Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

Clickbait Classification Model on Online News with Semantic Similarity Calculation Between News Title and Content

Hero Akbar Ahmadi*, Andry Chowanda

Fakultas Teknologi Informasi, Universitas Bina Nusantara, Jakarta, Indonesia
Email: [1,*]hero.ahmadi@binus.ac.id, [2]achowanda@binus.edu
Correspondence Author Email: hero.ahmadi@binus.ac.id
Submitted: **26/01/2023**; Accepted: **31/03/2023**; Published**: 31/03/2023**

Abstract−Clickbait is a sensational title that makes us click internet links to an article, image, or video. Online content providers use clickbait to gain user traffic, that leads to increasing income from the placed ads in their page. To attract more and more traffic, online content providers write sensational and hyperbolic titles, and even misleading and not telling the whole story. This can give us, the internet consumer, wrong perspective, and half-truth. And nowadays, clickbait titles are worse than ever. Modern clickbait titles are not hyperbolic nor ambiguous enough, and sometimes very hard to identify. This paper aims to classify clickbait titles, to help humans identify clickbait and stop sharing more online content that contains clickbait and misleading titles. This model classifies clickbait by calculating semantic similarity between the article title and the summary of the article content. The article content is summarized by T5 (Text-to-text Transfer Transformer) model. IndoBERT is then used to calculate semantic similarity score between generated summary and the article title. The article title, content, summary, and semantic similarity score are used for clickbait classification with various algorithms. The result shows that by adding article content alongside article title in the classification process improves F1-score by 7% when classified with IndoBERT. In another future research, this model can be integrated with another application such as twitter or telegram bot to send us warning every time a user consumes online content with clickbait title. Thus, it can prevent online communities from sharing misleading information caused by clickbait.

Keywords: Clickbait; Text Summarization; Semantic Similarity; Text-to-text Transfer Transformer; IndoBERT

1. INTRODUCTION

Clickbait is a sensational title that makes us click some links to an article, image, or video [1]. Clickbait titles are often biased, have duplicate meanings, and hyperbolic. This 'technique' of writing half-truth titles is used by writers because it can be used to increase their traffic, thus getting more income from online ads. Clickbait practice is not good for us, the internet consumer. Because the title is often misleading and not telling the whole story.

Some researchers have been trying to build a model to classify clickbait. But the main problem of most of the past research is the models only use the news title to classify. This can be a problem because nowadays, clickbait titles are trickier than ever. Some clickbait titles are not so obvious nor hyperbolic to be easily classified as clickbait.

For example, in 2018 [2] one of Indonesian online news portal posted news on twitter with the title: "Bendera China Dikibarkan di GBK Saat Closing Asian Games" which translated to Chinese Flag was Hoisted at GBK at the Closing of Asian Games. This title is not hyperbolic and not exaggerated. But this title can trigger some reader because different political views and sentiment about China for some people. But after reading its whole content, that title is in fact misleading and did not tell the whole story. The flag was in fact hoisted just because China is the next country to host Asian Game. If we classify this example with a model that only reads the title, most likely it will be classified as non-clickbait, where it should've been classified as clickbait. This type of clickbait titles can be hard to identify if the classification model only uses news title as the feature.

In 2006 [3], proposed a classification model that categorizes online news titles into: ambiguous, exaggeration, vulgar diction, and several other categories. Then the model was built with Gradient Boosted Decision Trees (GBDT). This model is then improved [4] by adding some categories like Coh-Matrix and Flesh-Kincaid grade level implementations to measure language formality. The model then performs clustering using t-Stochastic Neighborhood Embedding (SNE).

In another research [5], proposed a modeling using several neural network algorithms and comparing if these algorithms are combined with Word Embedding (WE), Character Embedding (CE), or a combination of both. As a result, the Bi-Directional Long-Short Term Memory (BiLSTM) algorithm gets the best performance. Another research in 2021 [6] presents Headline2Vec, that used an extracted layer from Convolutional Neural Network (CNN) on word2vec model. This research is done in Thai dataset that achieved about 93% accuracy.

A model of clickbait convolutional neural network (CBCNN) was also proposed [7]. The CBNN used consists of 1 + |T| Word2Vec model and a CNN model, where |T| is the number of article types. Then for testing, the Word2Vec model is used to convert news headlines to a weight matrix. Another recent study in 2022 [8] developed a model to automatically explain clickbait title and websites so the users does not need to read the entire text. This research fine-tuned an abstractive model of T5 and evaluate the model using ROUGE and BERT.

