Hoax Detection of Indonesian News Media on Twitter Using IndoBERT with Word Embedding Word2Vec

Pernanda Arya Bhagaskara S M, Sri Suryani Prasetyowati, Yulianti Sibaroni
School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email: 1bhagaskara@student.telkomuniversity.ac.id, s.suryani@telkomuniversity.ac.id, yulianti@telkomuniversity.ac.id
Correspondence Author Email: bhagaskara@student.telkomuniversity.ac.id

Abstract— Hoaxes is data that is added or deducted from the news that occurred. In the digital age, hoaxes are increasingly being spread, and people are very quickly affected by their spread, especially hoaxes circulating in Indonesian news media on social media. Disseminating information that has not been confirmed as accurate can cause public concern and anxiety. Virtual diversion has transformed into a correspondence key to begin thinking, talking, and moving around cordial issues. In this manner, exploration will be led by consolidating the IndoBERT model with the Word2Vec development highlight in arranging deception news in Indonesian news media. This model was constructed using K-Fold cross-validation to enhance model performance across extensive data sets. The information utilized comes from tweets shared on Twitter by the Indonesian public. The trials that have been carried out demonstrate that combining Word2Vec with IndoBERT is effective at detecting hoaxes, with an overall accuracy score of 88% for the entire dataset. This conclusion can be drawn from the classification results of Word2Vec with IndoBERT. Also, the best precision and inference for every cycle is almost 99%. In addition, the study's objective is to identify hoax news in Indonesian news media disseminated via social media. This will encourage individuals to be more cautious when reading and disseminating news, as untrue information will significantly impact certain individuals.

Keywords: Indonesian News Media; Hoax; IndoBERT; Word2Vec; Social Media.

1. INTRODUCTION

Lately, there has been a ton of deception news circling in the Indonesian news media. Hoax is information or news that contains facts that are not true or uncertain [1]. The hoax that is going around the Indonesian media is very bad and makes a lot of people think the wrong things. It says that the Indonesian media has become a place where many Indonesians can get the latest news. The hoax was due to the rapid spread of fake news in the Indonesian media, which provided an opportunity to reach out to customers in the informal community who had not seen the news recently [2].

Indonesian social media such as Twitter, Instagram, and Facebook are increasingly popular among the public [3]. Especially Twitter. Nowadays, most Indonesians use Twitter to obtain various data, particularly the most recent news. False information or news, whose integrity is only sometimes apparent, can be easily identified due to easy access. Every year, the development of Twitter clients accelerates. Comparatively, the number of users is anticipated to rise by 26%, reaching an average of 192 million in the fourth quarter of 2021 [4]. Accordingly, the utilization of tweets on Twitter is expanding, which can have both positive and unfortunate results. Due to the ease with which people interact on social media, particularly Twitter, it is difficult for most people to distinguish hoaxes from non-hoaxes due to the sheer volume of hoaxes that circulate. One of the side effects of social media is this.

In the study of detecting fraud on Twitter, SVM and other techniques were utilized. SVM is a calculation based on features that do a great job of grouping text with a lot of information about the text as an element. Although the word vector is unique in that the related words in a sentence can differ, most of the output sequence messages can be extracted directly, and the message report consists of many redundant parts. Hyperplanes that can delimit locales into subsets are created via SVM estimation. A hyperplane is a shape that separates two classes and considers the distance between their closest components. [5]. The best ratio, as determined by benchmark system accuracy tests, is 90:10, with a value of 78.33 percent. Highlights influencing the phony news class incorporate the Twitter retweet and URL elements and backing highlights. Twitter fake news can be identified with the help of the SVM classification. Create knowledge, explicitly setting off and battling [6]. Additionally, LSTM, or Long Short-Term Memory, is an iterative network design that avoids issues with long-term dependency by using memory cells and gate banks. LSTM is altered to defeat the deficiencies of RNNs in that they can't foresee words, recall long-put-away data, or eliminate information that is not generally required [7], despite the fact that CNN is a deep-learning model that computers frequently use. Nonetheless, as Kim's exploration has shown, the CNN model can likewise be utilized to arrange sentences [8]. Convolutions, or close associations, connect multiple neurons to a single neuron in the resulting layer in CNNs. Consequently, CNNs can separate spatial data from pictures. Its compositionality and immutability of location distinguish CNN [9]. LSTM-CNN in identifying COVID-19-related fraud. To find the best model bounds, we did some research. With 16 unitary layers, LSTM-CNN can achieve 79.71 percent accuracy in experiments that combine regularizers and dropouts. [4].

