Automatic Indonesian Hoax News Detection Using BERT

Reza Fuad Rachmadi
Department of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Sepuluh Nopember
Institute of Technology
Surabaya, Indonesia 60111
fuad@te.its.ac.id

Mauridhi Hery Purnomo
Department of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Sepuluh Nopember
Institute of Technology
Surabaya, Indonesia 60111
hery@ee.its.ac.id

Aufa Nabil Amiri
Department of Computer Engineering
Faculty of Intelligent Electrical
and Informatics Technology
Sepuluh Nopember
Institute of Technology
Surabaya, Indonesia 60111
aufa.17072@mhs.its.ac.id

***Abstrak*—Fake news or called hoax, is one of the things that still plaguing Indonesia. Even more so, with the rise of the social media, a fake news can spread wider and faster than ever before. Worse, Indonesian people have quite a high tendencies to share fake news. That is why, we are in a dire need of a method to detect fake news. This research is using BERT to automatically classifiy whether a news is a hoax or not. From a raw text, we applied a tokenization process before inputting the text to the BERT. Next, the pooled output of the BERT is being used as the input for Linear Regression, a tested-and-true method for classifying task. The output of the Linear Regression is then being used as a way to determine whether a news is a hoax or not. The purpose of this research is to create a machine learning model to help the people to determine whether a text is a fake news or not. The result of this research is a model to classify a hoax text with 89% in accuracy.**

***Kata kunci*—BERT, Hoax, Fake News Classification, Linear Regression**

I. Introduction

News is a report or a factual story, designed to be the fastest, has a good way of describing problems, and is just by nature to all problem in which it is choose to be published [1]. News also has a very important role in the public, not only because it is a good way to attain a new information, but also to broaden one knowledge.

Hoax or fake news is a way or method to try to deceive people so they believed something that is can't be considered correct and those incorrect things is more often than not is something only a mad-man would believe [2]. Not only reading a fake news will cost you your knowledge, hoax can have many other effects, ranging from the loss of reputation, money, up to even death threat.

Based on the data that we got from the Ministry of Communication and Informatics, there are a total of 5156 hoaxes that have been found only from a short range of August 2018 to March 2020. From January 2020 to March 2020, there are as many as 959 fake news that have been found [3]. Still based on the very same source, at Juny 2020, there are dozens of new hoaxes have been discovered every single day [4].

Nowadays, there are high chance everyone has a few social media account rather than those that are not. This in turn, has quite an effect on the spreading of the fake news, with those hoaxes spreading far and wide with speed never seen before. Based on the survey conducted by Khan and Idris, there are more than half of Indonesian people has a high tendency to share news links without feeling the need to do any validation of said news beforehand [5]. Another survey with similar topic conducted by Kunto with 480 respondence at East Java, shows that around 30% of the total of the responder has a tendency to share a fake news from mild to severe [6]. From those studies, it is safe to assume that Indonesian people in general, has a high tendency to share fake news through their social media accounts.

Neural networks is one of the many branches of machine learning study in which it is applying neurons, just like those that is usually found in human brain structure. Those neurons is used by neural network to process data which in turn resulting in an output. One of the newest things in neural network branch is a method called Bi-Directional Encoder Representations from Transformers or BERT in short. BERT is a method to get a context from a raw text in which it is inputted.

There are many previous works on this automatic hoax detection topic that have been done by other researchers in the past. Aggarway et al. has done an extensive research to see the difference between BERT, XGBoost and LSTM to classify fake news from english sources. Based on that research, turn out BERT has quite an edge to detect hoaxes compared to the other two method [7]. Another researcher under the name Bahad et al. has done another research to see which one is better between CNN, RNN, uni-directional LSTM RNN and bi-directional LSTM RNN when used also to classifying fake news. The result shows that LSTM coupled with attention-span, whether it is a uni-directional or bi-directional one, has quite a high accuracy compared to the other method like CNN or RNN [8]. From either of those two researchs, it can be concluded that if an algorithm is able to "remember" or know the context of the text, it will most likely has a higher accuracy if compared to the other non-"remember" approach.

