Classification of Censored Tweets in Chinese Language using XLNet

Shaikh Sahil Ahmed and Anand Kumar M.
Department of Information Technology,
National Institute of Technology Karnataka,
Surathkal, Mangalore, India
sahilahmed786001@gmail.com m_anandkumar@nitk.edu.in

Abstract

In the growth of today's world and advanced technology, social media networks play a significant role in impacting human lives. Censorship is the overthrowing of speech, public transmission, or other details that play a vast role in social media. The content may be considered harmful, sensitive, or inconvenient. Authorities like institutes, governments, and other organizations conduct Censorship. This paper has implemented a model that helps classify censored and uncensored tweets as a binary classification. The paper describes submission to the Censorship shared task of the NLP4IF 2021 workshop. We used various transformer-based pre-trained models, and XLNet outputs a better accuracy among all. We fine-tuned the model for better performance and achieved a reasonable accuracy, and calculated other performance metrics.

1 Introduction

The suppression of words, images, and ideas is known as Censorship. The government or the private organization can carry Censorship based on objectionable, harmful, sensitive, or inconvenient material. There are different types of Censorship; for example, when a person uses Censorship for their work or speech, this type of Censorship is known as self-censorship. Censorship is used for many things like books, music, videos, movies, etc., for various reasons like hate speech, national security, etc. (Khurana et al., 2017). Many countries in their law provide protections against Censorship, but there is much uncertainty in determining what could be censored and what could not be censored.

However, nowadays, we know that most of the data and the information are available on the internet, so many governments strictly monitor the disturbing or objectionable content on the internet. We could not use any method other than the software like fraud censorship detection and disturbing and objectionable content monitor, which works continuously and maintains the same accuracy for monitoring this vast data size.

This paper examines the methodologies and various machine learning domains that classify the censored and uncensored tweets associated with the workshop (Shaar et al., 2021). We used multiple models such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), DeBERTa (Decoding- enhanced BERT with disentangled attention) (He et al., 2020), ELECTRA (Clark et al., 2020), and XLNet (a generic autoregressive pre-training procedure) for binary classification of the tweets. "0" says that the tweet is uncensored, and "1" says that the tweet is censored. Also, we have experimented with various phases, such as data preprocessing, tokenization, and fine-tuning for model prediction. Further, we will go through various performance metrics such as accuracy, precision, and recall. We achieved a reasonable accuracy using XLNet as compared to other models.

2 Relevant Work

(Aceto and Pescapè, 2015) proposed a source for censoring procedures and a characterization of censoring systems and studied the tools and various censorship detection platforms. They also presented a characterization plan to analyze and examine multiple censored and uncensored data. They used their results to understand current hurdles and suggested new directions in the area of censorship detection.

(Ben Jones and Gill, 2014) presented an automated system that permits continuous measurements of block pages and filters them from generated. They claimed that their system detects 95% of the block pages, recognized five filtering tools, and evaluated performance metrics and various fingerprinting methods.

(Athanasopoulos et al., 2011) presented the idea and implementation of a web-based censorship monitor named "CensMon". CensMon works automatically and does not depends on Internet users to inform censored websites. Possible censorship is distinguished from access network breakdowns, and various input streams are utilized to define the type of censored data. They showed that their model detects the censored data favourably and points filtering methodologies efficiently used by the censor.

(Niaki et al., 2019) presented ICLab used for censorship research that is known to be an internet measurement platform. It can recognize DNS manipulation where the browser initially purposes its IP address with a DNS query and TCP-packed injection. ICLabs attempts to reduce false positives and manual validation

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 136–139

through performing operations and going through all the processing levels. They plotted various graphs, planned, and calculated metrics and concluded that ICLab detects different censorship mechanisms.

3 Dataset Description

The dataset of the shared task has been built using a web scraper (Kei Yin Ng and Peng, 2020) that contains censored and uncensored tweets gathered for a duration of 4 months (August 29, 2018, to December 29, 2018). The dataset attributes contain tweets (represented by the text in the dataset) and label, where the "text" field contains the information collected in the Chinese language, and "label" contains 0's and 1's where '0' signifies the tweet as uncensored and '1' signifies as a censored tweet. The first few lines and format of the dataset is shown in Fig. 1.

text	label
据说 卡塔尔 要 退出 石油输出国组织 ， 这是 川普 退群潮 的 连锁反应 ， 也 是 世...	0
早上 没 起床 先 各 渠道 看看 与 川普 在 阿根廷 的 会面 。 留给 中国 国家足球...	1
面对 加拿大 及 美国 的 无理 行为 ， 中国 也 可以 （ 无 任何理由 ） 把 加拿大...	0
如果 特朗普 宣布 美国 对 2000 亿 中国 出口 美国 商品 征收 10% 的 关税 ...	1
两个 白日梦 什么 时候 能 像 宣传 禁毒 一样 宣传 反 强奸 什么 时候 防空警报 能...	0

Figure 1: First few lines of dataset.