Connecting the impacts of running LSTM and the IndoBERT strategy utilizing Twitter-isolated datasets on the Covid for area coercion. LSTM accomplishes a typical precision of 87.54 percent considering experimental outcomes. In addition, the tests indicate that the IndoBERT model is accurate on average to 92.07 percent [7].
Therefore, the IndoBERT model outperforms the LSTM model in deception detection tasks and has been demonstrated to offer superior average accuracy results.

Word2Vec and IndoBERT perform exceptionally well in various fraud detection tests that place an emphasis on word expansion and deep learning calculations. IndoBERT can deliver superior results than other strategies. Consequently, Twitter data on public opinion regarding news will be used in research that combines Word2Vec and IndoBERT.

IndoBERT is a model that follows the BERT Base setup [10]. The number of datasets used in IndoBERT is limited because it uses more datasets than there are different strategies. In IndoBERT, fake news is detected by utilizing transfer learning from pretrained transformer models like customized native pretraining BERT, multilingual pretraining mBERT, and monolingual pretraining IndoBERT [11]. According to the findings of the research, mBERT base-cased finetuned had an accuracy of 97.93 percent; As a result, the model outperforms the competition [12]. Moreover, this last undertaking proposition, other than utilizing IndoBERT, likewise utilizes Word2Vec. Word2Vec doesn't cover vector spaces, embedding, analogies, similarity metrics, etc. first, last, or best. Yet, word2vec is basic and available [13]. Training a log-bilinear model based on a jump-gram or continuous bag-of-words (cbow) architecture, such as implemented in word2vec and fastText, is the standard method for learning word representation. [14][15]. The word2vec representation has been widely used in NLP pipelines to improve performance. Their impressive ability to transfer to new problems suggests they collect vital statistics about three training sets [16]. Then, hoax detection research using Word2Vec with IndoBERT is still rare. Therefore, this research will focus on detecting hoaxes using Word2Vec with IndoBERT by using Indonesian news data on Twitter from public opinion on news circulating in Indonesia. Besides that, this research can reduce and educate the public about hoax news circulating.

2. RESEARCH METHODOLOGY

2.1 System Design

This diagram Word2Vec's word embedding using the IndoBERT method.

![Diagram of Word2Vec with IndoBERT](image)

Figure 1. Diagram of Word2Vec with IndoBERT

The process depicted in Figure 1 begins with data preparation which includes data collection, eliminating data duplication, and labeling. Case folding will be the first step in preprocessing, followed by data cleaning, removing stopwords, and latest is word stemming. From then on, including development will be done using Word2Vec. Information sharing will be done to data splitting the dataset into data set and data train. Then the characterization uses IndoBERT and the framework is run. The system will then predict which ones are hoaxes and which are not, then an evaluation will be carried out.

2.2 Dataset Preparation

Three primary processes are used to set up data collection: data collection, data deduplication, and data labeling. Data is collected through the Twitter API for Indonesian language political news from 2020 to 2023 for this system. Once the data has been successfully collected, the author will label it, and duplicate values will be removed from the data set by the system. Data is classified based on whether hoax or non-hoax. Previous research determines the categories used in this study. The labeled dataset is as seen in Table 1.
Table 1. Dataset with Labels

<table>
<thead>
<tr>
<th>Number</th>
<th>Text</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Menurut survei indikator, kepuasan kinerja @Jokowi turun! Apa alasannya? Hal ini menunjukkan bahwa 44 responden merasa keadaan perekonomian semakin memburuk.</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>@KPK mengingatkan untuk tidak ikut negosiasi politik dalam kasus Azis Syamsuddin karena dana yang disita KPK terkait dengan tindak pidana pencucian uang (TPPU). Jaksa menuduh Rita kemudian memberi Robin dan Maskur Husain Rp 5.197 crore untuk menangani kasus tersebut.</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>@Megawati minta @SriMulyani jangan pelit dana wakaf untuk lunasi utang ke China!</td>
<td>1</td>
</tr>
</tbody>
</table>

