Personality Identification from Social Media Using Ensemble BERT and RoBERTa

Eggi Farkhan Tsani[1], Derwin Suhartono*[2]
[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
*Corresponding author
E-mail: eggi.tsani@binus.ac.id, dsuhartono@binus.edu

Keywords: social media, big five personality, data augmentation, BERT, RoBERTa

Recieved: March 30, 2023

Social media growth was fast because many people used it to express their feelings, share information, and interact with others. With the growth of social media, many researchers are interested in using social media data to conduct research about personality identification. The identification result can be used as a parameter to screen candidate attitudes in the company's recruitment process. Some approaches were used for research about personality; one of the most popular is the Big Five Personality Model. In this research, an ensemble model between BERT and RoBERTa was introduced for personality prediction from the Twitter and Youtube datasets. The data augmentation method also introduces to handling the imbalance class for each dataset. Pre-trained model BERT and RoBERTa was used as the feature extraction method and modeling process. To predict each trait in the Big Five Personality, the voting ensemble from BERT and RoBERTa achieved an average f1 score 0,730 for Twitter dataset and 0,741 for Youtube dataset. Using the proposed model, we conclude that data augmentation can increase average performance compared to the model without data augmentation process.

Povzetek: Članek uvaja model združevanja siestemov BERT in RoBERTa za napovedovanje osebnosti iz podatkov Twitterja (X) in Youtube, z izboljšanjem s pomočjo podatkovne augmentacije.

1 Introduction

Based on Leadership IQ's study [1] of 20.000 companies, 46% of new employees resign from their jobs within one and a half years, and 89% of their failures are because of attitudinal reasons. The recruitment process and high turnover because of resignations can be incur high costs for a company. Curriculum vitae screening and face-to-face interviews were not enough to make sure the candidate had a good attitude. One of the approaches for getting a candidate's attitude was to do personality identification. This identification can also use to determine which position of the job was particularly fit for the candidate [2].

Social media has grown so fast around the world. Currently, many people can use social media not only for communication but also to express their thoughts, expectations, and feelings [3]. In January 2021, datareportal survey noted that social media users in Indonesia reached 170 million or 61,8% of the total population [4]. It means more than half of the population in Indonesia uses social media in their daily activities.

Because users use social media to express their feelings, the researchers can use social media data to conduct research about personality prediction. Different approaches were introduced to predicting personalities such as Big Five Personality, MBTI (Myers-Briggs Type Indicator), and DISC (Dominance Influence Steadiness Conscientiousness) [5]. From these three approaches mentioned above, Big Five Personality is the most accepted model to describe personality structure and divide it into personal and group [6]. Table 1 below describes the advantages of the Big Five Personality approach compared to the other two. The Big Five Personality consists of five personality traits that are usually called OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) [7].

Table 1: Personality approach comparison

	Big Five	**MBTI**	**DISC**
Results	Unique	16 personalities	12 profiles
Predictive	Yes	No	No
Valid	Yes	Yes	No
Reliable	Yes	Yes	No

This research uses two social media datasets to build a prediction model. The first dataset is Twitter data, which consists of 508 users with around 46.000 posts collected manually, and the second dataset is Youtube data, consists of 10.000 clips extracted from 3.000 different videos of people speaking in English to the camera. This dataset is called the First Impression dataset and was downloaded from ChaLearn [8]. Both datasets are based on text and have multi-label cases. The prediction model was built using an ensemble BERT (Bidirectional Encoder Representation from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) classifier for the personality prediction case.

Compared to other personality prediction research, this research uses a different approach to increase classification performance. Data augmentation using the back translation method was introduced to increase the number of datasets and handle the imbalance class. As a result, classification performance increase around 3-5% compared to classification using the original dataset.

2 Related works

Research about personality prediction has been done previously using various social media data. Facebook, Twitter, and Youtube were three popular social media that were used for research about personality prediction. The dataset available for research either in public or private was collected and labeled to Big Five traits manually.

Research conducted by [9] uses a Facebook dataset called myPersonality consists of 250 users with around 10.000 statuses labeled with the Big Five traits label. Their research used LIWC (Linguistic Inquiry and Word Count), and SPLICE (Structured Programming for Linguistic Cue Extraction) features then used MLP (Multi-Layer Perceptron) classifier to produce 70,78% average accuracy. Another research that uses myPersonality dataset was conducted by [10] using LIWC, SPLICE, and SNA (Social Network Analysis) as feature extraction methods and XGBoost algorithm to achieve the best result with 74,2% average accuracy.

