Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic

Leonnardo B. Hutama[1*], Derwin Suhartono[2]
E-mail: [1]leonnardo.hutama@binus.ac.id, [2]dsuhartono@binus.edu
* Corresponding author
[1]Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Keywords: Hoax News Classification, Topic Distribution, XLM-R, mBERT, BERTopic, Indonesian Language

Received: August 10, 2022

Technology and information growth enable internet users to play a role in disseminating information, including hoax news. One way that to avoid hoax news is to look for sources of information, but valid news is not always perceived as 'true' by individuals because human judgments can lead to bias. Several studies on automatic hoax news classification have been carried out using various deep learning approaches such as the pre-trained multilingual transformer model. This study focuses on classifying Indonesian hoax news using the pre-trained transformer multilingual model (XLM-R and mBERT) combined with a BERTopic model as a topic distribution model. The result shows that the proposed method outperforms the baseline model in classifying fake news in the low-resource language (Indonesian) with accuracy, precision, recall, and F1 results of 0.9051, 0.9515, 0.8233, and 0.8828 respectively.

Povzetek: Raziskava se ukvarja z identifikacijo lažnih novic v Indoneziji s pomočjo modelov XLM-R and mBERT.

1 Introduction

Technology and information growth have make it easier for users to convey and consume information in recent years. Today, all internet users can play a role in disseminating information. However, the widely spread information is not all true and reliable, including hoax news. According to the expert [1], fake news is information that is intentionally intended to mislead people with a specific purpose. Meanwhile, according to [2] hoax news is misleading that imitates original content but has a different purpose. There are still many people who are victims of hoax news, including Indonesians. The Mastel Survey [3] about the Indonesia hoax outbreak, conducted on 941 respondents stated that as many as 34.60% of respondents received hoax news every day, and as many as 70.7% of hoaxes were received in text form and the largest hoax news distribution channel was through social media and websites. The survey also stated that 63.30% of respondents believed that hoax news was not a hoax because they got the news from trusted people and 24.5% of respondents believed hoax news because of convincing sentences.

Hoax news also has a negative impact, such as causing anxiety, triggering public panic, and can lead to manipulation and fraud that can bring down humans [4]. One way that can be done to avoid hoax news is to look for credibility or sources of information. The credibility of information is very important to avoid the risk of consuming hoax news [5]. However, valid news is not always perceived as 'true' by individuals because human judgments on the credibility of information can be influenced by the opinion of the individual and lead to bias. This bias can be reduced by implementing automated fake news detectors [6].

Several studies on automatic hoax detection have been carried out using various deep learning approaches. For example, by using Logistic Regression and Support Vector Machine (SVM) in research [7] and using a combination of CNN and LSTM in research (e.g. [6], [8], [9]) who managed to get an accuracy value of around 44-98%. These approaches generally use traditional word embedding methods such as word2vec which have limitations in overcoming the polysemy that occurs when the same word has different meanings. To overcome this, the researcher [10] created a model, named transformer, which uses a self-attention mechanism so that it can calculate a better representation of a word in a sentence. The pre-trained transformer models were provided by [11] in the library namely Huggingface. Transformer models such as BERT, ALBERT, and XLNet have been widely used in making automatic hoax detectors and have been shown to have better performance than traditional approaches (e.g. [12], [13], [14]). The transformer also has pre-trained multilingual models, such as mBERT, XLM, and XLM-R, which can be used for many languages. Compared to other multilingual transformer models, the XLM-R model proved to have the best performance because it was trained with a 2.5 TB dataset size which is larger than other multilingual models [15]. However, the model performs worse when implemented in a low-resource language, such as Indonesian, than in a high-

resource language. This happens because it is difficult to find data in low-resource languages.

In addition to using a pre-trained transformer model, the performance of automatic hoax news detectors can also be improved by using a topic modeling approach such as Latent Dirichlet Allocation (LDA). This is done by adding the results of the topic distribution from the LDA as input for the classification model [16]. The LDA, which uses a probability approach, is one of the most popular models used for topic modeling. However, other models with a transformer approach, such as BERTopic, have been shown to have better performance than LDA [17]. Besides having better performance, BERTopic also supports multilingual text. However, few still use topic modeling to maximize the performance of automatic hoax news detectors with low-resource languages.

Based on previous research, research using a multilingual model is proven to be able to overcome the problem of low-resource language. In addition, topic modeling can also improve performance on news classification. However, previous studies used this method only for single-language models, and its performance, when applied to low-resource languages, is still unknown. Addressing the aforementioned issues, this paper proposes a predictive model architecture using a multilingual transformer pre-trained model and a topic modeling model to detect hoax news automatically in a low-resource language (Indonesian). The model consists of a combination of a pre-trained multilingual transformer model to obtain contextual representation results and BERTopic in conducting topic distribution. In summary, the contribution of this work can be written as follows:

- We proposed a deep learning architecture with a pre-trained transformer multilingual model with additional topic word representation from topic modeling as a feature extraction method for the hoax news classification system in the Indonesian language.
- We evaluate its performance by comparing it with the previous research algorithm which gave the best performance in the classification of hoax news in the Indonesian language.
- We evaluate the performance of the topic distribution method by comparing the model using and without using the topic distribution method.
- We show that our method produces better performance than previous studies on the classification of hoax news in the Indonesian language.