2.3 Preprocessing

The preprocessing text means getting rid of noise like punctuation marks, stops, and terms that don't mean much in the context of the text. [17]. There are steps that need to be done before the data can be prepared. The data's quality will be improved when it is used to train the fraud data model as a result of this preprocessing step. Here are a few phases of preprocessing.

a. Case Folding
   This technique converts sentences into lowercase which the String library assists.

b. Data Cleaning
   Data cleaning step includes data cleanup, which includes unescaping HTML to get rid of HTML tags in the sentence. URL stripping to get rid of all links in the sentence, mention reduce to get rid of words with the "@" prefix and remove punctuation to get rid of all punctuation in the sentence.

c. Stopword Removal
   This technique removes commonly used words that have no special meaning, such as pronouns, prepositions, and conjunctions. Pronouncing stopwords in Indonesian uses the Python Natural Language Toolkit (NLTK) library.

d. Word Stemming
   This technique changes a word to its root form by removing the initial and final affixes. The stemming process uses a unique library for processing Indonesian, the Python Sastrawi library.

Table 2. Labelled Dataset

<table>
<thead>
<tr>
<th>Number</th>
<th>Text</th>
<th>Preprocessed text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Menurut survei indikator, kepuasan kinerja @Jokowi turun! Apa alasannya? Hal ini menunjukkan bahwa 44 responden merasa keadaan perekonomian semakin memburuk.</td>
<td>nurut survei indikator puas kerja jokowi turun apa alasan hal ini tunjuk bahwa responden rasa ada ekonomi makin buruk</td>
</tr>
<tr>
<td>2</td>
<td>@KPK mengingatkan untuk tidak ikut negosiasi politik dalam kasus Azis Syamsuddin karena dana yang disita KPK terkait dengan tindak pidana pencucian uang (TPPU). Jaksa menuduh Rita kemudian memberi Robin dan Maskur Husain Rp 5.197 miliar untuk menangani kasus tersebut.</td>
<td>kpk ngingat untuk tidak ikut negoisasi politik dalam kasus aziz syamsyuddin karena dana disita kpk kait tindak pidana cuci uang tppu jaksa nuduh rita kemudian beri robin maskur husain miliar nangani kasus tersebut</td>
</tr>
<tr>
<td>3</td>
<td>@Megawati minta @SriMulyani jangan pelit dana wakaf untuk lunasi utang ke China!</td>
<td>megawati minta sri mulyani jangan pelit dana wakaf lunas utang china</td>
</tr>
</tbody>
</table>

2.4 Modelling

K-Fold Cross Validation is used to measure model performance across datasets, which allows k training and testing. K-Fold cross-validation also validates the built model evaluation results. The models used in this system are IndoBERT and Word2Vec. This technique trains k models after dividing the data into k groups. According to previous research, this system employs five "k"s. Evaluation metrics like precision, accuracy, recovery, and the F1 score assess each model's performance after every k iteration [18]. The predicted results from the ranking report for each iteration are added together in calculations to determine the model's overall performance. Figure 2 shows a delegate's of K-Fold cross-validation.

Figure 2. K-Fold Cross Validation Illustration
a. IndoBERT

IndoBERT is a model of Bidirectional Encoder from Transformers Representations (BERT) which is prepared utilizing Indonesian language records [19]. IndoBERT has a similar architectural design to BERT but is conditioned on Indo4B in two 8 phases with a pool of 4 billion words and ~250 million sentences. The first stage uses a maximum sequence length of 128 for pertaining, and the second stage uses a full sequence length of 512. The difference in size in the second phase is intended so IndoBERT can learn more from the first phase.

![Figure 3. Illustration of BERT Sizes](image)

Figure 3 demonstrates that the base and large BERT models have the same number of encoder layers—twelve for the base model and 24 for the large model—which are referred to as transform blocks in this article. They additionally have more extensive feedforward networks (768 and 1024 secret blocks, separately) and more consideration heads (12 and 16, separately) than the default setup in the reference Transformer execution in the first article (6 encoder layers, 512 blocks). Furthermore, eight heads of consideration.