Another personality prediction research that uses social media datasets was also conducted by [11]. The dataset used was Twitter in Bahasa which consists of Twitter posts from 250 users labeled with Big Five traits label. Their research uses SGD (Stochastic Gradient Descent), XGBoost, and super learner to produce a good ROC-AUC (Receiver Operating Characteristic and Area Under Curve) score. Research conducted by [12] also uses Twitter dataset with more data that consists of tweets from 508 users. This research use word-n-gram and Twitter metadata to process using Random Forest classifier to produce 0,744 f1 scores on average. Research using Twitter dataset was also conducted by [13] using similar data to previous research. This research uses pre-trained models BERT, RoBERTa, and XLNet combined with TF-IGM statistical features. These methods used an averaging model to make better predictions result.

Popular social media dataset also used for research conducted by [14]. Their research uses Youtube vlog dataset which consists of 404 vlogs with audio-video features and transcripts. Decision tree and SVM algorithm were used and produced better results over baseline average performance. Another research using Youtube dataset was also conducted by [15]. This research uses Youtube translations only to create a model using Word2Vec, GloVe, and BERT as feature extraction methods then uses SVM and SVR for classification. This approach produces 0,612 f1 scores as the best prediction result.

Table 2: Previous personality research

Author	Dataset	Classifier	Findings
Tandera et al., 2017	Facebook	MLP	Accuracy 70,78%
Tadesse et al., 2018	Facebook	XGBoost	Accuracy 74,2%
Adi et al., 2018	Twitter	Super Learner	ROC AUC 0,992
Jeremy et al., 2019	Twitter	Random Forest	F1 score 0,744
Christian et al., 2021	Twitter	BERT, RoBERTa, XLNet	F1 score 0,757
Farnadi et al., 2016	Youtube	Decision Tree and SVM	RMSE 0,115
Lopez-Pabon et al., 2022	Youtube	SVM and SVR	F1 score 0,612

Previous research that uses the Big Five Personality approach processes the dataset with imbalance class. For optimizing performance, we modify the data with data augmentation, whereas the original dataset will add modified data so imbalance class can be minimized. With minimizing imbalance class in the data, dataset quality will increase, and modeling process will have a better performance.

The advantage of the ensemble model also reflects from previous research above. For Facebook, Twitter, and Youtube datasets, the best performance resulted from the ensemble model. The best model from Facebook

dataset resulted from ensemble boosting using XGBoost, Twitter dataset resulted from ensemble averaging using BERT, RoBERTa, and XLNet classifiers, and Youtube dataset resulted from SVM and SVR. With this consideration above, this research uses data augmentation to produce better dataset quality and ensemble model to increase classification performance.