2 Related work

Research on automatic fake news detection using the Transformer model has been carried out by several researchers. For example, research conducted by [12] and [13] used the BERT transformer model to detect hoaxes on English news datasets, namely LIAR and the FNC-1 datasets sequentially. The research shows that the pre-trained transformer model gets 15-20% better accuracy than the traditional CNN and LSTM models in classifying hoax news. Researchers [18], [19] also researched the classification of hoax news using the ConstraintAI'21 English news dataset with an ensemble model consisting of three transformer models (BERT, ALBERT, and XLNet). The ensemble method managed to get 98% accuracy in classifying hoax news. Although the accuracy value is good, this method cannot necessarily be applied to low-resource language news datasets, such as Indonesian, because the model used is only trained with an English corpus. This is proven by research conducted by [19] to detect hoax news in English and several low-resource languages (Hindi, Swahili, Vietnamese, and Indonesian). This study proves that the performance of the English news dataset produces an accuracy value of 84%, while for news with low-resource languages the accuracy value is 5% lower at 79%. This shows that the transformer model applied to low-resource language news produces less optimal performance due to the limited corpus available in low-resource languages.

The performance of the hoax news classification system is determined by the size and language of the dataset used when the model is trained. However, for low-resource languages, such datasets are still lacking. This can be overcome by using multilingual transformer models, such as mBERT because these models are trained in various languages. For example, research was conducted by [20] in detecting fake news on Arabic tweets using mBERT. The study compared the mBERT model with the Arabic-based BERT model (AraBERT) with and without fine-tuning. In this study, fine-tuned mBERT managed to get an F1 score above 0.92, which was better than the other baseline models (AraBERT, Distilbert) on all tasks. In addition to mBERT, research by [15] has succeeded in proposing a transformer-based multilingual model named XLM-RoBERTa (XLM-R) which has been pre-trained in 100 languages and with a 2.5TB dataset size. This model outperformed other multilingual models (mBERT and XLM) in performing classification, sequence labeling, and question answering. This research also applied the XLM-R model to a low-resource language and managed to improve performance by 15.7% in Swahili and 11.4% in Urdu.

Research using the XLM-R model for low-resource languages has also been carried out by several other researchers, such as using the XLM-R model for order classification of low-resource languages (Polish) which was conducted by [21]. This research has succeeded in proving that the XLM-R model has higher precision, recall, and F1 values compared to other multilingual models (mBERT and HerBERT) in performing sequence labeling. Similar studies were also conducted by [22] using the XLM-R model to conduct sentiment analysis on low-resource language (Korean) movie reviews. The study compared the XLM-R model with the mBERT and several pre-trained Korean models (KoBERT, KorBERT, and KR-BERT). In this study, the XLM-R model, combined with prune and Bi-GRU, managed to get a value of precision, accuracy, and recall that was 3.63% better than other models.

The XLM-R model is also used by several researchers such as [23] and [24] in building a classification system for hoax news. The study was conducted by [23] using the

XLM-R model to classify hoax news on English and Chinese tweets. In this study, the XLM-R model was fine-tuned and managed to get the best average accuracy compared to other traditional algorithms (Naïve Bayes, SVM, C4.5, Random Forest, CNN, BiLSTM, C-LSTM) with a value of 99% with only use raw text without being translated. In addition, despite having a large model size, the training time required by the XLM-R model is also comparable to other algorithms. In the research [24], the XLM-R model is used to classify hoax news in Spanish. This study compared XLM-R combined with CNN as a feature extractor with other pre-trained transformer models, namely BETO and XLM-R without CNN. The best results were obtained by the XLM-R model combined with CNN with an accuracy score of 0.96 followed by the XLM-R model without CNN with an accuracy score of 0.95 and BETO with an accuracy score of 0.93. These two studies prove that the XLM-R model can be used in a hoax news classification system in languages other than English, namely Chinese and Spanish. However, there has been no research using the XLM-R model for Indonesian, so its performance, when applied to Indonesian, is still unknown due to differences in the corpus.

The language model plays an important role in determining the performance of the hoax news classification system. However, the performance of the hoax news classification system can also be improved by adding a feature extraction method such as Latent Dirichlet Allocation (LDA). For example, the use of LDA as a feature extractor is carried out by [25] in classifying fake reviews using Logistic Regression and Multi-Layer Perceptron approaches. The study managed to get an accuracy score of 81% better than without using LDA. Research using LDA was also carried out by [16] who combined the pre-trained transformer model XLNet and LDA as a feature extraction method in classifying hoax news from English social media articles. The research was conducted by combining the results of the topic distribution from LDA and the results of contextual representation from XLNet. The LDA method has succeeded in increasing the performance of the XLNet model with an accuracy value of 96% better than the accuracy value without using the LDA of 94%. Besides LDA, there are several other topic distribution models such as BERTopic, a topic modeling technique based on BERT and TF-IDF. Research conducted by [17] in conducting topic modeling of Arabic news with BERTopic succeeded in proving that BERTopic has better performance than the Latent Dirichlet Allocation (LDA) topic modeling technique which uses a probability approach, and Non-Negative Matrix Factorization (NMF) which uses the matrix factorization method. These results are based on the Normalized Pointwise Mutual Information (NPMI) value where BERTopic gets a positive NPMI value while LDA and NMF get a negative NPMI value. Despite the success in improving performance, only a few researchers use feature extraction techniques with topic distribution to improve the performance of hoax news classification systems with low-resource languages.