A comparison of various Natural Language Processing (NLP) is done [9]. Researchers compared the performance of various machine learning algorithms: Gaussian Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes, Multilayer Perceptron, and Deep Learning with LSTMs. As a result, deep learning got the best score.

Researchers compared the performance of the Convoluted Neural Network (CNN) with word embedding from scratch (Click-Scratch) with word embedding that had previously gone through the learning process (Click-

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

Word2Vec). The result is that the Click-Word2Vec method is considered better for detecting clickbait titles [10]. Another research [11], is done by generating clickbait titles to be used later as training data. Using a method called Stylized Headline Generation (SHG) Framework. Other research also predicts the strength of clickbait in online social media [12]. This research modelled clickbait strength prediction as a regression problem.

The first ever transfer learning technique used in sensational headline generation [13]. This research is done without human annotated data and provide a novel loss function and auto tuned reinforcement learning. More research was done [14] by implementing various deep learning algorithms to detect clickbait on Youtube. The features used are then grouped into 3. The first is channel-related features (number of subscribers, total number of views on the channel, number of videos). Second, features related to video statistics (number of likes, number of dislikes, like-dislike ratio, number of comments, number of views). The third is a feature that contains text information from the video (title, description). As a result, the MLP (Multi-Layer Perceptron) algorithm with BERT embeddings produces the best performance.

Youtube clickbait research is still popular. It was done by developing the Clickbait Video Detector (CVD) method to detect clickbait on Youtube [15]. The features in the dataset are grouped into 3 groups, namely groups based on user profiling, video content, and audience consensus. As a result, the J48 algorithm produces the best score.

Research is done in 2021 also for classifying clickbait on Youtube [16]. The research is using data that is not fully labeled. Some of them, 787 data to be exact, are labeled. To label the rest of the data, the researcher generates a noisy label. This noise label is generated by considering the various features of the video.

Previous studies already have great model performance. But most of them are done in English dataset. In this paper, we will focus more to dataset with online news written in Bahasa Indonesia. Also, this paper proposed how to improve the clickbait classification in online news to not only use the title, but also the news content.

Connecting the news title and its content will give the model an advantage if the clickbait title is not so hyperbolic and quite reasonable. The proposed model summarizes the whole news content, and then calculates the semantic similarity with its title. News title, content, summarized content, and semantic similarity score are then used as features to run in various classification algorithm to determine whether the title is clickbait or not.

This model is then could be used in future work by integrating it with social media bots, for example twitter and telegram bots. This bot then can be developed to detect clickbait every time any online news is received or appears in the user's timeline. This can further prevent the spread of misleading content across social media.

2. RESEARCH METHODOLOGY

2.1 Proposed Method

The proposed clickbait classification method makes several important elements of online news as a variable. The elements used are news title and news content. These elements were chosen because they are closely related and can affect the classification result, whether the news is clickbait or not. The complete proposed method flowchart is shown in Figure 1.

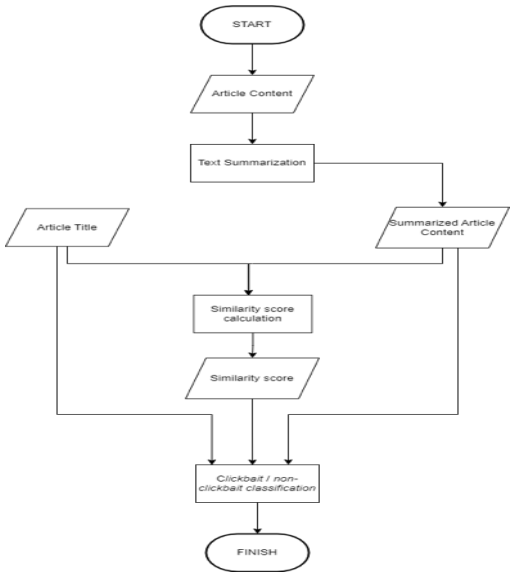

Figure 1. Proposed Method Flowchart

First, news content will be summarized using T5 (Text-to-text Transfer Transformer) [17] pre-trained model. This model is fine-tuned for summarization tasks. After that, the summary of this news content will be calculated for the similarity of the text, semantically to the news title. This text similarity score will become data which will later be

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

one of the inputs as a predictor variable to classify whether the news is clickbait or not. The news title and summary will also be the input for clickbait classification.

