IndoHateSpeech: A Deep Learning Approach to Automated Multi-class Multi-label Hate Speech Detection on Informal Indonesian Corpora

Athiya Deviyani
Carnegie Mellon University
adeviyan@cs.cmu.edu

Haris Widjaja
Carnegie Mellon University
iwidjaja@@cs.cmu.edu

Abstract

Millions of Indonesians rely on social media platforms to stay updated on current political, economics, education and health-related issues. Throughout the past decade, we have observed the rapid increase in social media content, and with it, the presence of online hate speech in the Indonesian language across various social media platforms has become more prominent. To cope with the massive scale of online platforms, there is an increased need for automated hate speech detection systems. In this paper, we will extend hate speech detection to the multi-class and multi-label setting involving the target of the hate speech, category and severity level using deep learning-based models. Additionally, we present a novel tool which allows data science enthusiasts and machine learning researchers to audit their dataset for the presence of any form of hate speech.

1 Introduction

The most commonly accepted formal definition of hate speech is adapted from the Encyclopedia of the American Constitution, defined by Nockleby (1994) as *any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic*. Davidson et al. (2017) suggests that hate speech is often accompanied with abusive language such as derogatory terms, however it does not follow that the presence of a derogatory term ensures that a certain piece of text classifies as hate speech. Additionally, hate speech detection remains to be a challenging task as there exists an ambiguity on what constitutes hate speech, and the decision highly varies from person to person.

In Indonesia, abusive words involve derogatory terms towards a minority ethnic (*cina*) or religious group (*kafir*), physical disability (*cacat*), sexual orientation (*bencong*), supporters of a political group (*cebong*), traditional way of life (*kampungan*), or simply addressing someone with animal names (*anjing*, *babi*) (Wijana and Rohmadi, 2010). Home to over 1,300 ethnic groups, Indonesia promotes unity in diversity by adopting the national motto "Bhinneka Tunggal Ika", which means "It is different, [yet] it is one". The Indonesian National Commission on Human Rights (Komnas HAM, 2015) asserts that hate speech is detrimental to the unity of a diverse nation, as it has the potential to exclude, discriminate and physically harm the target of the hate speech, which are often directed to minority groups. In extreme cases, hate speech might contain violent call-to-actions that may end up in social unrest.

Conversations in social media which tend to contain hate speech or abusive language usually involve sensitive topics such as politics. They also typically arise during high-tension situations such as during a presidential campaign. According to the most recent audit performed by DataReportal (Kemp, 2021), out of 274.9 million Indonesians, there were 202.6 million active internet users as of January 2021. Out of the 202.6 million, 170.0 million were active social media users. The report stated that the most frequently used social media platforms (with respect to the percentage of internet users aged 16 to 64 that have used each platform in January 2021) are YouTube, WhatsApp, Instagram, Facebook, and Twitter. Additionally, the report also presented that 12.5% of social media users in Indonesia are within the age of 13-17. This means that teenagers are frequently exposed to social media content that may contain hate speech and abusive language. This provides further motivation for an automated system to regulate harmful content present in social media platforms.

Furthermore, there has been an increase in research that focuses on building Indonesian language models used for tasks such as sentiment analysis (Miranda et al., 2019), traffic detection (Zulfikar et al., 2019), and personality prediction

(Adi et al., 2018) which utilize data collected from Twitter. This might be problematic as the data might contain offensive language that can affect the final model and propagate hateful sentiments in the downstream task. Recently, there has been more advanced research in the Indonesian natural language processing field such as the development of IndoBERTweet (Koto et al., 2021), a pretrained language model for Indonesian Twitter which extends the monolingually-trained Indonesian BERT model (Koto et al., 2020) with additive domain-specific vocabulary in the form of Tweets. It is highly likely that without filtering harmful tweets, harmful stereotypes might be encoded in the generated word vectors. Additionally, Cahyawijaya et al. (2021) has recently introduced IndoGPT, a generative Indonesian language model which follows the GPT-2 introduced by Radford et al. (2018). Oftentimes, since datasets used to train generative language models are huge, it is not rare for hate speech to seep into the data and get trained along with the other non-offensive data. In 2020, OpenAI (Brown et al., 2020) released a paper with tests that found GPT-3 has a generally low opinion of black people and exhibits sexism and other forms of bias. This, along with the previously discussed reasons, motivates the development of a tool that is able to annotate a dataset with respect to the presence of hate speech or any other form of abusive language. However, despite the overwhelming amount of research on tackling hate speech automatically, we were unable to find any public tool which does that.