The summary of the related works can be seen in table 1.

Table 1: Related Works Summary

No	Topic	Result
1	Hoax News Detection on English News Dataset using BERT [12][13]	BERT model gets 15-20% better accuracy than CNN and LSTM models.
2	Hoax News Detection on English News Dataset using the Ensemble model (BERT, ALBERT, XL-Net) [18]	The ensemble model gets 98% accuracy in classifying hoax news.
3	Hoax News Detection on Low-Resource Language Dataset (Hindi, Swahili, Vietnamese, Indonesian) using Transformer Multilingual model [19]	The multilingual transformer accuracy is 5% lower in low-resource languages (79%) than in English language (84%).
4	Detecting Hoax News on Arabic Tweet using mBERT and AraBERT [20]	Fine-tuned mBERT (multilingual model) gets 0.92 F1-score and it's better than AraBERT (baseline model).
5	Sequence Labeling on Low-Resource Language (Polish) Order using XLM-R [21]	XLM-R model gets higher precision, recall, and F1 compared to mBERT and HerBERT.
6	Sentiment Analysis on Low-Resource Language (Korean) Movie Review using XLM-R [22]	XLM-R model with prune and Bi-GRU managed to get 3.63% better precision, accuracy, and recall than other models.
7	Hoax News Classification on English and Chinese Tweets using XLM-R [23]	XLM-R gets the best accuracy value (99%) compared to Naïve Bayes, SVM, C4.5, Random Forest, CNN, BiLSTM, C-LSTM.
8	Hoax News Detection on Spanish Language using XLM-R combined with CNN [24]	XLM-R combined with CNN get accuracy score of 96% and it's better than BETO model with an accuracy of 93%.
9	Hoax News Detection using Logistic Regression (LR) combined with Latent Dirichlet Allocation (LDA) [25]	Logistic Regression combined with LDA gets better performance than without using LDA.

10	Hoax News Detection on English Social Media Articles using XLNet and LDA [26]	XLNet combined with LDA gets better accuracy (94%) than without using LDA (94%).
11	Topic Modeling on Arabic News using BERTopic, Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF) [17]	BERTopic gets a positive Normalized Pointwise Mutual Information (NPMI) value. LDA and NMF gets a negative NPMI value.

Based on previous research, the multilingual model has proven to be applicable to low-resource languages and the use of topic distribution models such as LDA can improve the performance of the classification model. However, there are still few studies that use the multilingual model for Indonesian news and the use of topic distribution models can be further improved by using a better model, such as BERTopic, compared to traditional topic distribution models, such as LDA. Therefore, this study focuses on the classification of Indonesian hoax news using the multilingual transformer model (mBERT and XLM-R) combined with the BERTopic topic distribution model to improve the performance of the classification model in the low-resource language used.

Transformer models are generally trained using English datasets so that they have good performance when used in several NLP tasks with English data. However, the Transformer model can also be used in NLP tasks other than English dataset by using the monolingual transformer model, which has been trained in a specific language, or can use the multilingual transformer model, which has been trained in several languages. This study will focus on two multilingual transformers models in classifying Indonesian hoax news, namely XLM-R and mBERT.

3.1.1 XLM-R

XLM-R (XLM-RoBERTa) is a multi-language model developed from the transformer architecture which has an architecture similar to the XLM model. The difference between the XLM-R model and the XLM model lies in the purpose of the training where XLM focuses on the Translation Language Model (TLM) while XLM-R has a training focus similar to the RoBERTa model, which focuses on the Masked Language Model (MLM). The XLM-R model uses Byte Pair Encoding (BPE), where most common XLM-R model uses Byte Pair Encoding (BPE), where the most common consecutive pair of data bytes is replaced with bytes that do not appear in the data, thereby improving vocabulary relationships between different languages. In addition to BPE, the XLM-R model is also trained to use the same words in different languages, thus enabling the XLM-R model to understand the context of one language from another. This causes the

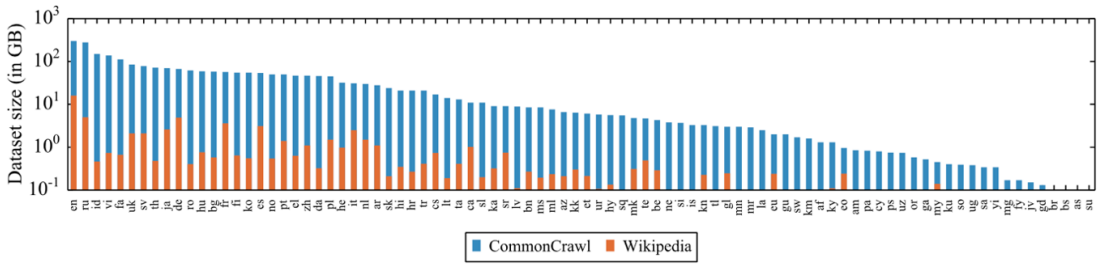

Figure 1: XLM-R Train Dataset Size (source: [15])

3 Fundamental theories

3.1 Multilingual transformer model

The transformer is a model architecture that focuses on self-attention mechanisms, without using convoluted convolutions like RNN [10]. Transformers allow for more significant parallelization and the use of self-attention can result in more interpretable models. The transformer consists of 2 main parts, namely an encoder, and a decoder. The encoder is a collection of several layers, each of which has 2 sub-layers, namely a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network, where there is a residual connection concept to maintain information based on the position of each. part to serve as input for the entire network. The decoder has a layer similar to the encoder with the addition of one section to process multi-head attention to the output of the encoder.