3 Methodology

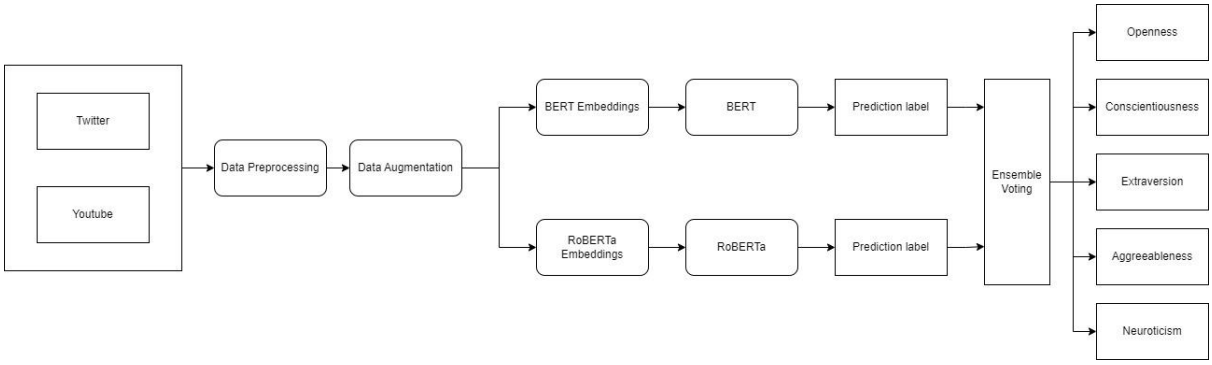

Figure 1: Architecture model

Using Twitter and Youtube as social media data, this research was composed of three phases: data collection, development, and evaluation. The details for each phase can be seen in figure 1. For the initial phase, data collected from previous research [9, 11, 12] has been collected for Twitter dataset. The label for defining the Big Five traits has been annotated by physiological experts. On the other hand, Youtube dataset was an open-source dataset from ChaLearn [7]. Before processing in the development phase, both datasets were preprocessed and augmented to get better-quality datasets.

For the development phase, pre-trained models BERT and RoBERTa were used as embedding and classification methods to produce prediction labels based on the Big Five personality traits. Predicted labels or results from each classifier were ensembled using the voting method to generate final personality labels. The ensemble method was used because ensembles can produce better predictive performance by combining multiple models [16]. After getting the prediction label, the confusion matrix was used as an evaluation metric during the evaluation phase.

3.1 Dataset

This research uses two social media datasets for the experiment. The first dataset is Twitter which was collected manually and consisted of 508 users with around 46.000 Twitter posts in Bahasa. The second dataset, first impression Youtube, consists of 10.000 short videos with text-based english transcripts. This dataset is public and downloaded from ChaLearn [7]. Both datasets are already labeled with Big Five Personality traits so they can be processed with the supervised learning method [17]. Text-based processing was applied in this research ignoring other types like video and audio. Dataset distribution for Twitter and Youtube were described in table 3 and 4.

Tabel 3: Twitter dataset distribution

values	O	C	E	A	N
High	27.921	14.365	36.187	27.572	22.871
Low	22.191	35.747	13.925	22.540	27.241

Tabel 4: Youtube dataset distribution

values	O	C	E	A	N
High	6.617	5.615	4.407	6.553	5.509
Low	3.281	4.283	5.491	3.345	4.389

3.2 Data preprocessing

Today's real-world data are highly too noisy, lost, and unsteady [18]. Because of this reason, we need to conduct data preprocessing before modeling the data. Data preprocessing was conducted to remove noise, missing values, and inconsistent data [19]. Data preprocessing consists of several steps such as data cleaning, transformation, and reduction.

Steps to perform preprocessing data in this research are:

1. Remove URL
2. Remove the symbol
3. Translate Bahasa into English Language
4. Converting letters into lowercase

5. Remove stop words
6. Lemmatization

3.3 Data augmentation

From dataset distributions in table 3 and table 4, it can conclude that both datasets have an imbalanced class. In Twitter dataset, the Conscientiousness label has 14.365 high class and 35.747 low class while the Extraversion label has 36.187 high class and 13.925 low class. So, Twitter dataset has an imbalance class in Conscientiousness and Extraversion labels. In Youtube dataset, the Openness label has 6.617 high class and 3.281 low class while the Agreeableness label has 6.553 high class and 3.345 low class. So, Youtube dataset has an imbalance class in Openness and Agreeableness labels.

In this research, the imbalance class can be handled by implementing data augmentation using the back-translation method. This method translates Twitter posts and youtube transcripts from English to Germany and then translates them back to English using T5 (Text to Text Transfer Transformer) model. The T5 model was chosen for language translation because this model can generate good paraphrases from the original language [20]. The translation was used for getting different sentences that are paraphrased from the original data. So, the additional data generated from data augmentation have paraphrased sentences to get better modeling process in the transformer classifier.

3.4 Features

In this research, a pre-trained model for feature extraction was used before modeling with ensemble BERT and RoBERTa. Because embedding was processed using BERT, the feature generated by token embedding, segment embedding, and positional embedding which part of BERT embedding process. This approach was designed to do modeling of two-way representation from left to right and right to left to get the context of the sentence.

Token embedding processed social media status concatenates with special token called classification [CLS] and separator [SEP]. The CLS token was inserted at the beginning of the sentence, and the SEP token inserted at the end of the sentence [21]. The aim of this step is to get input representation for the classification task and separate each sentence. Each word in the sentence was tokenized and mapped to corpus dimension size. Each sentence consists of 12 token representations with 768 fixed dimensions [22]. The second embedding layer used in this feature extraction is segment embedding. This embedding layer was designed to create vector representation of a sentence. If the input is only one sentence, then the segment embedded only the corresponding vector with index zero. The third embedding layer is positional embedding. This embedding layer was designed as a lookup table representing the number of long sentences. Each row of the table was a position of vector representation of the word.