Given the above, we will highlight our contributions in this paper as the following:

- A comparison of deep learning-based multi-class multi-label hate speech detection models suitable for text obtained from Indonesian Twitter and other social networks with Indonesian Language, which does not only classify a tweet as hate speech or abusive or neither, but also the target, category, and severity level of the hate speech.
- A novel automated hate speech detection tool which allows users to upload a dataset to be audited with respect to the user-selected annotations (hatefulness and abusiveness, target, category, level). We present the tool with an intuitive user interface, allowing an easy-access to information on how the tool works as well as an example usage. For explainability purposes, we have included information of what model was used, the metrics that we obtained (accuracy and F1-score) using the model, as well as a link to the dataset that was used to train the model.

In section 2, we will go through several works regarding hate speech classification for the Indonesian language. In section 3, we will go in detail about the dataset that we have chosen for our task. We will then explain our methodology in approaching the multi-label multi-class hate speech classification task in section 4, including a description of the baseline models and deep learning models that we have used. We will discuss the corresponding results of our experiments in section 5. In section 6, we will provide a detailed explanation on how we build and deploy our hate speech detection tool. We will discuss our findings and limitations of the system in section 7. Finally, section 8 will contain our conclusion and future work.

2 Related work

A closely related work is Ibrohim and Budi (2019), which presents the multi-label dataset on which we perform our experiments. This dataset contains labels for the target, category, and level of hate speech in Indonesian Twitter. The authors conducted preliminary experiments on multi-label abusive language and hate speech detection using basic machine learning models such as Support Vector Machines (SVM), Naive Bayes (NB), Random Forest Decision Tree (RFDT), with Binary Relevance, Label Power-set and Classifier Chains; as feature representations, they use term frequency, orthography, and lexicon features. We expand upon their experiments by introducing deep learning-based models as well as tokenization and word embeddings learned by IndoBERTweet (Koto et al., 2021).

Hana et al. (2020) attempts deep learning on a "flattened" version of the dataset we use - they train a single model to simultaneously predict all 12 labels available in the dataset and found that SVM models perform better than deep learning. We instead split the original multilabel task to four separate tasks in a structured and principled way, which we postulate allows deep learning models to learn better, revealing that deep learning can indeed improve on previous performance on this task.

Rohmawati et al. (2018) provides a similar online dataset auditing tool in the form of an API and a WordPress plugin; however, their classification model is a simple binary classification (offen-

sive/not offensive). We build on their work by introducing a richer model into our auditing tool, capable of multilabel classification across four separate dimensions of hate speech (hate speech vs abusiveness, target, category, and level of hate speech).

Lastly, the aforementioned multidimensionality of our classifiers improves on an existing body of work which studies hate speech in the Indonesian language in a simple binary fashion (hate speech vs. no hate speech) (Alfina et al., 2017) (Fauzi and Yuniarti, 2018). Furthermore, we investigate the performance of deep learning models, which have not been explored extensively in existing literature for the task of hate speech detection in the Indonesian language.

3 Data

Compared to other languages, there are significantly fewer datasets available for Indonesian hate speech classification, with only one existing dataset which tackles the task of multilabel hate speech and abusive language detection. In this paper, we will be using the dataset presented by Ibrohim and Budi (2019). They have conducted a focus group discussion with the staff in an agency that is responsible for investigating cybercrimes in Indonesia to obtain the different characterization of hate speech. They have found that hate speech has a particular target, category, and level.