XLM-R model to be used to increase the use of low-resource languages by utilizing other languages that have high resources. The most prominent development of the XLM-R model lies in the size of the dataset used where XLM-R is trained using 2.5TB of Common Crawl data in 100 languages, making the XLM-R model superior to previous multi-language models such as BERT and XLM-100 which have Weaknesses in low resource languages. The comparison of the size of the dataset used by XLM-R (blue bar) with multi-language model BERT and XLM-100 (orange bar) against 88 languages can be seen in figure 1.

3.1.2 mBERT

Multilingual-BERT (mBERT) is a version of BERT, which has a transformer-like architecture [10] and is pre-trained with unlabeled raw text. In contrast to BERT that only trained in a single language, mBERT was trained in

104 languages, including Indonesian, with the largest Wikipedia dataset using a masked language modeling (MLM) objective [26]. In dealing with unbalanced data, oversampling is carried out on small languages and undersampling for large languages. This allows the mBERT model to be used for low-resource languages dataset such as Indonesian.

3.2 BERTopic

BERTopic is a topic modeling technique based on BERT and TF-IDF in creating clusters that produce easy-to-interpret topics and important words that describe the topic. The BERTopic model uses BERT in word processing which produces extraction results that match the context of the word. Besides BERT, BERTopic also supports several other word extraction models such as XLM-R which supports more than 50 languages for text extraction other than English. In addition, several advantages of BERTopic that can support this research are that BERTopic supports a variety of topic modeling, hierarchical topic reduction, and can find the number of topics automatically [27].

For document grouping, BERTopic uses two algorithms, namely the UMAP algorithm to reduce the dimensions of the word insertion results and the HDBSCAN algorithm for document grouping [17]. The grouping of documents in the BERTopic model is based on the value of the class-based variant of TF-IDF (c-TF-IDF) in determining the uniqueness of a document compared to other documents.

3.3 Hyperparameter tuning using Optuna

The transformer models used in this study, XLM-R and mBERT, require hyperparameters to obtain optimal training results. The hyperparameters used in the model include learning rate, weight decay, and training epochs. The value of these hyperparameters can be determined manually, but it is very inefficient and time-consuming. Therefore, in this study, hyperparameter tuning was carried out in determining the hyperparameter values automatically using a framework called the Optuna. Unlike other optimization frameworks, Optuna provides a Define-by-run API that dynamically constructs hyperparameter search [28].

4 Research methodology

This research will be divided into three stages, namely the initiation stage, implementation stage, and evaluation stage. These stages can be seen in figure 2. The dataset that has been collected will go through a preprocessing process before being input into the topic distribution model. The cleaned data will be processed using a topic distribution model to get a contextual words representation of the news which will then be combined with news articles as input. The combined news and topic distribution data will then be input to the transformer model for hyperparameter tuning to produce hyperparameters that will be used for model evaluation. The evaluation results will also be compared with other models to determine the performance of the proposed method.

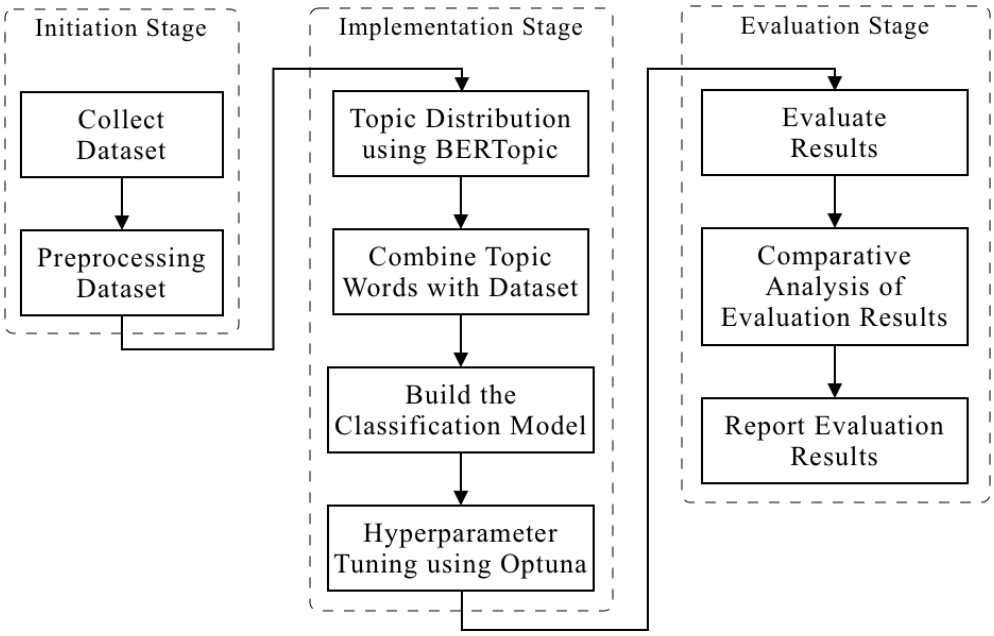

Figure 2: Research Methodology Stages

4.1 Proposed method

The proposed deep learning method will focus on 2 main parts, which are the feature extraction stage using the pre-trained multilingual model (XLM-R, mBERT) and the topic distribution stage using the BERTopic Model.