But, if we are talking about Indonesia news detection state nowadays, there are not that many researcher has been doing that topic. There has been a research, done by Prasetijo et al. , that try to use SVM and SGD to detect Indonesian hoax news and resulting in a model with the accuracy of 85% [9]. Another research by Rahutomo et al. on the same topic but using naive bayes as the method, has been succesfully attain 80% accuracy on the same task [10].

The purpose of this research is to develop a model to automatically detect Indonesian hoax news by using BERT. The reason being is that by using BERT, hopefully, there will be an increase in efficiency and accuracy of Indonesian hoax

news detection.

II. Design and Implementation

This research is explaining about the implementation one of the branch of deep learning studies with the aim to automatically detect Indonesian hoax news by leveraging BERT method. This detection method is trained by using a combination of dataset from https://data.mendeley.com/datasets/p3hfgr5j3m/1 and dataset that we made ourself for this paper alone by using web crawling technology. Picture 1 is the outline of this research in a nutshell.

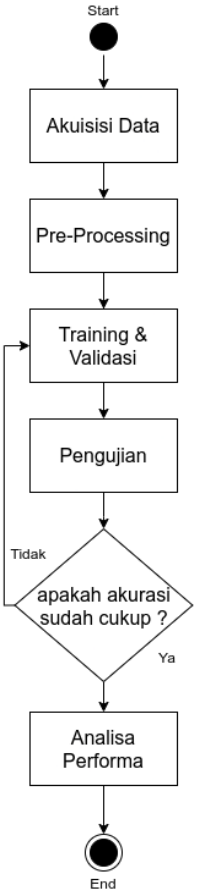


Fig. 1. This research method in a nutshell.

A. Material and Tools Specification

The dataset that is being used in this research is a dataset originated from https://data.mendeley.com/datasets/p3hfgr5j3m/1 coupled with our own made dataset in which we have create it using web crawling technology. Both of these dataset combined, is resulting in total of 1621 data with the exact details can be seen at table II. Meanwhile, table I is the starting point of our dataset which we gotten from https://data.mendeley.com/datasets/p3hfgr5j3m/1 alone.

Each of these dataset is containing the content of the news along with its label which can be either "Valid" or "Hoaks". We took the news from accredited and verified news sources for the valid news, while on the other side, we took all of the hoax news mostly from https://turnbackhoax.id, a website that contains the list of user reported hoax news from many sources.

TABLE I
Total of news from data.mendeley.com

Label	Total Data
Hoaks	228
Valid	372
Total	**600**

TABLE II
Total of training dataset

Label	Total Data
Hoaks	885
Valid	736
Total	**1621**

B. Akuisisi Data

Because the dataset that we get from https://data.mendeley.com/datasets/p3hfgr5j3m/1 feel severely lacking for our purpose because it only consist of 600 data, and because there are no web crawling which outputting its result into a convenient CSV file from Indonesian news sites, we took on our hand a task to create a webcrawling program to take news content from many Indonesian news sites, those sites included but not limited to liputan6.com, detik.com, tempo.com and others. As all of those sites is rightfully accredited and verified by the government, it is used for our valid news dataset. Our hoax news site however, only has one source from turnbackhoax.id, this is mainly because said site has quite an active forum behind it in which lots of people can report their finding of hoax text, seen and checked by lots of other people, before lastly, will be uploaded to the turnbackhoax.id site. But, the biggest factor in choosing that site compare to others is mainly because turnbackhoax.id wrote the original hoaks text in their website, this coupled with the fact that their website has some kind of structure into it has shorten our task significantly. For this research, the webcrawling process has took news from varied dates, ranging from April 2018 as the oldest to April 2021.

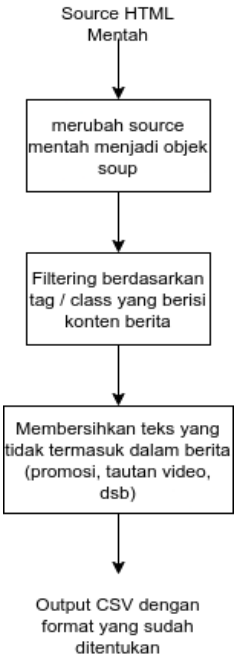


Fig. 2. Garis besar alur program *web crawl*.