The target of a hate speech can either be an individual, such as a political figure, or a group that shares a specific characteristic. This characteristic can be further broken down into different categories, such as religion, race/ethnicity, physical characteristics/disabilities (which might include mental disabilities), gender/sexual orientation, or other forms of slander. For the target and category, a particular tweet can have multiple targets and multiple categories, for example it can be hateful towards an individual and a group by targeting a political figure and their supporters. Additionally, a tweet can be hateful towards different categories by targeting a gender group belonging to a particular religion.

Finally, they were also able to divide hate speech into three severity levels, from weak, moderate, to strong. The authors of the dataset suggest that weak hate speech is usually targeted towards a particular individual while strong hate speech is targeted to an individual or a group with incitement or violent call-to-actions to bring open conflict. This entails that the misclassification of strong hate speech is detrimental to preventing widespread social unrest and conflict within the country. They also have additional labels for general hatefulness and abusiveness.

The dataset consists of over 13,000 tweets which are hand-labeled with the labels described above. The usernames and URLs in the tweets have been obfuscated to preserve the privacy of the users. Table 1 shows the overall distribution between the labels. We can see that majority of the classes are imbalanced, with more items classified as FALSE or does not contain hate speech with respect to the label. Table 2 shows some examples taken from the dataset with its accompanying English translation, which may contain offensive speech.

Labels	TRUE Count	FALSE Count
INDIVIDUAL	3575	9594
GROUP	1986	11183
RELIGION	793	12376
RACE	566	12603
PHYSICAL	323	12846
GENDER	306	12863
OTHER	3740	9429
WEAK	3383	9786
MODERATE	1705	11464
STRONG	473	12696
HATEFUL	5561	7608
ABUSIVE	5043	8126

Table 1: Distribution of labels in the dataset. For example, if a Tweet contains hate speech towards an individual belonging to a particular gender, the INDIVIDUAL and GENDER label will be marked TRUE.

4 Methodology

4.1 Text preprocessing

For preprocessing, we have adopted methods from the authors of the dataset that we are using. First, we removed unnecessary characters such as endline characters and noisy tokens that are Twitter-specific such as emoticons, 'RT' (retweet), 'URL' (obfuscated URL) and 'USER' (obfuscated username). We also removed any additional symbols or trailing spaces. The authors have also provided an additional file which allows us to map various spellings of colloquial text to their formal counterpart. This is done to reduce the total number of unique words

Label group	Labels	Text
Target	INDIVIDUAL	ID: 'lihatlah Rakyat NKRI ini mending mundur aja pak **Jokowi**' EN: 'look at the people of NKRI you should just step down Mr. **Jokowi**'
	GROUP	ID: '**Budha** bunuh Muslim di Myanmar' EN: '**Buddhists** kill Muslims in Myanmar'
Category	RELIGION	ID: '**KRISTEN** membunuh di Afghanistan.' EN: '**CHRISTIANS** kill in Afghanistan.'
	RACE	ID: '**Cina** perusak bangsa!! Usir !! Stuju??' EN: '**Chinese** people destroy the nation!! Banish !! Agreed?'
	Physical	ID: 'Kalo **cacat** otak jangan dipelihara' EN: 'If you have a **brain defect**, don't keep it'
	GENDER	ID: 'Jilbab tapi akhlak bagai **pecun**' EN: 'Hijab-wearer but morals like a **prostitute**'
	OTHER	ID: 'ada teman semuanya **brengsek**' EN: 'all my friends are **assholes**'
Level	WEAK	ID: 'Dasar brengsek' EN: 'You asshole'
	MODERATE	ID: 'rezim tukang nipu. hoax. anti kritik' EN: 'fraudulent regime. Hoax. anti-criticism'
	STRONG	ID: 'lengserkan jokowi sekarang juga!!!" EN: 'remove Jokowi right now!!!'
Hatefulness and Abusiveness	HATEFUL	ID: 'Bacot la njer. Sok intelek. Kek telek' EN: 'That is bullshit. Fake intellectual. Like shit'
	ABUSIVE	ID: 'admin kok dendaman, gausa jd admin tolol' EN: 'how come the admin is vengeful, don't be an admin idiot'

Table 2: Sample sentences with the corresponding labels marked as TRUE. The word in the text that are highly indicative of each label is emboldened.

present in our corpus. We have also removed stop words based on the list provided by Tala (2003).