Similar to BERT, RoBERTa uses token, segment, and positional embedding for extracting features. RoBERTa provides improvement from BERT because RoBERTa uses dynamic masking patterns instead of static masking and separates the segments with separation token </s>.

3.5 Model prediction

Deep learning with the transformer model was a popular method for creating the personality prediction system [23]. In transformer, each identical layer in the encoder first computes multi-headed attention between a given token and then run position to the feed-forward network [24]. The latest research was to build a model for personality prediction using transformer and produce a good result. Based on that success, this research used two transformer classifiers BERT and RoBERTa combined with an ensemble method to predict personality traits.

Input resulted from embedding processed using each classifier BERT and RoBERTa. Both classifiers use 16 batch sizes for Twitter dataset and 32 batch sizes for Youtube dataset. We use Adam optimizer with learning rate $1e^{-5}$ because the performance produces better on learning rate $1e^{-5}$. For epochs and loss function, we use 10 epochs with saving the best model to get the best performance from every epoch and binary cross entropy with logit loss which uses sigmoid activation function.

After getting the predicted class from each classifier, we use voting ensemble to produce an average combination to decide the final label for each trait. The final label will be evaluated using a confusion matrix to produce an f1 score as the evaluation result.

3.6 Evaluation metric

Classification system performance describes how good the system classified the data. The confusion matrix is one of the methods that can be used to measure the classification system performance [25]. Basically, the confusion matrix contains information that compares the results of the classification performed by the system with the predicted result.

In this research, the performance measure as a prediction result was f1 score because of imbalanced data

on the dataset. F1 score obtained from the confusion matrix with combining precision and recall formula.

4 Result

In the experiment result, f1 score performance metric was shown for each trait and average. It was shown for each model that we used in this research for Twitter and Youtube datasets.

Tabel 5: Experiment result using Twitter dataset

	M1[a]	M2[b]	M3[c]	M4[d]	M5[e]	M6[f]
O	0,672	0,655	0,671	0,645	**0,705**	0,701
C	0,500	0,443	0,721	0,704	0,509	**0,724**
E	0,813	0,827	0,734	0,759	**0,849**	0,793
A	0,705	0,671	0,809	0,803	0,709	**0,826**
N	0,610	0,563	0,570	0,526	**0,641**	0,605
Avg	0,660	0,632	0,701	0,687	0,683	**0,730**

[a]M1 represent BERT model
[b]M2 represent RoBERTa model
[c]M3 represent BERT model with augmentation
[d]M4 represent RoBERTa model with augmentation
[e]M5 represent BERT + RoBERTa model
[f]M6 represent Proposed model

Table 5 shows all scenario results including classification using one classifier, voting ensemble from two classifiers, and data augmentation for Twitter dataset. The table shows that the proposed model produced better results than BERT or RoBERTa on average. The highest result for each trait produces by ensemble BERT and RoBERTa without data augmentation and proposed models. Openness, Extraversion, and Neuroticism traits produce the best results from ensemble BERT and RoBERTa. For Conscientiousness and Agreeableness traits produce the best results from the proposed model.

From the experiment result, we can conclude that data augmentation on the proposed model produces balanced f1 score for each trait, so it can produce better performance on average results with 0,730 compared to other models. This result is around 5% higher than the model without data augmentation process. Meanwhile, the highest f1 score result was produced by the ensemble BERT and RoBERTa model for the Extraversion trait with 0,849 f1 scores.