The dataset used in this study is the CLICK-ID [18] dataset. This dataset contains online news news from 12 online news portals in Indonesia, namely: detikNews, Fimela, Kapanlagi, Kompas, Liputan6, Okezone, Posmetro-Medan, Republika, Sindonews, Tempo, Tribunnews, and Wowkeren.

There are more than 46,000 raw data, and more than 15,000 labeled data. Labeling is carried out by 3 people, where this labeling process considers the title of the news. In this study, the data that will be used is data that has been labeled. The distribution of data in the dataset can be seen in Figure 2, and the data structure and content of the dataset is shown at Table 1.

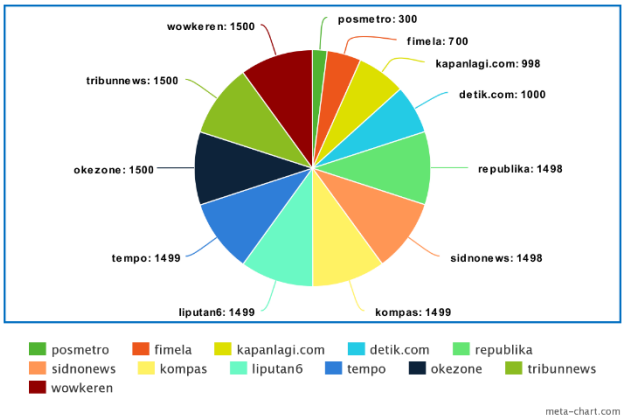

Figure 2. Click-ID data source distribution

Figure 2 shows the CLICK-ID dataset is obtained from top 12 online news portal in Indonesia. From the data source perspective, the data distribution is quite balanced. Of all 12 online news portals, 8 of them contribute about 1500 data. Only Posmetro, Fimela, Kapanlagi, and Detik.com have far less than that. Although in this research the data source will not be one of the variables to help clickbait classification, it is good to know that this dataset is obtained from reputable sources.

Table 1. CLICK-ID Dataset Column Structure

Column Name	Example Data
title	Kocak! Maling di Rumah Mewah Jakut Terekam CCTV Bingung Cari Jalan Kabur
source	detik.com
date	9/10/2019
time	18:02
category	detikNews
sub-category	Berita
content	Jakarta - Sebuah rekaman CCTV viral di media sosial menunjukkan maling di rumah mewah yang mencoba kabur setelah aksinya dipergoki pemilik rumah. Pelaku sempat kocar-kacir mencari jalan keluar dari rumah mewah itu <dan seterusnya>
url	https://news.detik.com/berita/d-4700751/kocak-maling-di-rumah-mewah-jakut-terekam-cctv-bingung-cari-jalan-kabur
label	Clickbait

Table 1 shows the data structure of CLICK-ID dataset. This data is annotated with Clickbait and Non-clickbait label. The columns that are utilized in this research are title, content, and label. Unused columns are deleted in the pre-processing phase.

2.1.1 News Content Summarization

In the proposed method, the first step is to summarize the content of the news. Another separate dataset is used to build the summarization model. In this experiment, we use IndoSum dataset to fine-tune and evaluate our model. IndoSum dataset [19] was `compiled by kata-ai in 2018 from Indonesian online news and article that publicly available. This dataset contains 71.353 training data, 3.742 validation data, and 2.347 test data. The dataset structure is shown in Table 2.