We have avoided lemmatization and stemming the words as inflections in the Indonesian language indicate possession and sometimes gender. For example, "steward" is "pramugar*a*" in Indonesian while "stewardess" is "pramugar*i*". Additionally, adding possessive inflection may change the hatefulness of a sentence. For example, "muka jelek" means "ugly face", while "mukanya jelek" means "their face is ugly". The latter sentence has a potential of being classified as offensive while the former doesn't, as it is not targeted towards anyone.

4.2 Model architecture

For each of the label groups target, category, level, and hatefulness and abusiveness, we have built a simple pipeline for multi-label multi-class classification using popular methods available in `scikit-learn`. The results of these models will be used as a baseline to compare our deep learning-based models with. Prior to that, the tweets are tokenized and transformed into feature vectors using `CountVectorizer` and passed through a `TfIdfTransformer` to downscale weights of words that occur in many tweets, and finally for our classification models we evaluated the Logistic Regression classifier (`LogisticRegression`), Multinomial Naive Bayes classifier (`MultinomialNB`), and Stochastic Gradient Descent Classifier (`SGDClassifier`), along with the ensemble Random Forest Classifier (`RandomForestClassifier`). The aforementioned classification methods are known to be binary classifiers, which means they cannot be used off-the-shelf for our multi-class multi-label classification task. To mitigate this issue, we use the Classifier Chains (CC) (El Kafrawy et al., 2015) method to transform our data. Classifier Chains is a machine learning method for data transformation in multi-label classification which combines the computational efficiency of the Binary Relevance method while still being able to take the label dependencies into account for classification (Read et al., 2011). For the purpose of obtaining baseline classifiers to compare our deep learning models with, we have kept the model parameters as the default settings.

For our deep learning models, we employ the IndoBERTweet tokenizer (Koto et al., 2021), which handles all text pre-processing necessary for our neural language models. This preprocessing includes splitting texts into words or subword-

level tokens, lowercasing, and transforming out-of-vocabulary words into a special "unknown" token - for full details, we direct the reader to Koto et al. (2021). We also use the embeddings learned by the IndoBERTweet model as the word embeddings for our deep learning models. IndoBERTweet is trained on a larger Indonesian tweet corpus, and we think it's likely that this pre-training will improve the deep models' performance on this smaller dataset.

We experiment with the following deep learning models, all of which employ the IndoBERTweet tokenizer and word embeddings, listed in order of complexity:

1. Single Layer Neural Network: a fully feedforward neural network which takes in a global average representation of the tokens in the input and with hidden size 20.

2. GRU: a single-layer Gated Recurrent Unit (GRU) recurrent network (Chung et al., 2014) with hidden size 64. Gated Recurrent Units are a gating mechanism in recurrent neural networks, which tend to perform better on smaller and less frequent datasets than LSTMs.

3. LSTM: a single-layer Long Short-Term Memory (LSTM) recurrent network (Hochreiter and Schmidhuber, 1997) with hidden size 64. Both GRUs and LSTMs are great at interpreting temporal or sequential information, such as a structured sentence or tweet. This is the main reason why they are most commonly used as a base architecture for text classification problems.

4. Bidirectional GRUs and LSTMs: similar to the models above, but with bidirectional recurrent connections instead, which allows the models to incorporate information from both the left and right when contextualizing the vector representation of an input token. This has been shown to improve performance for many language tasks which involve classification (Schuster and Paliwal, 1997) (Devlin et al., 2018).

5. Recurrent CNN (R-CNN): In a similar fashion to Zhang et al. (2018), this is a hybrid model consisting of 1D-convolutional layers stacked on top of three bidirectional GRU cells. The GRU preserves historical information in long text sequences, while the final convolution layer (CNN) extracts local features, leading to a better representation for the classification task. We use a hidden size of 64 for both the GRU cells and the convolution layer.

