Fake news classification for Indonesian news using Extreme Gradient Boosting (XGBoost)

J P Haumahu*, S D H Permana and Y Yaddarabullah
Universitas Trilogi, Kalibata, Pancoran, Jakarta Selatan, Indonesia, 12760

*johnhaumahu@trilogi.ac.id

Abstract. Fake news or commonly known as a hoax has become one of the most visible cybercrime. Hoax news dissemination harms the social community, such as raising hatred towards something both individuals and groups. This paper is to classify amongst hoaxes and valid news utilizing Extreme Gradient Boosting (XGBoost) method in this research based on Indonesian news. The dataset used is Indonesian news about Indonesia itself and the world from 2015 to early 2020. The study used 500 news data including 250 valid news and 250 hoax news, divided into 80% training data and 20% test data. The result of this study shows that the machine learning model created using XGBoost has an accuracy value of 89%, with the precision value of 90% and recall value 80%.

1. Introduction
The increasing number of internet users has an impact on the increasing spread of deception. Social media is one means for certain parties to take action to spread false information. The lack of sorting information that is spread on various social media makes it easy for fraudsters to carry out their actions, and finally, there is so much fake news spread [1]. Hoax information has a negative influence on the real world, such as creating hatred towards something both individuals and groups. False information with the objective of creating confusion and dispute is intolerable in principle and should be combatted at any prevalence greater than zero [2]. Classification becomes a problem that can be solved by machine learning, one of them in the case of text classification [3]. In this study, researchers intend to classify deceptive news using machine learning, for text-based news in Indonesian.

Previous research into the literature is from Pratiwi [4] using Naïve Bayes Classifier for hoax classification, shows the value of accuracy is 78.6%. The study from Prasetyo [5] Modified K-Nearest Neighbor for hoax classification shows the value of accuracy 75%, recall 0.75, precision 0.83, f-measure 0.79. The study from Handayani [6] to compare XGBoost, SVM, and MLP for Breast Tumor Classification. The conclusion form this study is XGBoost and SVM become the algorithm chosen by researcher, because it gives good results. The study conducted Santhanam [7] about experimenting XGBoost algorithm for prediction and classification of different datasets shows that XGBoost give good result with high level of accuracy based on 4 datasets from two categories (classification and regression). In study by Syahrani [8] about comparing XGBoost and Random Forest algorithm for DNA sequence classification give conclusion that XGBoost provides better accuracy, both by default and in parameter tuning.
This study uses Extreme Gradient Boosting (XGBoost). XGBoost is a variant of the Tree Gradient Boosting algorithm. Boosting method focuses on the new learning process on data with a weak learner value than the previous process [8]. XGBoost is used in supervised learning cases such as regression, classification, and ranking [9]. Tianqi Chen and Carlos Guestrin [10], who are the makers of this algorithm write that XGBoost can be used well in handling real cases both small and large scale by using a minimal amount of resources. They also wrote that the computations carried out on this algorithm are distributed and parallel to make the learning process of the system faster, and also able to adjust the flexibility of a case. This algorithm is improved in two aspects: accelerate the tree construction and proposing a new distributed algorithm for tree searching. Optimizing the value of the objective function is the core of XGBoost [11]. In this research entitled “Fake News Classification for Indonesian News Using Extreme Gradient Boosting (XGBoost)” hope can be to classify hoax nor valid news from Indonesian news.

2. Research methodology
The research consists of several stages, from dataset preparation, text preprocessing, modeling with XGBoost and model evaluation.

2.1. Problem analysis
The Study utilizes the concept of supervised machine learning. The dataset is the main ingredient in the training of machine learning as knowledge of the program. News collected into a dataset is made in 2 labels, namely hoax data and valid data. The collected news is taken from several Indonesian news websites, such as kompas.com, detik.com, cnnindonesia.com, liputan6.com, and turnbackhoax.id. The period of news taken is 2015-2020. The dataset that will be used in this study contains 500 data. The dataset will be divided into two parts, namely training data and test data. The scale of the dataset is 80% training data and 20% test data, or 400 data into training data and 100 data in the dataset will be test data used to measure model performance.

