazilian Portuguese language. This dataset has 21000 tweets classified in 6 different hate speech categories, with

around 40% of the song categorized as hate speech. In this dataset, 21K tweets were manually annotated by the candidates covering multiple demographic groups.

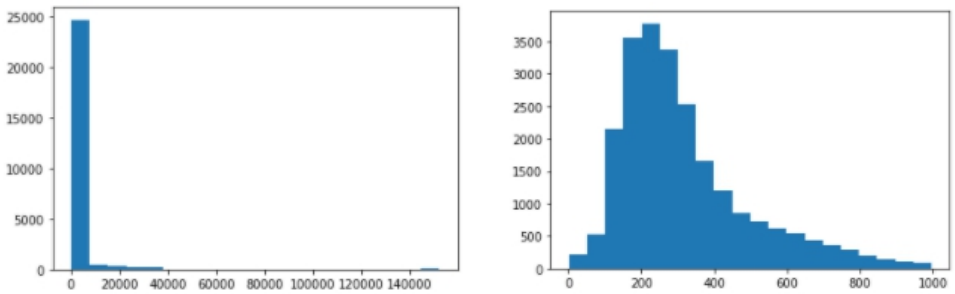


(a) (b)

Figure 1: Data distribution in (a) OLID dataset (b) ToLD-Br dataset

5. Proposed Approach

The model has been used to evaluate the Billboard Hot 100 Songs. Multilingual BERT (M-BERT) has been fine-tuned for training and testing purposes. Transformers have been installed from the Hugging Face, which gives a PyTorch interface for working with BERT. To use BERT text embeddings as input to train the text classification model, we need to tokenize the dataset's songs. Tokenization refers to breaking a sentence into individual words. To tokenize our text, we have used the BERT tokenizer. The songs have word tokens of variable length with a maximum of 152,084 tokens. BERT has a maximum limit of 512 tokens, and therefore we needed an optimum number of word tokens for tokenization. On visualizing each song's length, it was noticed that the majority of the songs had tokens less than 400. Figure 2 shows the distribution of word tokens of songs using a histogram. Songs with more than 400 tokens have been truncated, and with less than 400 have been padded to achieve equal sequence size. It is essential first to tokenize then truncate as there are instances where the word is not present in the dictionary, and the BERT tokenizer breaks the word into subwords. Attention masks have been added as we have varying lengths of a sequence that we do not want the model to attend. The special tokens, i.e., [CLS] and [SEP], have been added to every input.



(a) (b)

Figure 2: (a) Word tokens range (b) Word tokens between 0 to 1000

The dataset that is being used has more than 25000 rows. Working with such large datasets requires loading them into memory all at once. This leads to memory outage and slowing down of programs. PyTorch offers a solution for parallelizing the data loading process with automatic batching by using DataLoader. Dataloader has been used to parallelize the data loading as this boosts up the speed and saves memory.

According to BERT[6] Paper section 3: The first token of every sequence is [CLS], a special classification token. For classification tasks, the final hidden state corresponding to this token is used as the aggregate sequence representation. Therefore, the token [CLS] 's last hidden state has been extracted for the classification task. BERT-base has a hidden size of 768, which acts as an input layer. To this input layer, a layer of the hidden size of 50 has been added, and an output layer of 2 as the binary classification is being performed.

The output layer gives results in the form of logits. Logits are non-normalized prediction that the classification model generates, which is then passed to a normalization function. Logits are the values in the range R, which is used to map the probabilities in [0,1].

$$L = ln\frac{p}{1-p} \qquad p = \frac{1}{1+e^{-L}}$$

where L stands for logits and p stands for probability.

These values are used to input in the softmax function. Logits sometimes also refer to the element-wise inverse of the sigmoid function. The softmax function generates a vector of (normalized) probabilities with one value for each possible class.

Using the pre-trained BERT model for any NLP task fine-tuning model in 2 epoch produces good results.

6. Evaluation metrics

The metrics used for evaluation of the model is classification accuracy, precision, recall and f1-score represented in the following equations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$f1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

$$macro - avg\ f1 - score\ = \frac{\sum_{i=0}^{n} Precision\ of\ class\ i}{n} \tag{5}$$

$$weighted - avg\ f1 - score\ = \frac{\sum_{i=0}^{n} (Precision\ of\ class\ i * Wi)}{n} \tag{6}$$

where,

TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

False positives are crucial to know, so the F1 score has been calculated rather than the accuracy score. The optimizer used for training is AdamW[7], which fixes weight decay in Adam. The AdamW variant was proposed in Decoupled Weight Decay Regularization.

7. Result and discussion

On evaluating the mBERT model on the OLID dataset, a macro avg f1-score of 76% has been achieved.

On analyzing the trends of hate speech on streaming songs, it has been found that there has been an extensive increase in the use of hate speech in songs over the years. By graphical visualization of the past six years, the increasing percentages of songs containing hate speech have been observed. Also, while plotting the word cloud for songs categorized as hate speech content, the frequency of the most occurring words can be quickly concluded that the model could identify the songs with bad lyrics. However, according to the paper HateXplain[4], the word “nigga” is used every day by the African-American community, which is offensive but does not qualify hate speech.

word	frequency
nigga	2287
bitch	2125
shit	1976
fuck	1935
ass	553
fuckin’	524
pussy	345
smoke	326

Table 1: Words with associated frequency.