In the first stage, news data that has passed the preprocessing stage will be extracted for its features before being used as classification input. The Feature Extraction stage at this stage uses the pre-trained multilingual model (XLM-R, mBERT) from the

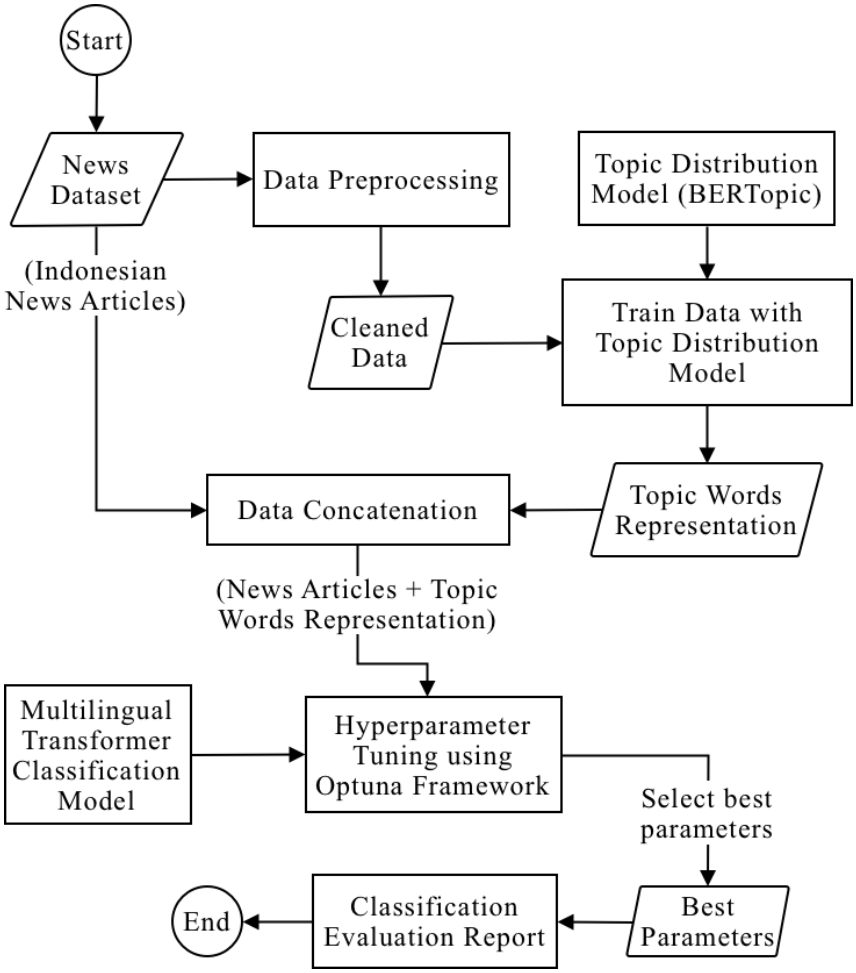

Figure 3: Proposed Deep Learning Method Flowchart

Hunggingface library because the model supports text input in multi-languages, making it suitable for use in Indonesian news datasets. The second stage aims to produce topic word representations that will be used as input, along with the results of feature extraction, in the classification model. BERTopic also supports multilingual language text as input.

The proposed method will be implemented using the python programming language and the *Huggingface* transformers library which can be done using the PyTorch library. If illustrated, the proposed deep learning architecture can be seen in figure 3. Figure 3 briefly describes the flow of the proposed method architecture. The dataset will first go through a preprocessing process which consists of removing URLs, converting words to lowercase, removing stop words, removing excess spaces, and stemming. The results of the preprocessing will go through the topic modeling process using the BERTopic model to get the topic words representation, and the feature extraction process using the pre-trained transformer multilingual model (XLM-R, mBERT). The results of topic distribution and feature extraction will then be combined and used as input to the classification model to produce a SoftMax score which will be used as predictive output.

To get maximum performance results, a parameter tuning process will be carried out using the Optuna framework to find the optimal parameters. In addition, the size of the number of topics used in the topic modeling stage will also be divided into 5 and 10 words. After getting a good performance, a performance evaluation is carried out.

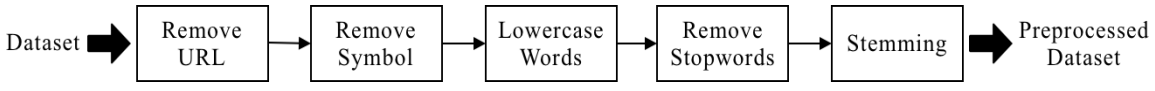

Figure 4: Preprocessing Progress Flow

4.2 Data

The dataset that will be used in the study is a dataset that contains news articles in Indonesian that have been labeled as valid or hoaxes originating from two sources. The first dataset comes from research (Rahutomo et al., 2019) with a total of 600 news data consisting of 228 hoaxes and 372 valid news data downloaded from https://data.mendeley.com/datasets/p3hfgr5j3m. The dataset has been labeled by three experts with a voting system as the result which is used as the dataset label. The second dataset comes from a GitHub repository called Pierobeat, which was downloaded from https://github.com/pierobeat/Hoax-News-Classification, with a total of 500 data consisting of 250 valid news data and 250 labeled hoax news data collected from official news sites (Turnbackhoax, liputan6, Kompas, Detik and CNN Indonesia). These datasets have two attributes, which are the news attribute and the label attribute. In detail, the number of data used can be seen in table 2.