Tabel 6: Experiment result using Youtube dataset

	M1[a]	M2[b]	M3[c]	M4[d]	M5[e]	M6[f]
O	0,735	**0,800**	0,748	0,787	0,790	0,786
C	0,649	0,693	0,599	0,695	**0,721**	0,713
E	0,493	0,406	0,687	0,700	0,573	**0,717**
A	0,744	0,791	0,731	0,793	0,788	**0,793**
N	0,624	0,681	0,601	0,679	**0,707**	0,697
Avg	0,649	0,674	0,673	0,731	0,716	**0,741**

[a]M1 represent BERT model
[b]M2 represent RoBERTa model
[c]M3 represent BERT model with augmentation
[d]M4 represent RoBERTa model with augmentation
[e]M5 represent BERT + RoBERTa model
[f]M6 represent Proposed model

Meanwhile, table 6 shows experiment results for Youtube dataset. As shown in the table above, the result for the proposed model can outperform the result from BERT or RoBERTa on average. The highest f1 score result varies for each trait. Extraversion and Agreeableness traits produce the best results from the proposed model with 0,717 and 0,793 f1 scores, respectively. Conscientiousness and Neuroticism traits produce the best results from ensemble BERT and RoBERTa models with 0,721 and 0,707 f1 scores, respectively. For the Openness trait, it produces the best results from RoBERTa classifier with 0,800 f1 scores.

Similar to Twitter, Youtube dataset also concludes that data augmentation on the proposed model produces a balanced f1 score for each trait and produces better average performance with 0,741 compared to other models. This result is around 3% higher than the model without the data augmentation process. Meanwhile, the highest performance result was produced by RoBERTa model for the Openness trait with 0,800 f1 scores.

5 Discussion

This research uses two datasets, which are Twitter and Youtube. For Twitter dataset, this research achieves an average f1 score 0,730. Although this result is still below the previous result [13], this research showed that ensemble using only two classifiers and modified dataset using the data augmentation method can provide a good f1 score. While for Youtube dataset, this research achieves an average f1 score 0,741. Compared to previous results that used Youtube dataset also, this result provided a good result and was better than the research done by [15] that

resulted best f1 score 0,612. One of the reasons for the result is this research modified the dataset to minimize imbalance class with the data augmentation method. Using data augmentation, the result of the f1 score increase about 3% for Youtube dataset and 5% for Twitter dataset.

6 Conclusion

This research shows that personality prediction using text data from social media can produce good results. Although two datasets used in this research have an imbalanced class, they can be fixed with data augmentation using the back translation method. A result from the experiment shows that the proposed model with ensemble BERT and RoBERTa as feature extraction and pre-trained model, back translation as data augmentation method can produce 0,730 average f1 scores for Twitter dataset and 0,741 average f1 scores for Youtube dataset. The back translation method using T5 increase the average f1 score performance on both datasets compared to the processing dataset without data augmentation.

For future development, the dataset used for personality prediction should have a balance class. If the dataset already balances, we don't need to have more processing time to do data augmentation. Besides that, another pre-trained model like ALBERT can be used to reduce the memory and training speed of BERT and RoBERTa [26].

Acknowledgement

We would like to thank Bina Nusantara University for supporting and assisting this research. Also thank to our colleagues who provide insight and contributions for this paper.

References

[1] Sprockets, "The Importance of Recruiting for Personality: Everything You Need to Know," https://sprockets.ai/the-importance-of-recruiting-for-personality.

[2] R. Shilpa, V. Supriya, P. Sweta, V. R. Vinaya, and S. S. Uday, "Personality prediction using Machine Learning," *Science and Engineering Journal*, vol. 25, no.7, pp. 46-53, 2021. https://saejournal.com/wp-content/uploads/2021/07/Personality-Prediction-Using-Machine-Learning.pdf

[3] R. Valanarasu, "Comparative analysis for personality prediction by digital footprints in social media," *Journal of Information Technology and Digital World*, vol. 3, no. 02, pp. 77-91, 2021. https://doi.org/10.36548/jitdw.2021.2.002

[4] Datareportal, "Digital 2021: Indonesia," https://datareportal.com/reports/digital-2021-indonesia.