Table 2. IndoSum Dataset Column Structure

Column Name	Example Data
category	tajuk utama

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

gold_labels	[[false, true], [true, true], [false, false, false], [false, false], [false, false], [false, false], [false, false], [false], [false, false]]
id	1501893029-lula-kamal-dokter-ryan-thamrin-sakit-sejak-setahun
parahraphs	[[["Jakarta", ",", "CNN", "Indonesia", "-", "-", "Dokter", "Ryan", "Thamrin", ",", "yang", "terkenal", "lewat", "acara", "Dokter", "Oz", "Indonesia", ",", "meninggal", "dunia", "pada", "Jumat", "(", "4", "/", "8", ")", "dini", "hari", "."], ...
source	cnn indonesia
source_url	https://www.cnnindonesia.com/hiburan/20170804120703-234-232443/lula-kamal-dokter-ryan-thamrin-sakit-sejak-setahun-lalu/
summary	[["Dokter", "Lula", "Kamal", "yang", "merupakan", "selebriti", "sekaligus", "rekan", "kerja", "Ryan", "Thamrin", "menyebut", "kawannya", "itu", "sudah", "sakit", "sejak", "setahun", "yang", "lalu", "."], ["Lula", "menuturkan", ",", "sakit", "itu", "membuat", "Ryan", "mesti", "vakum", "dari", "semua", "kegiatannya", ",", "termasuk", "menjadi", "pembawa", "acara", "Dokter", "Oz", "Indonesia", "."].....

The IndoSum dataset consists of 7 columns as shown in Table 2. In this research, the utilized columns are paragraphs and summary. Paragraphs column contains a complete list of words that forms sentences and finally forms a complete article. And the summary column is a label that contains a summarization of paragraphs column.

The model used in this summarization process is based on t5-base-indonesian-summarization-cased model by Cahya [20]. This model has been fine-tuned with Liputan6 dataset, which is a huge dataset consist of online news in Indonesian. Also, this model specifically fine-tuned for Indonesian text summarization task. We fine-tuned this model using the IndoSum dataset. Adam optimizer is used with a learning rate of 3e-5. Because of the limitation in computational power, we only fine-tune the summarization model in 1 epoch.

The model is then evaluated with ROUGE score. Rouge is Recall-Oriented Understudy for Gisting Evaluation. It is a package of metrics used to evaluate text summarization model. It can compare an automatically produced summary and human generated summary. ROUGE works by computing numbers of matching N-gram between the texts. After fine-tuning the model, we finally can use it to summarize the CLICK-ID dataset. The summarization result is then appended to the original dataset and saved to be used in the next process, which is semantic similarity calculation.

2.1.2 Similarity Calculation between Title and News Content Summary

The next step is to compute the semantic similarity between the title of the news and its summarized content. This will result in a similarity score between 0-100%. This method uses semantic similarity, so it can capture the text meaning to compute the similarity.

In this experiment, we used cosine similarity with IndoBERT as the sentence embedding, to compute the similarity between news title and the summarized news content. The result is expected to be an input for the next step which is a classification model.

2.1.3 Clickbait Classification

The final step of this entire model is to classify the data. This will be a binary classification, whether a news is a clickbait, or not. In this experiment, we evaluate and compare some classification methods. For example, classification with only similarity score as the input. Or including title and summary to the input. Also, we use some several classification algorithms including deep learning algorithms like LSTM.

Before the classification process is done, the news content needs to be preprocessed by removing any irrelevant information, such as city name in the first word of the news. For some classification methods like LSTM and Random Forest, the input needs to be preprocessed more, such as case folding and stop words removal. This is done to improve the model's performance. The Classification models are then applied to the data, with 80% of them are used for training. The model is then evaluated by calculating precision, recall, and F1-score.

3. RESULT AND DISCUSSION

3.1 News Content Summarization Result

The first process in this experiment is news summarization. Before generating summarization of the news content, the t5-base-indonesian-summarization-cased model is fine-tuned with IndoSum dataset. This fine-tuning model then evaluated by ROUGE score, with the result shown in Table 3.

Table 3. Summarization Model Testing Evaluation

ROUGE	Precision	Recall	F1-Score
ROUGE-1	77%	71%	74%
ROUGE-2	70%	63%	66%
ROUGE-L	75%	69%	72%

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

From the result that shown at Table 3, the fine-tuning process produce a model with F1-Score of 74% for ROUGE-1 evaluation, 66% for ROUGE-2 and 72% of ROUGE-L.

After fine-tuning process, this model is then used to generate the news content summary of the main dataset, CLICK-ID. The generated summary then analyzed. This analyzation is done to prevent any generated summary that is unexpected. From this analyzation process, we found 48 data that failed the summarization process. This caused by the data is incomplete. The examples of the incomplete news content found in the CLICK-ID dataset is shown at Table 4.