Table 2: News Dataset Distribution

Dataset	**Label**		**Total**
	Hoax	**Valid**	
Rahutomo	228	372	600
Pierobeat	250	250	500
Combined (Rahutomo + Pierobeat)	**478**	**622**	**1100**

4.3 Preprocessing

The preprocessing stage aims to make the feature extraction process more contextual. The steps taken during preprocessing can be seen in figure 4.

Preprocessing begins by removing URLs, which are news references and symbols that have no meaning in a sentence. The punctuation symbol is not removed because it can provide context information to the Transformer model. The next stage is normalization by changing the word to lowercase. After that, redundant spaces and stop words are removed which did not contain meaningful information. The last stage is stemming, which is removing the affix of a word so that it becomes a basic word that contains the essence of a word. The stop words and stemming removal phase will use the NLTK Library which provides methods for stemming and stop words dictionary.

4.4 Experiment

To compare the proposed methodology, comparisons will be made with several architectures with different types of topic distribution methods using several models in classifying Indonesian fake news. This research will be divided into several scenarios where each scenario will use different algorithms and topic distribution methods. This study will use two types of multilingual pre-trained transformer models to be compared (mBERT, XLM-R). Each model will be trained without using the word distribution of topics. Then proceed by using 5 and 10 topics words representation from the BERTopic model. Hyperparameter tuning using the Optuna framework will be used to get the optimal performance. A total of 25 experiments will be carried out using the Optuna framework where each experiment will use different parameter values. At the end of the iteration will display the experimental results and the most optimal hyperparameter values. The range values specified in the Optuna framework can be seen in table 3.

Table 3: Optuna Hyperparameter Range

Hyperparameter	**Range Value**
Learning Rate	4e-5 - 0.01
Weight Decay	4e-5 - 0.01
Training Epochs	2-5

5 Result and discussion

In our research, the BERTopic model is used to distribute topics whose results are used as input along with news articles. The BERTopic model automatically determines the number of topics from the news dataset used, which is 24 topics, and each topic has a different word representation. The results of BERTopic can be seen in table 4 below.

Table 4: BERTopic Topic Words Representation Result

Topic	**Top 5 Words**	**Top 10 Words**
0	permen, dot, narkoba, surabaya, mengandung	permen, dot, narkoba, surabaya, mengandung, sekolah, anak, diduga, razia, makanan
1	tahanan, brimob, petugas, rutan, kerusuhan	tahanan, brimob, petugas, rutan, kerusuhan, mako, rikwanto, blok, sel, keributan
2	pokemon, bahasa, yahudi, go, game	pokemon, bahasa, yahudi, go, game, monster, arti, anak, pikachu, permainan
3	traveloka, ananda, anies, kanisius, derianto	traveloka, ananda, anies, kanisius, derianto, out, walk, aksi, acara, nilai
4	facebook, data, rudiantara, indonesia, pengguna	facebook, data, rudiantara, indonesia, pengguna, kominfo, pemerintah, memblokir, konten, akun
5	lele, ikan, kanker, mengandung, kotor	lele, ikan, kanker, mengandung, kotor, tubuh, sel, limbah, kolam, manusia
6	stroke, darah, jarum, penderita, pertolongan	stroke, darah, jarum, penderita, pertolongan, pasien, jari, pembuluh, sakit, otak

7	reog, davao, kjri, ponorogo, pembakaran	reog, davao, kjri, ponorogo, pembakaran, city, filipina, reyog, kesenian, budaya
8	masjid, istiqlal, aksi, 212, abu	masjid, istiqlal, aksi, 212, abu, 21, subuh, peserta, masuk, bppmi
9	iphone, plus, apple, melengkung, pengguna	iphone, plus, apple, melengkung, pengguna, mudah, bengkok, smartphone, saku, layar
10	bulu, sikat, babi, gigi, bristle	bulu, sikat, babi, gigi, bristle, kuas, bahan, rambut, produk, terbuat
11	ahok, tni, dik, karoseri, arahan	ahok, tni, dik, karoseri, arahan, vs, lokal, konflik, monas, polri
12	palestina, israel, muslim, arab, gap	palestina, israel, muslim, arab, gap, natal, canada, haji, snack, jco
13	presiden, gaji, wakil, kenaikan, rp	presiden, gaji, wakil, kenaikan, rp, dokumen, beredar, mulyani, sri, diusulkan
14	maluku, guru, jokowi, tpg, partai	maluku, guru, jokowi, tpg, partai, tim, 2019, pengawas, esports, pemilu
15	isis, irak, bom, serangan, as	isis, irak, bom, serangan, as, suriah, orang, pasukan, luka, kelompok
16	luhut, china, indonesia, kau, freeport	luhut, china, indonesia, kau, freeport, kapal, minggu, jokowi, ahad, negaraku
17	gaji, presiden, rp, penghasilan, juta	gaji, presiden, rp, penghasilan, juta, pejabat, rpp, pns, indeks, negara
18	korban, mirna, meninggal, pelaku, pembunuhan	korban, mirna, meninggal, pelaku, pembunuhan, jessica, polisi, ditemukan, sakit, bom
19	pesawat, penumpang, penerbangan, lion, bandara	pesawat, penumpang, penerbangan, lion, bandara, maskapai, pilot, air, lambertus, kopilot
20	gempa, banjir, longsor, bencana, gunung	gempa, banjir, longsor, bencana, gunung, agung, erupsi, desa, kabupaten, gorontalo
21	iklan, hago, guru, unj, ika	iklan, hago, guru, unj, ika, siswa, paslon, sosok, profesi, peserta
22	gula, yg, formalin, kristen, masuk	gula, yg, formalin, kristen, masuk, mohon, nanas, jawa, pake, nabi
23	presiden, yg, kereta, orang, ini	presiden, yg, kereta, orang, ini, 2019, korban, wakil, papua, mahasiswa