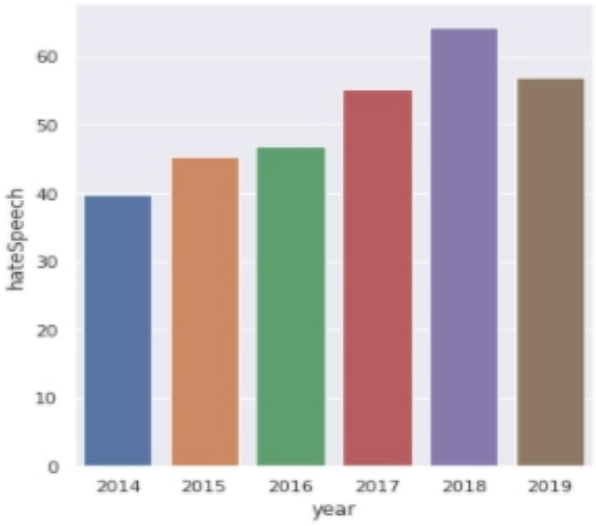


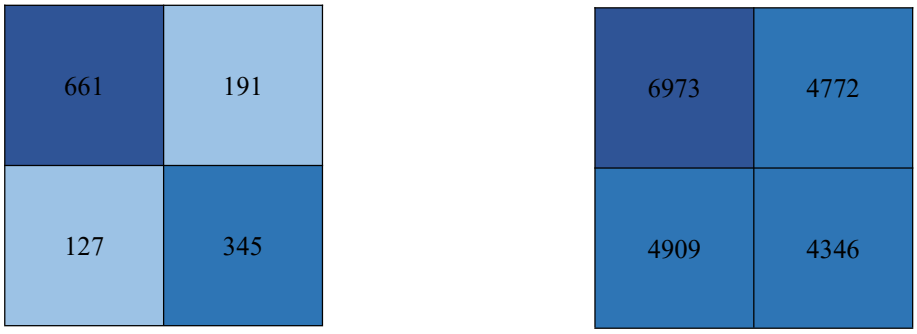
Figure 3 : Songs considered as hate speech

	precision	recall	f1-score	support
0.0	0.63	0.56	0.59	11745
1.0	0.51	0.58	0.54	9255
accuracy			0.57	21000
macro avg	0.57	0.57	0.57	21000
weighted avg	0.58	0.57	0.57	21000

Table 1: M-BERT OLID dataset

	precision	recall	f1-score	support
0	0.84	0.77	0.80	852
1	0.64	0.73	0.68	472
accuracy			0.76	1324
macro avg	0.74	0.75	0.74	1324
weighted avg	0.77	0.76	0.76	1324

Table 2 : M-BERT ToLD-Br dataset



(a) (b)

Figure 4: Confusion matrix on evaluating on (a) OLID dataset (b) ToLD-Br dataset

8. Conclusion and future scope

On dissecting the trends of hate speech on hit songs, it has been found that there has been extensive progress in the use of hate speech in songs over the years. Results reveal that around 50% of the songs are classified under the hate speech category. It shows that music is no longer about art; it is about making money. Lyricist no longer cares about the consequences of the messages their songs generate as long as they are paid. Results also highlight the significance of language-specific datasets. Zero-shot learning achieved a macro-F1 of 57%, which is comparatively less to transfer learning, which achieved a macro-F1 score of 76%. This shows that monolingual approaches still outperforms multilingual experiments.

In future work, we can make a song scoring system to categorize the type of songs people would prefer to listen to that would retain them motivated. We can categorize songs in a broader way that would help us to make a generalization based on song context which songs would be suitable for listening to.

9. References

1. Jacob Devlin and Ming - Wei Chang and Kenton Lee and Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, http://arxiv.org/abs/1810.04805

2. Zampieri, Marcos and Malmasi, Shervin and Nakov, Preslav and Rosenthal, Sara and Farra, Noura and Kumar, Ritesh, SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval), Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval), 2019

3. João A. Leite, Diego F. Silva, Kalina Bontcheva, Carolina Scarton (2020): Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. To appear in the Proceedings of the AACL-IJCNLP 2020.

4. Mathew, Binny and Saha, Punyajoy and Yimam, Seid Muhie and Biemann, Chris and Goyal, Pawan and Mukherjee, Animesh, HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection, arXiv preprint arXiv:2012.10289, 2020

5. Multilingual and Multi-Aspect Hate Speech Analysis, Ousidhoum, Nedjma and Lin, Zizheng and Zhang, Hongming and Song, Yangqiu and Yeung, Dit-Yan, Proceedings of EMNLP, 2019, Association for Computational Linguistics

6. Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization. In Proceedings of ICLR 2019.

7. Medium article by Salim Zubair

https://towardsdatascience.com/sentiment-analysis-of-all-billboard-hot-100-songs-over-time-1958-2019-3329439e7c1a

8. Mazieh Marzieh Mozafari and Reza Farahbakhsh and Noel Crespi A, BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media, https://arxiv.org/abs/1910.12574, 8th International Conference on Complex Networks and their Applications

9. Ojasv Kamal and Adarsh Kumar and Tejas Vaidhya, Hostility Detection in Hindi leveraging Pre-Trained Language Models, https://arxiv.org/pdf/2101.05494v1.pdf

10. Ashish Vaswani and Noam Shazeer and Niki Parmar and Jakob Uszkoreit and Llion Jones and Aidan N. Gomez and Lukasz Kaiser and Illia Polosukhin, Attention is all you need