Table 4. Example of Incomplete News Content at IndoSum Dataset

No	Title	Content
1	Mengenal Jenis Masker yang Direkomendasikan untuk Korban Kebakaran Hutan	-
2	Menpora Harap Delegasi Pemuda Indonesia Promosikan Olahraga dan Budaya di Jepang	<Nan>
3	Alasan Black Widow Merelakan Diri di 'AVENGERS ENDGAME'	-,,,
4	Putrinya Sienna Wisuda, Marshanda Foto Keluarga Bareng Istri Ben Kasyafani	-,adalah salah satu sosok ibu yang selalu memberikan kasih sayang berlimpah untuk sang putri, Sienna Ameerah Kasyafani. Putri hasil perkawinannya dengan presenter dan aktor,itu baru saja wisuda.,Lewat akun Instagram-nya, wanita yang biasa disapa Caca ini pun menunjukkan kebahagiaan dan kebanggaannya pada prestasi Sienna. *(truncated)*

Table 4 shows four examples of incomplete news content inside IndoSum dataset. The example data number 1, its content is only consist of dash (-) character. The content of example data number 2 only consists of a word <Nan>. Other incomplete data is also similar with example number 3, which only contains some symbols. Another example of incomplete data is example data number 4, which the starting word is off and incomplete.

All data with incomplete news content is then removed from the processed dataset. After the removal process, the final dataset contains 14.790 data remaining.

3.2 Similarity Calculation Between Title and News Content Summary Result

After generating news content summary, we compute the semantic similarity between the title and the generated summary. This process is done by calculating cosine similarity. We use IndoBERT [21] as the sentence embedding. The news title, content, summarized content, and similarity score are then used as an input to the classification model.

3.3 Clickbait Classification Result

In this research, the classification process is done through various experiments. Every experiment uses at least one of these features: news title, news content, the similarity score, or a combination of two or more features. Every experiment also classified using several different classification algorithms. All the models are then evaluated by calculating precision, recall, and F1-score.

3.3.1 Experiment: Classification with similarity score as the only feature

The first experiment in this research is done using similarity score as the only feature for classification. For this first experiment, there are 7 classification algorithm that are applied: Logistic Regression, SVM, Naive Bayes, Decision Tree, Random Forest, Threshold, and LSTM.

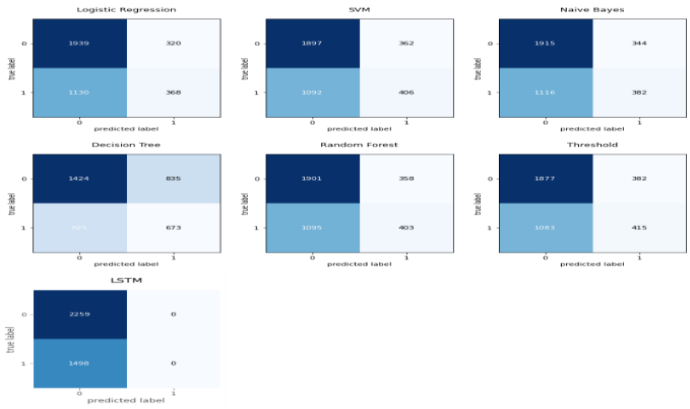

Figure 3. Confusion Matrix of Classification with Similarity Score as the Only Feature

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

From the confusion matrix in Figure 3, there are a lot of false negatives and some less number of false positives. From this first experiment, decision tree is the best performer with F1-score of 0,45 (Table 5). LSTM always produces non-clickbait label, so it is algorithm with the worst F1-score. The accuracy is ranged between 0,56 to 0,61. This result shows that similarity score with the only feature is not good enough to be used for clickbait classification for this research. Other features are needed to improve the classification result.

3.3.2 Experiment: Classification with news title as the only feature

The next experiment is to classify the data using the news title as the only feature. The data classified with three different algorithms: Random Forest, IndoBERT, and LSTM.