6. Transformer Encoders: a stack of three transformer encoder layers similar to the one introduced in Vaswani et al. (2017). We modify the first encoder layer to downsize the word embeddings from size 768 to 128, in order to avoid overfitting to our small dataset. All layers have 8 attention heads.

4.3 Evaluation and metrics

We will evaluate the model accuracies of both the baseline models and the deep-learning based models to provide a statistical comparison. However, given the class imbalances shown in Table 1, accuracy will not be a sufficient indicator of model performance with respect to our multi-class multi-label hate speech detection task. Therefore, we will utilize the F1-score as an additional metric, which is known to be a better indicator of model performance trained on an imbalanced dataset. The F1-score combines precision and recall into a single metric, formally defined as follows:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

In addition to understanding how the deep models fare against the baseline classification models, we will use the best model with respect to the F1-score for our proposed hate speech detection tool. Further information on the tool and design decisions can be found in the deployment section.

5 Results

5.1 Quantitative analysis

From Table 3, we can see that the best performing baseline model is the Random Forest Classifier, obtaining an average accuracy of 77.21% and an average F1 score of 0.6508 over all the tasks. This is not surprising as the Random Forest Classifier is an ensemble classifier and therefore is the most complex model among the other baseline models. The classifier performs best on the hatefulness and abusiveness classification, with an F1-score of 0.8305,

	Hatefulness and Abusiveness		Category		Target		Level		Average	
	Acc. (%)	F1	Acc. (%)	F1	Acc. (%)	F1	Acc. (%)	F1	Acc. (%)	F1
Baseline models										
LogisticRegression	77.45	0.8131	76.20	0.5734	76.96	0.5974	74.37	0.5184	76.25	0.6256
MultinomialNB	76.12	0.8172	64.09	0.2279	66.02	0.2841	62.03	0.1476	67.07	0.3692
RandomForestClassification	**77.94**	**0.8305**	**77.41**	**0.6107**	**78.40**	**0.6182**	**75.09**	**0.5439**	**77.21**	**0.6508**
SGDClassifier	74.60	0.7819	72.44	0.4743	74.45	0.528	70.58	0.4077	73.02	0.5480
Deep learning models										
Single Layer Neural Network	68.68	0.772	71.03	0.4935	70.27	0.4285	68.45	0.4025	69.61	0.5241
GRU	78.47	0.8440	76.84	0.6287	77.30	0.6723	74.18	0.5629	76.70	0.6770
LSTM	77.45	0.8316	76.65	0.6293	76.65	0.6519	75.13	0.6009	76.47	0.6784
Bidirectional GRU	77.94	0.8389	76.73	0.6378	77.98	0.677	74.94	0.6116	76.90	0.6913
Bidirectional LSTM	78.74	0.8432	76.61	0.6268	78.21	0.6768	74.64	0.5959	77.05	0.6857
R-CNN (with Bi-GRU)	**81.74**	**0.8691**	**77.15**	**0.6150**	**79.42**	**0.7131**	**75.70**	**0.6168**	**78.50**	**0.7035**
Transformer	79.31	0.8510	75.21	0.6416	77.03	0.6874	74.91	0.6156	76.62	0.6989

Table 3: Performance metrics of the various models trained to perform the task of multi-label multi-class hate speech classification

and performs the worst on the level classification, with an F1-score of only 0.5439. This trend is also shared with all the other baseline models.

Between all the deep learning models, the Recurrent CNN with Bidirectional GRU (R-CNN) performs best, achieving an average accuracy of 78.50% and an average F1 score of 0.7035 over all the tasks. This presents an improvement of 5% (absolute) over the best F1 score among the basic models. Additionally, we can see that all the deep learning models other than the Single Layer Neural Network has achieved an average F1-score that is higher than the best baseline model. This indicates that deep learning-based methods can be employed to improve the multi-class multi-label hate speech classification task on Twitter data in the Indonesian language.