TABLE III
DATASET SAMPLE

news	tagging
Wakil Gubernur DKI Jakarta Sandiaga Uno menargetkan pengerjaan tahap awal Stadion BMW dilakukan pada Oktober. Stadion ini diperuntukkan bagi klub Persija....	Valid
"Komisi II bersama KPU dan Bawaslu masih membahas ketentuan wajib cuti bagi petahana presiden yang maju Pilpres 2019. Mekanisme pengambilan.....	Valid
Jaksa penuntut Ulumum (JPU) pada Komisi Pemberantasan Korupsi (KPK) mencecar Pejabat Pembuat Komitmen (PPK) reguler pada Direktorat Perlindungan Sosial Korban Bencana Sosial Kemensos Victorious Saut Hamonangan Siahaan soal...	Valid
"Halo Kak! Aku Winda Dari Team Giveaway BAIM WONG Anda Memenangkan Hadiah Uang 100Jt dari kami info klik: https://wa.me/+6285796306857"	Hoax
"Apa yang terjadi dengan hewan dalam penelitian? Teknologi ini telah dicoba pada hewan, dan pada hewan penelitian yang dilakukan, semua hewan mati , tidak langsung dari suntikan...	Hoax
"Kadrun istilah dr PKI alias KOMUNIS ditujukan buat islam. Kl mau jd komunis pake aja istilah kadrun buat umat islam. Auto lsg Komunis"	Hoax

Picture 2 is the outline flow of the webcrawling program. Starting with inputting raw HTML code into the program, changing said code into an easier-to-process objects, get the news teks and do some post-cleaning on the text, lastly, create a .CSV file to store all of the obtained news text with the appropriate format.

```
...
<div class="detail__body␣itp_bodycontent_wrapper">
<div class="detail__body-text␣itp_bodycontent">

<strong>Jakarta</strong> - Koalisi <a href="https://
detik.com/tag/jokowi" target="_blank">Jokowi</a>
sedang menyusun visi-misi jagoannya. Setelah
menerima masukan dari <a href="https://detik.com/
tag/muhammadiyah" target="_blank"> Muhammadiyah</a>,
 ...
Dan kita pun membuka diri untuk menerima
masukan untuk penyempurnaan,"␣imbuhnya.<br><br><!--
s:parallaxindetail--><div␣class="clearfix"></div><style>
...
```

Listing 1. Penggalan Kode Sumber HTML detik.com.

Firstly, we need to determine tag or class of the HTML code for our first filter. If we look into listing 1 as a reference, we can see `detail__bodytext` class is the one that containing our desired news text. We filtered that class by inputting the class name into the appropriate parameter.

More often than not, our filtering result will contain some garbage or unrelated teks resulting in the need to refine it further by post-clean it after the filter process. Usually, those teks is writer or editorial notes, ad, or related news links which we don't need at all.

Finally, is outputting all of the acquired news teks as a .CSV. There are no particular reason on the article of why we chose CSV format compared to other famous format unless CSV format is easier to use in our training program and because it is an open format that can be open by nearly any spreadsheet program.

As the general interface and improving user experience for our webcrawling software, we use a .json format file to configure what news sources that we want to get, how much is it, and when is it. All of those configuration will be processed by the program and the program will take the news in accordance with said configuration.