![Figure 1. Display one of the news portals.](image)

Based on the dataset, valid news generally reports an event or event that has been proven to occur and usually has a clear source (has reliable references). In contrast to non-hoax news, hoax news that is obtained from a lot of broadcasts whose contents are inciting and tends to ask to be disseminated as much as possible, but also not infrequently the news without references or sources.
Table 1. Examples of hoax and valid news used.

| Hoax News Example                                                                 | Valid News Example                                                                 |
|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| "WARNING!!! Drugs model baru LSD. Atau di kenal dgn nama “kertas Happy” model dan bentuknya menarik, rasa manis seperti permen. efek halusinasi tinggi. Di jual Rp.10rb. Kalau sudah ketagihan harga Rp.36jt/gram. Sebarkan untuk menyelamakan generasi penerus kita dr drugs. Target anak2 sekolah." | Jakarta, CNN Indonesia -- Striker Barcelona, Lionel Messi berhasil memenangkan Ballon d’Or 2019. Messi berhasil terpilih sebagai pemain terbaik tahun ini dan menyingkirkan sejumlah bintang lapangan hijau lain macam Virgil van Dijk, Sadio Mane, dan Cristiano Ronaldo. Mohamed Salah ada di posisi kelima. Pemain asal Mesir itu sukses membuktikan kehebatannya tahun lalu bukan sekadar sensasi satu musim. |

2.2. Text preprocessing
Preprocessing text (preprocessing text) is the beginning of the text classification flow. Preprocessing text is useful in processing the data used in a form suitable for the classification process [12]. Text Preprocessing is divided into several stages, which in this study the stemming stage is not used. Stemming is the process of breaking down existing words into basic forms of words, such as the example word in Bahasa, the word "mempengaruhi" is changed to "pengaruh". This process is not used because it is based on the research of Rasywir [13] wrote that the process of text classification will produce better results without stemming the preprocessing process. The preprocessing stages used in this study are case folding, tokenization, and filtering.

Case Folding is the process of converting entire text from data used into a standard. The standard used will make all words lowercase (lowercase). Case folding is used to help simplify the data used into one form (all small letters), because if one word has two different writing patterns (for example in Bahasa "Buku" and "buku") then the computer will consider it different, even though it has the same meaning. This process is used to make the computer does not discriminate between words that are actually the same despite different writing patterns.

At the stage of tokenization, data (in the form of text) is processed into parts called tokens. This process has the purpose of eliminating certain characters such as punctuation and symbols because they are considered to have no use in processing text, and dividing each word that is useful in the processing of text data. The references to the separator between tokens are spaces and punctuation marks.

Desain rumah satu lantai yang unik !

desain rumah satu lantai yang unik !

Figure 2. Tokenization illustration.

Filtering takes words that are considered important in the classification taken from tokenization results. In this study, the filtering used is to remove stopwords and punctuation (punctuation). Stopwords are words that appear in the data used but do not have any contribution to the meaning of the document [14].
2.3. Modeling with XGBoost

This study uses XGBoost in classifying whether the news is classified as a hoax or not. The dataset after going through the process of preprocessing and sharing data will create a classification model. Before arriving at the classification process, the dataset will enter the feature selection stage. Feature selection is the process of selecting a subset of words (terms) that appear in the data used and are considered to represent important information from a data, which later is useful as a feature that is only used in text classification. Feature selection aims to make the training data used simpler, and improve the accuracy of the classification system due to the strengthening of features that are considered to be strong in the document [15]. This study uses the Term Frequency-inverse document frequency (tf-idf) method in feature selection. This method combines two concepts for weight calculation, namely Term Frequency (tf) which measures how often the term appears in the data, and inverse document frequency idf is used to measure how important the term is, which is like the example of the word in bahasa "adalah" often appears but does not have a significant meaning in the text. The formula of tf-idf is tf times idf.