The dataset comprises three sets, i.e. train, validation and test set. The train set comprises 1512 tweets, and the validation set comprises 189 tweets. The test set only comprises 189 tweets with no labels.

4 Methodology

The XLNet (Yang et al., 2019) is a transformer-based machine learning method for Natural Language Processing tasks. It is famous for a generalized autoregressive pretraining method which is one of the most significant emerging models of NLP. The XLNet consists of the recent innovations in NLP, stating the solutions and other approaches regarding language modelling. XLNet is also known for the auto-regressive language model that promotes joint predictions over a sequence of tokens on transformer design. It aims to find the possibility of a word token's overall alterations of word tokens in a sentence.

The language model comprises two stages, the pre-train phase and fine-tune phase. XLNet mainly concentrates on the pre-train phase. Permutation Language Modeling is one of the new objectives which is implemented in the pre-train phase. We used "hfl/chinese-xlnet-base" as a pre-trained model (Cui et al., 2020) for Chinese data that targets enhancing Chinese NLP resources and contributes a broad category of Chinese pre-trained model selection.

Initially, the dataset is preprocessed, and the generated tokens are given input to XLNet pre-trained model. The model trains the data over 20 epochs and further

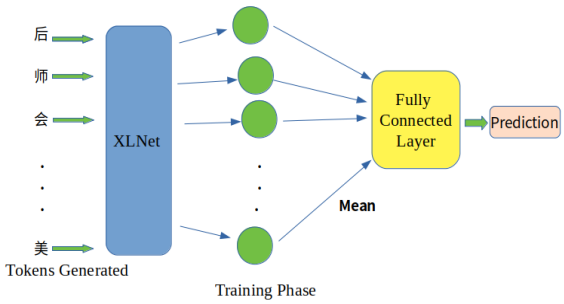

Figure 2: Architecture of XLNet.

goes through a mean pool, passing through a fully connected layer for fine-tuning and classification, and predicts the data over a given test set. Fig. 2 shows the architecture of the XLNet model.

4.1 Data Preprocessing

The dataset contains fields like "text" and "label" only, extra attribute "id" is added to the dataset for better preprocessing. Also, the noisy information from the dataset has been filtered out by using the "tweet-preprocessor" library. After preprocessing the dataset with the first few lines is shown in Fig. 3.

text	label	id
据说 卡塔尔 要 退出 石油输出国组织 ， 这是 川普 退群潮 的 连锁反应 ， 也 是 世...	0	1
早上 没 起床 先 各 渠道 看看 与 川普 在 阿根廷 的 会面 。 留给 中国 国家足球...	1	2
面对 加拿大 及 美国 的 无理 行为 ， 中国 也 可以 （ 无 任何理由 ） 把 加拿大...	0	3
如果 特朗普 宣布 美国 对 2000 亿 中国 出口 美国 商品 征收 10% 的 关税 ...	1	4
两个 白日梦 什么 时候 能 像 宣传 禁毒 一样 宣传 反 强奸 什么 时候 防空警报 能...	0	5

Figure 3: First few lines of dataset after preprocessing.

4.2 Tokenization

Tokenization breaks down a text document into a phrase, sentence, paragraph, or smaller units, such as single words. Those smaller units are said to be tokens. All this breakdown happens with the help of a tokenizer before feeding it to the model. We used "XLNetTokenizer" on the pre-trained model, as the models need tokens to be in an orderly fashion. The tokenizer imports from the "transformers" library. So, word segmentation can be said to break down a sentence into component words that are to be feed into the model.

4.3 Fine-Tuning

A pre-trained model is used to classify the text, where an encoder subnetwork is combined with a fully connected layer for prediction. Further, the tokenized training data is used to fine-tune the model weights. We have used "XLNetForSequenceClassification" for sequence classification. It consists of a linear layer on the pooled output peak. The model targets to do binary classification on the test data.

5 Experiments and Results

We have used Adam optimizer to fine-tune the pre-trained model and performed label encoding for output labels. The softmax over the logits used for prediction and the learning rate is initialized with 2e-5, and twenty epochs were used for training. After training the data with XLNet, we achieved a training accuracy of 0.99.