![Figure 4. Model Architecture of IndoBERT](image)

The model's architecture has 13 layers, as shown in Figure 4. 12 IndoBERT stowed away layers and 1 classifier on top. The use of words from the previous level is the focus of each level. Size for each level: [number of tokens * 768 * bundle size]. In the wake of doing tests in view of the learning misfortune capability,
I observed that the ideal number of ages is 8, the clump size is 8 examples, and the learning rate is 2e-5. The greatest number of tokens in each succession is restricted to 200 [20].

b. **Word2Vec**

Word2Vec is a neural network (NN) model that encodes semantic information for each term in a corpus using untagged training data. To comprehend semantic similarity, evaluate the cosine similarity between the word arrays. The vectors of words with similar meanings are the same, but the vectors of different words are different. They are utilized in document parsing, NER, sentiment classification, and other controlled language processing tasks. It comes in two flavors: Continuous Bag of Words (CBOW) and Skipgram (SG). [21]. Word2vec's neural network architecture is a projection layer-hidden neural network trained with stochastic gradient descent and a backpropagation algorithm. Words in the context of n-grams are displayed in continuous vectors by the projection layer. A correlation between words occurs when words in the context of N-grams appear simultaneously or repeatedly and are activated by the same weight. The hidden layer and the input layer are connected by the weights. A W x N matrix, where V is the dimension of the input layer, and N is the dimension of the hidden layer, is used to represent the weights between the two layers. An N x V matrix represents the matrix W in this instance, which is located between the output layer and the hidden layer. [13].

![CBOW and Skip-gram Models](image)

**Figure 5.** Word2vec Illustration

Word2vec is restricted by a window in the CBOW model and uses the words to the left and right of the target word to bind the word. Skipgram also uses the word to predict words to the left and right of the windowed word at the same time. Each term used as input is encoded into a one-hot vector. The difference between the two models is the word prediction model. In CBOW, an intermediate layer will average the input word vectors because CBOW accepts a number of n words as input.

### 2.5 Evaluation Performance Measurement

Performance measurement is a step to analyze whether the system is sound or not built and has good predictive results. This final design research measures performance using accuracy, f1-score, precision, recall, and error rate.

#### a. Confusion Matrix

The classification model's accuracy prediction is evaluated using the test accuracy confusion matrix. As shown in Table 3, the accuracy testing matrix has four conditions.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Positive Prediction</th>
<th>Negative Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certainly Positive</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Certainly Negative</td>
<td>FP</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Explanation:**
- TP = The quantity of positive and unsurprising information is right.
- TN = The anticipated number of negative information is right.
- FP = The dataset is negative, but should be positive.
- FN = The quantity of positive information yet anticipated negative information.

#### b. Accuracy

The precision of its ability to correctly classify the model is its accuracy. (1).

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\%
\] (1)
c. F1 – Score
   F1 - Score is a comparison of the average weighted precision and gain scores (2).
   \[ F1 – Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  
   \[ (2) \]

d. Precision
   Precision is the comparison value between the model's predicted results and the requested data (3).
   \[ \text{Precision} = \frac{TP}{TP + FP} \times 100\% \]  
   \[ (3) \]
e. Recall
   Recall is a value that shows how successful it is to retrieve information. (4).
   \[ \text{Recall} = \frac{TP}{(TP + FN)} \times 100\% \]  
   \[ (4) \]
f. Error Rate
   The error rate is the proportion of anticipated blunders utilizing the accompanying condition (5).
   \[ \text{Error Rate} = \frac{(FP + FN)}{(TP + FP + FN + TN)} \times 100\% \]  
   \[ (5) \]

3. RESULT AND DISCUSSION

3.1 Indonesian News Media on Twitter

From the data obtained, there are 3939 non-hoax labels and 3689 hoax labels, with the label "0" indicating non-hoax and "1" indicating hoax, which can be seen in Figure 6. Therefore this data will be used for modelling, namely IndoBERT with a combination of Word2Vec with a percentage of 51.6% : 48.4% for the non-hoax and hoax labels.

![Figure 6. Label Distribution](image)

3.2 Word2Vec Modelling

Word2Vec transforms words into vectors for IndoBERT in this modeling. One of the Word2Vec architectures, skip-gram, will be used in this study. A straightforward neural network with a hidden layer that has been trained...
to predict the probability of a word based on the input word is the skip-gram model. Instinctively, the skip grams model is something contrary to the CBOW model. This architecture tries to accurately predict the words before and after the current word using the current word as its input. The goal of this model is to learn and predict context words that are related to an input word. With a wider range of word vectors, it was discovered that the quality of predictions improves, but computational complexity also rises, based on experiments assessing this model's accuracy. Word2Vec shows an example of the word "gagal" below. In this example, the word "gagal" will be turned into a vector, which will continue the IndoBERT model, as shown in Table 4. This will be finished for all words in the informational index utilized in this research.