\[ W = tf \times idf \]  

The dataset is then divided into training data and test data using the k-fold cross validation method. K-fold cross validation is one of the statistical methods used in seeing the success of a system. The method used to repeat where each input attribute is randomized, and the entire data acts as training data and test data [16]. In the k-fold cross, training and testing are carried out k times, according to the number of models to be made. Each data has the same training time and once becomes test data. According to Gareth [17] the common ideal k value taken is 5 - 10, with the use of a small dataset, is only 5 models or k = 5, this is to minimize the biased that occurs in the model so that it can do a better classification.

This study is to apply XGBoost to classify hoax news texts. XGBoost made major improvements to the traditional Gradient Boosting Decision Tree (GBDT) algorithm in computing speed, generalization performance, and scalability. XGBoost is a variant of the Tree Gradient Boosting algorithm. Boosting is a form of algorithm that converts or enhances learning of weak systems into strong ones [18].
2.4. Model evaluation

Evaluation of the model is seen using a confusion matrix based on test data. The results of the confusion matrix will see several values such as precision, recall, f-1 score to accuracy.

| Classified as valid | Valid data | Hoax data |
|---------------------|------------|-----------|
| True positive (tp)  |            | False positive (fp) |
| False negative (fn)|            | true negative (tn)  |

Precision and Recall is a calculation matrix used to measure the effectiveness of information retrieval [15]. Precision measures the accuracy of the system in classifying text whether it is hoax or valid. The precision formula is used as follows.

$$Precision = \frac{tp}{(tp+fp)}$$  \hspace{1cm} (2)

Recall is the level of success of introducing a class that must be recognized. Recall measures the accuracy of the system producing relevant values so that the text can be classified.

$$Recall = \frac{tp}{(tp+fn)}$$  \hspace{1cm} (3)

F1-score is a comparison of the average precision or precision with recall. f1-score is used to see the balance between precision and recall.
\[
F1-score = \frac{1}{2} \left( \frac{1}{\text{precision}} + \frac{1}{\text{recall}} \right)
\]  

(4)

Accuracy calculations are used to see how accurately the system classifies. The formula used to get the calculation of the accuracy value of the research conducted, here is the equation for getting the accuracy value.

\[
\text{Accuracy} = \frac{\sum \text{match (tp+tn)}}{\sum \text{tp}} \times 100
\]  

(5)

3. Implementation and results

Data for dataset take narratives from the news that turnbackhoax.id declared hoaxes. While valid news sources are taken from news media, provided the news taken is not an opinion article or news about the opinions of people/figures. The dataset is then employed to the preprocessing stage. The results of preprocessing show each of the words that often appear each label both valid and hoax, after being deleted the words are classified as stopwords. The word “person” appears the most on both labels based on the dataset used, of which 186 were found on valid labelled data and 80 times were found in hoax labelled data.

![Figure 5. Most frequent word by labels.](image)

The results of the test evaluation showed an accuracy rate of 89%. Visualization of the evaluation model can be seen in the confusion matrix below which shows how much data from 100 test data (20% of 500 data in the dataset) are properly classified. Confusion matrix shows positive true value (tp) is 44, true negative (tn) is 2, false negative (fn) is 6, and false positive (fp) is 2.
Figure 6. Confusion Matrix model test results against test data.

Based on the confusion matrix above, the values of precision, recall, and f1-score are taken as follows:

- \[ \text{precision} = \frac{44 \text{(TP value)}}{44 \text{(TP Value)} + 2 \text{(FP Value)}} = 0.96 \]
- \[ \text{recall} = \frac{44 \text{(TP Value)}}{44 \text{(TP Value)} + 6 \text{(FN Value)}} = 0.88 \]
- \[ \text{f} - 1 \text{score} = \frac{1}{2} \left( \frac{1}{\text{precision}} + \frac{1}{\text{recall}} \right) = 0.92 \]

The accuracy given in the model built is:

- \[ \text{Accuracy} = \frac{44 \text{(TP Value)} + 48 \text{(TN Value)}}{100 \text{(jumlah data uji)}} = 0.92 \]

4. Conclusion

In this paper, the study tries to classify text for Indonesian fake news with XGBoost. Based on the results of the study conducted, the conclusion drawn that hoax text classification for Indonesian news using XGBoost can be implemented in text classification with high value of accuracy. The High accuracy value in this paper is obtained by tuning parameters. The model is built using XGBoost give the high evaluation value, the value accuracy is 92 % with a comparison of datasets there are 80 : 20 out of a total of 100 randomly shared data.

References

[1] Septanto H 2018 Pengaruh HOAX dan Ujaran Kebencian Sebuah Cyber Crime Dengan Teknologi Sederhana di Kehidupan Sosial Masyarakat J. Sains dan Teknol. Kalbiscientia 5(2) p 157–162
[2] Allen J Howland B Mobius M Rothschild D and Watts D J 2020 Evaluating the fake news problem at the scale of the information ecosystem Sci. Adv. 6(14) p 1–7
[3] Hania A A 2017 Mengenal Artificial Intelligence, Machine Learning, Neural Network, dan Deep Learning J. Teknol. Indones. June
[4] Pratiwi I Y R Andrie R and Rahutomo F 2017 Study of Hoax News Detection Using Naive Bayes Classifier in Indonesian Language Int. Conf. Inf. Commun. Technology Syst. p 73–78
[5] Prasetyo A R Indriati and Adikara P P 2018 Klasifikasi Hoax Pada Berita Kesehatan Berbahasa
Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya 2(12) p 7466–7473

[6] Handayani A Jamal A and Septiandri A A 2017 Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara J. Nas. Tek. Elektro dan Teknol. Inf. 6(4) p 394–403

[7] Santhanam R Uzir N Raman S and Banerjee S 2017 Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets Int. J. Control Theory Appl. 9 March p 651–662

[8] Syahrani I M 2019 Analisis perbandingan teknik ensemble secara boosting (xgboost) dan bagging (random forest) pada klasifikasi kategori sambatan dna iswaya maalik syahrani

[9] Beeravalli V 2018 Comparison of Machine Learning Classification Models for Credit Card Default Data, medium.com. [Online]. Available at: https://medium.com/@vijaya.beeravalli/comparison-of-machine-learning-classification-models-for-credit-card-default-data-c3cf805c9a5a. [Accessed: 02-Sep-2019]

[10] Chen T and Guestrin C 2016 XGBoost: A Scalable Tree Boosting System p. 785–794

[11] Ma B Meng F Yan G Yan H Chai B and Song F, 2020 Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data Comput. Biol. Med. p. 103761.

[12] Nikhath A K Subrahmanyam K and Vasavi R 2016 Building a K-Nearest Neighbor Classifier for Text Categorization Int. J. Comput. Sci. Inf. Technol. 7(1) p 254–256

[13] Rasywir E and Purwarianti A, 2015 Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin J. Cybermatika 3(2) p 1–8

[14] Chasbullah S, 2018, Analisis Berita Hoax Menggunakan Klasifikasi Multinomial Naive Bayers Dengan Fitur N-Gram, INSTITUT PERTANIAN BOGOR

[15] Manning C D, 2009 An Introduction to Information Retrieval Cambridge UP

[16] Darmawidjuna I G M Wirawan I M A and Kusumayanti L M D, 2016 Pengembangan Sistem Pendukung Keputusan Kredit Konsumtif Dengan Metode Naive Bayes (Studi Kasus: PT. Bank Pembangunan Daerah (BPD) Bali Cabang Tabanan) J. Nas. Pendidik. Tek. Inform. 5(3) p 108–114

[17] Gareth James D W and Trevor Hastie R T, 2013 An Introduction to Statistical Learning 7th

[18] Adi G Y N N Tandio M H Ong V and Suhartono D, 2018 Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia Procedia Comput. Sci. 135 p 473–480