[5] N. A. Utami, W. Maharani, and I. Atastina, "Personality classification of Facebook users according to Big Five Personality using SVM (Support Vector Machine) method," *Procedia Computer Science*, vol. 179, pp. 177-184, 2021. https://doi.org/10.1016/j.procs.2020.12.023

[6] N. Abood, "Big five traits: A critical review," *Gajah Mada International Journal of Business*, vol. 21, no. 2, pp. 159-186, 2019. https://doi.org/10.22146/gamaijb.34931

[7] S. Basaran, and O. H. Ejimogu, "A Neural Network Approach for Predicting Personality from Facebook Data," *Sage Journals*, vol. 11, no. 3, 2021. https://doi.org/10.1177/21582440211032156

[8] Computer Vision Center and University of Barcelona, "ChaLearn Looking at People," https://chalearnlap.cvc.uab.cat/dataset/24/description/

[9] T. Tandera, D. Suhartono, R. Wongso, and Y. L. Prasetio, "Personality Prediction System from Facebook Users," *Procedia Computer Science*, vol. 116, pp. 604-611, 2017. https://doi.org/10.1016/j.procs.2017.10.016

[10] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Personality predictions based on user behavior on the Facebook social media platform," *IEEE Access*, vol. 6, pp. 61959-61969, 2018. https://doi.org/10.1109/ACCESS.2018.2876502

[11] G. Y. Adi, M. H. Tandio, V. Ong, and D. Suhartono, "Optimization for Automatic Personality Recognition on Twitter in Bahasa Indonesia," *Procedia Computer Science*, vol. 135, pp. 473-480, 2018. https://doi.org/10.1016/j.procs.2018.08.199

[12] N. H. Jeremy, C. Prasetyo, and D. Suhartono, "Identifying Personality Traits for Indonesian User from Twitter Dataset," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 4, pp. 283-289, 2019. https://doi.org/10.5391/IJFIS.2019.19.4.283

[13] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from

multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 68, 2021. https://doi.org/10.1186/s40537-021-00459-1

[14] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M. Moens, and M. D. Cock, "Computational personality recognition in social media," *Springer Science and Business Media LLC*, vol. 26, pp. 109-142, 2016. https://doi.org/10.1007/s11257-016-9171-0

[15] F. O. Lopez-Pabon, and J. R. Orozco-Arroyave, "Automatic Personality Evaluation from Transliterations of Youtube Vlogs using Classical and State-of-the-Art Word Embeddings," *Ingeneria e Investigacion*, vol. 42, no. 2, 2022. https://doi.org/10.15446/ing.investig.93803

[16] Machine Learning Mastery, "A Gentle Introduction to Ensemble Learning Algorithms," 2021.

[17] H. Zheng, and C. Wu, "Predicting personality using facebook status based on semi-supervise learning," in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 59-64. https://doi.org/10.1145/3318299.3318363

[18] A. Souri, S. Hosseinpor, and A. M. Rahmani, "Personality classification based on profiles of social network users and the five-factor model of personality," *Human-centric Computing and Information Sciences*, vol. 8, no. 24, 2018. https://doi.org/10.1186/s13673-018-0147-4

[19] M. U. Maheswari, and J. G. R. Sathiaseelan, "Text Mining; Survey on Techniques and Applications," *International Journal of Science and Research*, vol. 6, no. 6, pp. 1660-1664, 2017. https://www.ijsr.net/get_abstract.php?paper_id=ART20174656

[20] Towards Data Science, "Paraphrase any question with T5 (Text-To-Text Transfer Transformer) - Pretrained model and training script provided," https://towardsdatascience.com/paraphrase-any-question-with-t5-text-to-text-transfer-transformer-pretrained-model-and-cbb9e35f1555.

[21] P. Kaur, G. S. Kohli, and J. Bedi. "Classification of Health-Related Tweets Using Ensemble, Zero-Shot and Fine-Tuned Language Model," in *Proceedings of the 29th International Conference on Computational Linguistic*, 2022, pp. 138-142.

[22] J. Liu, C. Xia, X. Li, H. Yan, and T. Liu, "A BERT-based Ensemble Model for Chinese News Topic Prediction," *ACM Digital Library*, pp. 18-23, 2020. https://doi.org/10.1145/3404512.3404524

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *31st Conference on Neural Information Processing Systems*, 2017. https://arxiv.org/abs/1706.03762

[24] M. E. Peters, M. Neumann, L. Zettlemoyer, and W. T. Yih, "Dissecting contextual word embeddings: Architecture and representation," in *Proceedings of the 2018 Conference of Empirical Method in Natural Language Processings*, 2020, pp. 1499-1509. https://doi.org/10.18653/v1/d18-1179

[25] A. Kulkarni, D. Chong, and F. A. Batarseh, Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering, pp. 83-106, 2020. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

[26] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv*, 2019. https://arxiv.org/abs/1909.11942