III. LOREM IPSUM

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

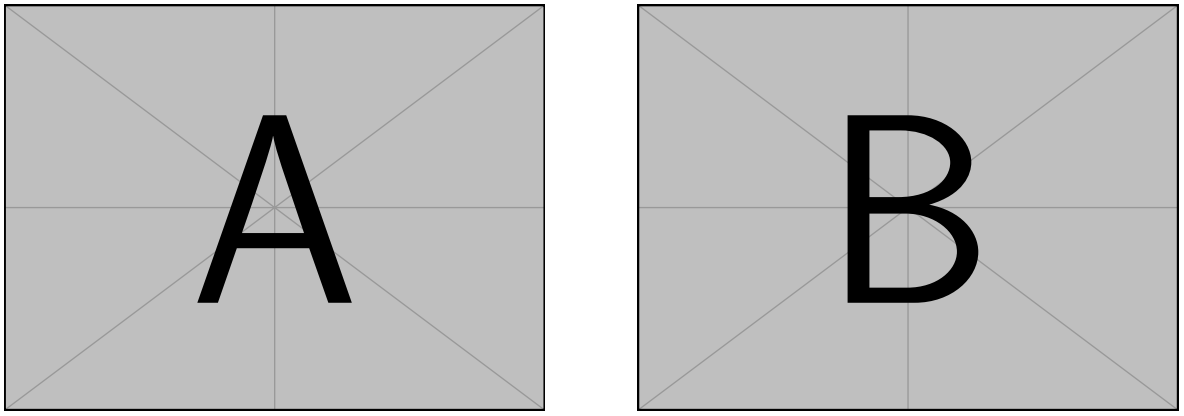
Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

```
def apakahBilanganPrima(nilai):
    if nilai > 1:
        for i in range(2,nilai):
            if (nilai % i) == 0:
                return False
        else:
            return True
    else:
        return False
```

Listing 2. Program perhitungan bilangan prima.

Nulla non mauris vitae wisi posuere convallis. Sed eu nulla nec eros scelerisque pharetra. Nullam varius. Etiam dignissim elementum metus. Vestibulum faucibus, metus sit amet mattis rhoncus, sapien dui laoreet odio, nec ultricies nibh augue a enim. Fusce in ligula. Quisque at magna et nulla commodo consequat. Proin accumsan imperdiet sem. Nunc porta. Donec feugiat mi at justo. Phasellus facilisis ipsum quis ante. In ac elit eget ipsum pharetra faucibus. Maecenas viverra nulla in massa.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede.



(a) Hasil A (b) Hasil B

Fig. 3. Contoh input beberapa gambar.

Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

IV. KESIMPULAN

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

REFERENCES

[1] Rani and N. L. R. Maha, "Persepsi jurnalis dan praktisi humas terhadap nilai berita," 2013.

[2] Wikipedia, "Berita bohong," diakses 27 November 2020. [Online]. Available: https://id.wikipedia.org/wiki/Berita_bohong

[3] K. Kominfo, "Temuan isu hoaks," 03 2020. [Online]. Available: https://eppid.kominfo.go.id/storage/uploads/2_12_Data_Statistik_Hoax__Agustus_2018_-_31_Maret_2020.pdf

[4] ——, "Laporan isu hoax juni 2020," 07 2020. [Online]. Available: https://eppid.kominfo.go.id/storage/uploads/2_31_Laporan_Isu_Hoaks_Bulan_Juni_2020.pdf

[5] M. L. Khan and I. Idris, "Recognize misinformation and verify before sharing: A reasoned action and information literacy perspective," *Behaviour and Information Technology*, 01 2019.

[6] K. Wibowo, D. Rahmawan, and E. Maryani, "Penelitian di indonesia: umur tidak mempengaruhi kecenderungan orang menyebarkan hoaks," 2019, diakses 27 November 2020. [Online]. Available: https://theconversation.com/penelitian-di-indonesia-umur-tidak-mempengaruhi-kecenderungan-orang-menyebark

[7] A. Aggarwal, A. Chauhan, D. Kumar, M. Mittal, and S. Verma, "Classification of fake news by fine-tuning deep bidirectional transformers based language model," p. 163973, 04 2020.

[8] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional lstm-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 02 2020.

[9] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on indonesian news sites based on text classification using svm and sgd," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 10 2017.

[10] F. Rahutomo, I. Pratiwi, and D. Ramadhani, "Eksperimen naïve bayes pada deteksi berita hoax berbahasa indonesia," *JURNAL PENELITIAN KOMUNIKASI DAN OPINI PUBLIK*, vol. 23, 07 2019.