After that, hyperparameter tuning was carried out using the Optuna framework for the pre-trained multilingual model (XLM-R and mBERT), with and without the topic word representation, to get maximum performance. The best hyperparameter results can be seen in table 5.

Table 5: Optuna Hyperparameter Tuning Best Hyperparameter

Model	**Learning Rate**	**Weight Decay**	**Epochs**
mBERT	3.6642e-05	0.00021	4
XLM-R	2.067e-05	3.116e-05	4
mBERT + 5 topic words	3.6524e-05	0.009	5
XLM-R + 5 topic words	2.5452e-05	0.00041	5
mBERT + 10 topic words	1.996e-05	0.0041	4
XLM-R + 10 topic words	1.7388e-05	1.0221e-05	4

By using these hyperparameters, an evaluation of the training results is carried out using the accuracy, precision, recall, and f1 metrics. The evaluation was carried out by comparing the proposed model with the BERT model [12][13] and the ensemble model, consisting of BERT, ALBERT, and XLNet [18]. Each model will be trained using the news dataset with a ratio of 7:3 between the training and testing data. The training results of each model can be seen in table 6.

Table 6: Training Result

Model	**Accuracy**	**Precision**	**Recall**	**F1**
BERT [12][13]	0.7454	0.72	0.6067	0.6585
Ensemble Method [18]	0.7545	0.4606	0.8723	0.6029
mBERT	0.7863	0.8181	0.6067	0.6967
XLM-R	0.8	0.64	0.826	0.72
mBERT + 5 topic words	0.8242	0.8392	0.7014	0.7642
XLM-R + 5 topic words	0.8121	0.7812	0.7462	0.7633

mBERT + 10 topic words	**0.9051**	**0.9515**	0.8233	**0.8828**
XLM-R + 10 topic words	0.8935	0.8818	**0.8712**	0.8765

Table 6 shows the results of the evaluation of this study. From these results, the proposed method, which uses a multilingual model (XLM-R, mBERT) combined with a topic distribution model using BERTopic outperforms the BERT model [12][13] and the ensemble model [18]. This happens because those models (BERT, ALBERT, XLNet) are only trained in English and are not suitable for low-resource languages such as Indonesian. The proposed method also proves that the BERTopic model successfully improves the performance of the individual pre-trained multilingual model. The mBERT model with 10 topic words got the best accuracy, precision, recall, and F1 values of 0.9051, 0.9515, 0.8233, and 0.8828 respectively. While the XLM-R model with 10 topic words managed to get accuracy, precision, recall, and F1 values which were not much different from mBERT, which are 0.8935, 0.8818, 0.8712, and 0.8765 respectively. These results are much better than using only individual pre-trained models which only produce accuracy, precision, recall, and F1 values of 0.7863, 0.8181, 0.6067, and 0.6967 for the mBERT model and 0.8, 0.64, 0.826, and 0.71 for the XLM-R model.

This study also proves that the more words from the topic distribution are used, the higher the performance of the multilingual pre-trained model. It can be seen in the table above, that the results of multilingual model training using 10 topic words succeeded in increasing the performance of accuracy, precision, recall, and F1 by 0.08, 0.11, 0.12, and 0.11 respectively compared to using only 5 topic words.

6 Conclusion and future work

This study shows a comparison of pre-trained multilingual models in building a classification system for hoax news against low-resource language news (Indonesian Language). The deep learning model proposed in this study uses a pre-trained multilingual model, namely mBERT and XLM-R, which are added topic-representative words from the distribution of topics using the BERTopic model. The proposed model is proven to be successful in improving the performance of the individual pre-trained multilingual model with the results of accuracy, precision, recall, and F1 of 0.9051, 0.9515, 0.8233, and 0.8828 respectively for the mBERT model with 10 topic words and 0.8935, 0.8818, 0.8712, and 0.8765 for the XLM-R model with 10 topic words. These results are proven to increase accuracy significantly when compared to using only individual pre-trained multilingual models in classifying hoax news on news with low-resource language (Indonesian Language). In addition, this study also proves that the more topic-representative words used, the higher the performance of the model. This can be seen from the performance of using 10 topic representative words which resulted in better accuracy, precision, recall, and F1 values of 0.08, 0.11, 0.12, and 0.11 respectively compared to using only 5 topic representative words.

In the future, development can be done by exploring datasets related to fake news using other low-resource languages in large numbers, because this study has not used large amounts of data. In addition, development can be done by trying different learning methods and models, such as using the ensemble method combined with different topic distribution models.