The R-CNN model outperforms the more complex Transformer-based models, even though the Transformer models outperform almost all other models in our experiments. This is surprising as Transformer-based models have been dominating the state-of-the-art for most Natural Language Processing tasks, including text classification (Yang et al., 2019). We speculate that this is because a simpler model generalizes better for a smaller dataset, and our experiments agree with this tendency. Furthermore, Wolf et al. (2020) confirms that the Transformer architecture is particularly conducive to pretraining on large text corpora, while our training dataset consists of around 10,000 entries.

5.2 Qualitative analysis

We decided to investigate further by performing a detailed error analysis on non-hateful text that were misclassified as hateful by our best model. At first, we hypothesize that most of the misclassified non-hateful text contains instances of abusive language which are used in a non-hateful context. An example would be the sentence "saya suka *anjing*", which means "I like *dogs*" in English. We suspect that since the word "anjing", although bearing the literal meaning of "dog", is often used as an insult towards someone, our model is most likely to classify a text which contains the word as hate speech. This is an observation made by Davidson et al. (2017), where they noted that not all tweets containing offensive language in the form of profanity are hateful or offensive per se, and the task of determining which tweet is which is not trivial.

However, after a more careful examination of the misclassified tweets, we observe that the tweets seem to share several words in common. The most noticeable ones are the words "presiden", which means president, and "Jokowi", which is the name of the Indonesian president during the time when the dataset was collected and also a presidential candidate for the 2019 elections. We speculate that since the authors of our dataset (Ibrohim and Budi, 2019) collected tweets during the time of the 2019 presidential campaign in Indonesia, there would be a large number of tweets containing the names of the presidential candidates, possibly used in a hateful context. To confirm this hypothesis, we decided to plot the counts of the top 15 words in tweets in the training set which are labeled as hate speech alongisde the counts of the top 15 words in tweets in the test set which are misclassified as hate speech. From the plot in Figure 1, we can observe that the word "Jokowi" occurs more than 500 times in tweets originally labeled as hate speech, and consequently appears as the 8^{th} most frequent word in non-hateful tweets which are misclassified as

hate speech. This trend is also observed for the word "presiden".

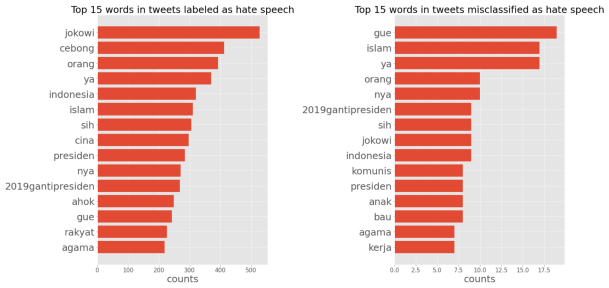

Figure 1: Counts of top 15 words in tweets in the training set which are labeled as hate speech (left) and counts of top 15 words in tweets in the test set which are misclassified as hate speech (right)

Although often true, we found out that the frequency of a word corresponding to a particular individual or group occuring in a tweet originally labeled as hate speech does not directly mean that they are often the target of hate speech. For example, we can see that the word 'Islam' occurs frequently in both tweets actually labeled as hate speech and tweets misclassified as hate speech. After taking a closer look at the tweets containing the word 'Islam', we can see that they are hateful towards entities which are anti-Islamic, not Islam itself. For example, the tweet 'rezim represif anti Islam' means 'anti-Islamic repressive regime', which criticizes the political regime to be anti-Islamic. These kind of occurences of tweets along with other ambiguities continue to prove that hate speech classification remains to be a challenging task, even for complex deep-learning based models.

6 Deployment

Given the results in Table 3, we chose to use the R-CNN with bidirectional GRU layers as the base model for deployment. After evaluating several tools and platforms, we have decided to deploy our model and build our system using the web framework, Streamlit. Singh (2021) argues that the biggest advantage of using Streamlit is that it allows you to use HTML code within the application Python file without having to have separate templates and CSS formatting for the front-end user interface. Additionally, it comes with a pre-built front-end styling which supports writing text in the markdown format and readily-available widgets such as a text input and file uploader functionalities. The resulting web application is also optimizable for different devices across different platforms and operating systems.