![Image](image.png)

Table 4. Word2Vec Example

<table>
<thead>
<tr>
<th>Text</th>
<th>Word2Vec Vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>gagal</td>
<td>-0.003040998000000215</td>
</tr>
</tbody>
</table>

3.3 IndoBERT Modelling

Table 5. IndoBERT Parameter After Word2Vec

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch size</td>
<td>12</td>
</tr>
<tr>
<td>epoch</td>
<td>3</td>
</tr>
<tr>
<td>df eval : df test</td>
<td>0,8</td>
</tr>
<tr>
<td>random state</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5 shows the boundaries used to construct the IndoBERT model in this review, so the batch size is twelve, where batch-size is how much information is utilized in preparation. The amount of memory used and the default BERT epoch usage of 3 increase with batch size. for df eval, which is a split dataset-like function. The proportion utilized is awesome for the informational index. Random state is set to zero with K-Fold cross-validation test data at a ratio of 20%:80%, and the data is not random. Besides, Table 6 presents the attributes of the model.

Table 6. IndoBERT Performance in K-Fold cross-validation

<table>
<thead>
<tr>
<th>K</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>87</td>
</tr>
<tr>
<td>4</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>87</td>
</tr>
</tbody>
</table>

This model always performs well when cross-validating the K-Fold, as shown in Table 6. Over 80% of the time, this model is accurate. However, not all predictions for each class are always accurate. In the principal cycle, the model obviously has the most noteworthy exactness. Consequently, the general exhibition of the model can be determined by adding all anticipated outcomes from every cycle of the positioning report. The exactness of the IndoBERT model with Word2Vec for the whole dataset is assessed at 88%.

![Image](image.png)

Figure 7. IndoBERT Confusion Matrix k = 5

As referenced over, this model gives the best exactness from the primary cycle with a worth of close to 100% and a general precision worth of 88%. The figure depicts the entire connection matrix for this model. 7. 2737 non-fraudulent actions and 2648 fraudulent ones are correctly predicted by the model. Along these lines, this conjecture has 718 missing pieces of information.
3.4 Overall Performance

<table>
<thead>
<tr>
<th>Measures Performance</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>88%</td>
</tr>
<tr>
<td>Precision</td>
<td>89%</td>
</tr>
<tr>
<td>Recall</td>
<td>86%</td>
</tr>
<tr>
<td>Error Rate</td>
<td>11%</td>
</tr>
<tr>
<td>F1-Score</td>
<td>88%</td>
</tr>
</tbody>
</table>

Based on the results of the experiments, average values of accuracy, precision, recall, error rate and f1-value are obtained for all data sets. Accuracy is how precisely it can correctly classify the model, accuracy is the value of the comparison between the requested data and the model’s prediction results, recall is the value that describes the success of the model in obtaining information, error rate is the comparison between the data that was wrong predicted and actual data, and for f1-score is the average comparison of the weighted precision value and the acquisition value, as shown in Table 7, the precision is 88%, the recall is 86%, the error rate is 11%, and the f1 score is 88%.

4. CONCLUSION

Based on the results and discussion of experiments that were carried out with 7628 Twitter data labeled fraudulent and non-fraudulent news compared to Indonesian news, this study can draw the following conclusion: 48.4%. This dataset is also used to construct a Word2Vec model with IndoBERT using K-Fold cross-validation. The model with the best exactness was produced in the principal cycle with an IndoBERT precision of almost 100% for each given k emphasis. A precision value of 88% was acquired for the informational index. The Word2Vec model with IndoBERT shows great execution for each emphasis. The total score values are accurate at k=1 99%, k=2 80%, k=3 87 percent, k=4 88 percent, and k=5 87 percent for each iteration and all datasets. They accomplished 89% execution estimation precision, 86% review, 11% blunder rate, and 88% f1 result. When combined with the Word2Vec word insertion model and IndoBERT, this value yields a relatively high accuracy value for fraud detection and zero fraud in Indonesian news. Extra information in the dataset with additional names could be utilized for additional exploration to give more proof to this end, particularly with model consideration. After that, find a value that works for each model by utilizing the other parameters and adding additional words to fine-tune the model and attain high accuracy.

REFERENCES