Models	Validation Set		
	Precision	Recall	F1-measure
BERT	0.544	0.544	0.544
DeBERTa	0.476	0.476	0.476
ELECTRA	0.624	0.624	0.624
XLNET	**0.634**	**0.634**	**0.634**

Table 1: Performance of the system on validation data.

We calculated precision, recall and F1-measure for the validation set with all the four models used in our investigation, as shown in Table 1. We got a precision of 0.634 and a recall of 0.634, which is far better than other models. Fig. 4 shows the plot for different epochs vs. validation accuracy during the training phase.

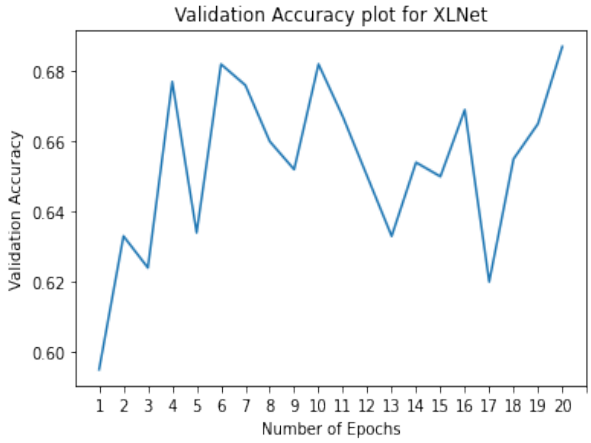

Figure 4: Validation Accuracy plot.

Class	Test Set		
	Precision	Recall	F1-Measure
0	0.61	0.73	0.66
1	0.69	0.56	0.62

Table 2: Performance of the system on test data using XLNet.

Class	Accuracy
Majority baseline	49.98
Human baseline	23.83
XLNet	0.64

Table 3: Accuracy.

Moving ahead with test data, we achieved a precision of 0.65 and recall of 0.64 using XLNet. Table 2. shows the precision, recall, and F1-Measure for test set using XLNet. Also, we found majority class baseline as 49.98 and human baseline as 23.83 as shown in Table 3.

Finally, we made one CSV file where the file contains test data tweet with label attribute. Fig. 5 shows the test data prediction, where the tweets are classified as censored and uncensored tweets.

Tweet id	Tweet	label
1	20.一位机械工程专家讲过这样一件事："文革"中，他在某地...	0
2	警惕看不到内心的人，虚荣的人，心狠的人，没有是非观的人，...	1
3	这个国在计划生育的时候对人都能下杀手，何况现在杭州政策杀狗...	1
4	罗伯特.所罗门《大问题——简明哲学导论》在我们这样一个多元的...	0
5	特朗普，输了！○特朗普，输了！	1

Figure 5: First few lines of test data after prediction.

6 Conclusion and Future Work

In the paper, we investigated various pre-trained models and achieved a reasonable accuracy for XLNET. We cleaned the dataset during preprocessing, which is further given input to the model. XLNet seems to be influential in the classification problem moving deep into censorship detection. XLNet performs better than BERT, DeBERTa, and ELECTRA having its improved training methodology, where it uses permutation language modelling predicting the tokens randomly. The future work is to examine other NLP models and fine-tune them censorship detection in other languages.

References

Giuseppe Aceto and Antonio Pescapè. 2015. Internet censorship detection: A survey. *Computer Networks*, 83.

Elias Athanasopoulos, Sotiris Ioannidis, and Andreas Sfakianakis. 2011. Censmon: A web censorship monitor. In *USENIX Workshop on Free and Open Communications on the Internet (FOCI 11)*, San Francisco, CA. USENIX Association.

Nick Feamster Ben Jones, Tzu-Wen Lee and Phillipa Gill. 2014. Automated detection and fingerprinting of censorship block pages. *Stony Brook University*.

Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Anna Feldman Kei Yin Ng and Jing Peng. 2020. Linguistic fingerprints of internet censorship: The case of sina weibo. volume 34, pages 446–453. Proceedings of the AAAI Conference on Artificial Intelligence.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2017. Natural language processing: State of the art, current trends and challenges.

Arian Akhavan Niaki, Shinyoung Cho, Zachary Weinberg, Nguyen Phong Hoang, Abbas Razaghpanah, Nicolas Christin, and Phillipa Gill. 2019. Iclab: A global, longitudinal internet censorship measurement platform.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, Alex Nikolov, Wajdi Zaghouani, Preslav Nakov, and Anna Feldman. 2021. Findings of the NLP4IF-2021 shared task on fighting the COVID-19 infodemic and censorship detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.