On the front-end, we provide a short description on what the app does, what the various annotations mean and what the sub-labels are for each label group (hatefulness and abusiveness, target, category, level). We also provide information on what model is currently being deployed, which dataset was used for training and the relevant metrics corresponding to the model.

For users who are interested in the tool without an available dataset, we have provided a text box where a user can type a sentence and receive annotation results consisting of all the labels in the label groups as a demo on how the tool works. On the other hand, if a user has a dataset ready which contains Indonesian text, we have provided specific instructions on how to receive an annotated version of their dataset. We have also provided an example dataset so that they could format their existing dataset accordingly. After the dataset is successfully uploaded to the web application, we will show them a preview of the dataset so they will be able to confirm that they have uploaded the correct dataset. Then, the user will need to specify which column contains the text that will be analyzed. Paying attention to privacy concerns, we will not be storing the uploaded dataset at all, and all of the preprocessing and inference are done on-the-fly. We have made this statement clear on our front-end.

Then, the users will be able to select which label groups would they like their dataset to be annotated with using a checkbox. Finally, after clicking on the 'Audit' button, they will be able to preview the annotated version of their dataset and download it as a CSV file. Instead of providing binary values for each label, we have provided the score (a float between 0 and 1) output by the model instead. This is because we would like to give the user the freedom of choosing various thresholds when filtering out hateful text from their dataset.

The files relevant to the web application are hosted on a public Github repository. Streamlit allows a functionality to automatically deploy an application straight from a Github repository, so we decided to move forward with this implementation. Additionally, Streamlit will also provide a link to the Github repository on the front-end, which further supports our aim to make the tool as transparent as possible. Screenshots of the web application showing the previously mentioned functionalities

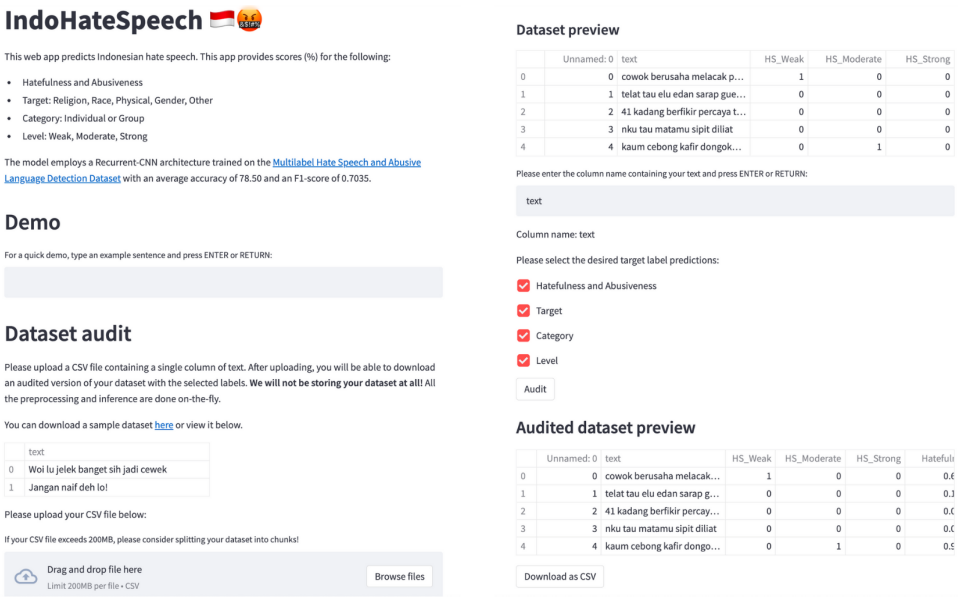

Figure 2: Screenshots of the IndoHateSpeech web application

is available in Figure 2.