References

[1] Werme, "Pemberdayaan Masyarakat Mengenai Isu Hoax Vaksinasi Terhadap Kesehatan Masyarakat di Masa Pandemi Covid-19," 2016. https://www.kompasiana.com/nanda82966/622f94e3bb448645a54718b2/pemberdayaan-masyarakat-mengenai-isu-hoax-vaksinasi-terhadap-kesehatan-masyarakat-dimasa-pandemic-covid-19 (accessed Mar. 21, 2021).

[2] D. M. J. Lazer *et al.*, "The science of fake news," *Science (1979)*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[3] Mastel, "Hasil Survey Wabah Hoax Nasional 2019," 2019. https://mastel.id/hasil-survey-wabah-hoax-nasional-2019/ (accessed Jun. 21, 2021).

[4] Kemenkeu, "Jangan Mudah Termakan Hoax, Saring Sebelum Sharing," Jun. 22, 2020. https://www.djkn.kemenkeu.go.id/artikel/baca/13206/Jangan-Mudah-Termakan-Hoax-Saring-Sebelum-Sharing.html (accessed Jun. 21, 2021). https://doi.org/10.15548/amj-kpi.v2i1.486.

[5] M. Viviani and G. Pasi, "Credibility in social media: opinions, news, and health information—a survey," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 7, no. 5, p. e1209, 2017. https://doi.org/10.1002/widm.1209.

[6] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models," *Procedia Computer Science*, vol. 179, pp. 704–712, 2021. https://doi.org/10.1016/j.procs.2021.01.059.

[7] A. Stöckl, "Detecting Satire in the News with Machine Learning," *arXiv preprint arXiv:1810.00593*, 2018.

[8] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNN-LSTM)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020. https://doi.org/10.1109/access.2020.3019735.

[9] A. Roy, K. Basak, A. Ekbal, and P. Bhattacharyya, "A deep ensemble framework for fake news detection and classification," *arXiv preprint arXiv:1811.04670*, 2018.

[10] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[11] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.

[12] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019. https://doi.org/10.3390/app9194062.

[13] M. Qazi, M. U. S. Khan, and M. Ali, "Detection of fake news using transformer model," in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2020, pp. 1–6. https://doi.org/10.1109/icomet48670.2020.9074071.

[14] D. J. M. Pasaribu, K. Kusrini, and S. Sudarmawan, "Peningkatan Akurasi Klasifikasi Sentimen Ulasan Makanan Amazon dengan Bidirectional LSTM dan Bert Embedding," *Inspiration: Jurnal Teknologi Informasi dan Komunikasi*, vol. 10, no. 1, pp. 9–20, 2020. https://doi.org/10.35585/inspir.v10i1.2568.

[15] A. Conneau *et al.*, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[16] A. Gautam, V. Venktesh, and S. Masud, "Fake news detection system using xlnet model with topic distributions: Constraint@ aaai2021 shared task," in *International Workshop onCombating Online Hostile Posts in Regional Languages during Emergency Situation*, 2021, pp. 189–200. https://doi.org/10.1007/978-3-030-73696-5_18.

[17] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique," *Procedia Computer Science*, vol. 189, pp. 191–194, 2021. https://doi.org/10.1016/j.procs.2021.05.096.

[18] S. Gundapu and R. Mamidi, "Transformer based Automatic COVID-19 Fake News Detection System," *arXiv preprint arXiv:2101.00180*, 2021.

[19] A. De, D. Bandyopadhyay, B. Gain, and A. Ekbal, "A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–20, 2021. https://doi.org/10.1145/3472619.

[20] M. S. H. Ameur and H. Aliane, "AraCOVID19-MFH: Arabic COVID-19 Multi-label Fake News & Hate Speech Detection Dataset," *Procedia Computer Science*, vol. 189, pp. 232–241, 2021. https://doi.org/10.1016/j.procs.2021.05.086.

[21] J. Radom and J. Kocoń, "Multi-task Sequence Classification for Disjoint Tasks in Low-resource Languages," *Procedia Computer Science*, vol. 192, pp. 1132–1140, 2021. https://doi.org/10.1016/j.procs.2021.08.116.

[22] N. R. Shin, T. Kim, D. Y. Yun, S.-J. Moon, and C. Hwang, "Sentiment analysis of Korean movie reviews using XLM-R," *International Journal of Advanced Culture Technology*, vol. 9, no. 2, pp. 86–90, 2021.

[23] A. Zervopoulos, A. G. Alvanou, K. Bezas, A. Papamichail, M. Maragoudakis, and K. Kermanidis, "Deep learning for fake news detection on Twitter regarding the 2019 Hong Kong protests," *Neural Computing and Applications*, vol. 34, no. 2, pp. 969–982, 2022. https://doi.org/10.1007/s00521-021-06230-0.

[24] Z. Guan, "TSIA team at FakeDeS 2021: Fake news detection in spanish using multi-model ensemble learning," 2021.

[25] S. Jia, X. Zhang, X. Wang, and Y. Liu, "Fake reviews detection based on LDA," in *2018 4th International Conference on Information Management (ICIM)*, 2018, pp. 280–283. https://doi.org/10.1109/infoman.2018.8392850.

[26] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" *arXiv preprint arXiv:1906.01502*, 2019. https://doi.org/10.18653/v1/p19-1493.

[27] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," *Frontiers in Sociology*, vol. 7, 2022. https://doi.org/10.3389/fsoc.2022.886498.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.