7 Discussion

One of the main limitations of our system is that even though our proposed Recurrent CNN with Bidirectional GRU (R-CNN) has achieved an average accuracy and F1-score that surpasses all the baseline models, an accuracy of 78.50% entails that there is more than a 20% chance that a hateful text sequence will be misclassified as non-hateful. Thus, when a user uses our tool to audit their dataset, they will not be able to filter out all the hateful entries completely. However, we hope that by providing the prediction scores instead of the binary label, the user will be able to choose their own thresholds for filtering out hateful text such that they are able to maintain as much of their original data as possible while also keeping their dataset clean of offensive speech.

Additionally, while performing qualitative analysis, we have noticed that the presence of specific words makes the model more probable of classifying a text as hateful. This might be an indication of the occurence of shortcut learning, where our model attempts to identify the simplest solution or a 'shortcut' to solve a given problem. Future work should explore methods to mitigate the effect of the existence of such words to the final prediction of the model, such as how Kumar et al. (2019) proposed a method to demote latent confounds for the task of native language identification. For example, if the text contains the word 'Sweden', the classifier will most likely predict that the text is Swedish even though it is not. Given the nature of our task and our observations on the presence of confounding variables, we believe that this method has the potential to improve the overall model performance.

8 Conclusion

In this paper, we propose an automated hate speech detection tool powered by a Recurrent CNN with Bidirectional GRUs which achieve an average accuracy of 78.50% and an F1-score of 0.7035 across all classification tasks involving the hatefulness and abusiveness of a text, as well as determining the category, target, and severity level of the hate speech. We hope that our tool will allow researchers to train language models on datasets free of hate speech as well as provide a solution to create an online space which is safe and peaceful for Indonesians.

References

Gabriel Yakub NN Adi, Michael Harley Tandio, Veronica Ong, and Derwin Suhartono. 2018. Optimization for automatic personality recognition on twitter in bahasa indonesia. *Procedia Computer Science*, 135:473–480.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Com-*

puter Science and Information Systems (ICACSIS), pages 233–238. IEEE.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, et al. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. *arXiv preprint arXiv:2104.08200*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Passent El Kafrawy, Amr Mausad, and Heba Esmail. 2015. Experimental comparison of methods for multi-label classification in different application domains. *International Journal of Computer Applications*, 114(19):1–9.

M Ali Fauzi and Anny Yuniarti. 2018. Ensemble method for indonesian twitter hate speech detection. *Indonesian Journal of Electrical Engineering and Computer Science*, 11(1):294–299.

Karimah Mutisari Hana, Said Al Faraby, Arif Bramantoro, et al. 2020. Multi-label classification of indonesian hate speech on twitter using support vector machines. In *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pages 1–7. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57.

Simon Kemp. 2021. Digital 2021: Indonesia. *DataReportal*.

Komnas HAM. 2015. Buku saku penanganan ujaran kebencian (hate speech).

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization. *arXiv preprint arXiv:2109.04607*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian NLP. *CoRR*, abs/2011.00677.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. *CoRR*, abs/1909.00453.

Eka Miranda, Mediana Aryuni, Ricky Hariyanto, and Edwin Satya Surya. 2019. Sentiment analysis using sentiwordnet and machine learning approach (indonesia general election opinion from the twitter content). In *2019 International Conference on Information Management and Technology (ICIMTech)*, volume 1, pages 62–67. IEEE.

John T Nockleby. 1994. Hate speech in context: The case of verbal threats. *Buff. L. Rev.*, 42:653.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359.

Umu Amanah Nur Rohmawati, Sari Widya Sihwi, and Denis Eka Cahyani. 2018. Semar: An interface for indonesian hate speech detection using machine learning. In *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 646–651. IEEE.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Pramod Singh. 2021. *Machine Learning Deployment as a Web Service*, pages 67–90. Apress, Berkeley, CA.

Fadillah Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

I Dewa Putu Wijana and Muhammad Rohmadi. 2010. Sosiolinguistik: Kajian, teori, dan analisis.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Jiarui Zhang, Yingxiang Li, Juan Tian, and Tongyan Li. 2018. Lstm-cnn hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1675–1680. IEEE.

Muhammad Taufiq Zulfikar et al. 2019. Detection traffic congestion based on twitter data using machine learning. *Procedia Computer Science*, 157:118–124.