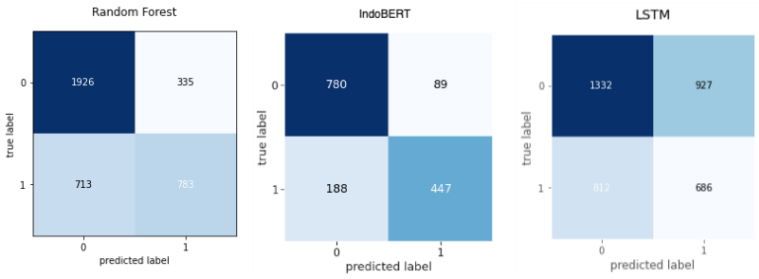

Figure 4. Confusion Matrix Of Classification With News Title As The Only Feature

In the classification with news title as the only feature, IndoBERT reached the highest score of 0.76 (Table 5). Meanwhile, the LSTM algorithm got the worst F1-score with 0.44. The accuracy ranges from 0.52 to 0.82. From the confusion matrix in Figure 4, the IndoBERT algorithm shows that the number of false predictions decreases to 188 for false positive, and 89 for false negative. Or in total, there is about 18% false predictions with IndoBERT. That's better compared to previous experiment, where the classification only use similarity score as the only classification feature. BERT also outperformed other algorithms in F1-score.

3.3.3 Experiment: Classification with news summary as the only feature

In this experiment, the summarized news content is used as the only feature to classify. The algorithm used in this experiment is Radom Forest, IndoBERT, and LSTM.

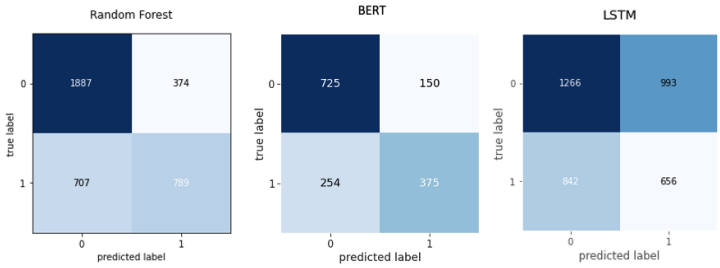

Figure 5. Confusion Matrix Of Classification With News Summary As The Only Feature

The highest F1-score is reached with IndoBERT algorithm with 0.65 (Table 5). Meanwhile, the LSTM algorithm got the worst F1-score with 0.42. The accuracy obtained from various classification algorithms ranges from 0.51 to 0.73. The confusion matrix in Figure 5 shows once again that the IndoBERT algorithm has a better percentage of true predictions and F1-score compared to other algorithms.

3.3.4 Experiment: Classification with title and news content as the features

In this experiment, more than one feature is used to train the classification model. News title and the entire content is two features that used in this experiment. The algorithm used in this experiment is Radom Forest and IndoBERT.

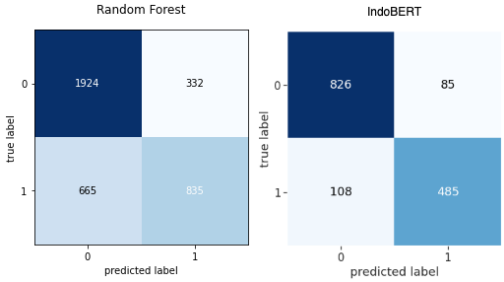

Figure 6. Confusion Matrix Of Classification With News Title And Content As Features

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

While combining news title and the entire content for classification, the classification result shown that once again IndoBERT outperforms other algorithms. The F1-score obtained by the IndoBERT algorithm is 0.83 (Table 5), which is 0.20 points higher than the Random Forest algorithm. The confusion matrix in Figure 6 also shows the decreasing number of false positives and false negatives in the IndoBERT algorithm which is in line with the high F1-score.

3.3.5 Experiment: Classification with title and summarized news content as the features

Once again, a combination of two features is tested to train the classification model in this experiment. This time, the two features are the news title and the summarized news content that is used. In this experiment, the model trained by three algorithms: Random Forest, IndoBERT, LSTM.

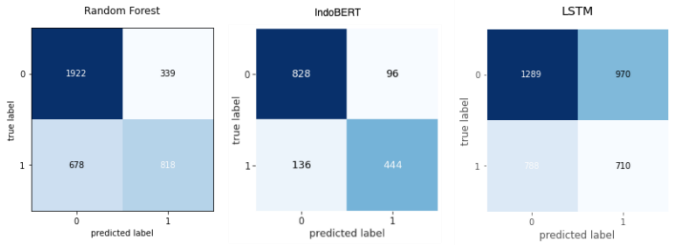

Figure 7. Confusion Matrix Of Classification With News Title And Summarized Content As Features

Combining title and the summary of news content as features again shows the superiority of the IndoBERT compared to other classification algorithms. IndoBERT gets the highest score with an F1-Score of 0.79 (Table 5). The confusion matrix picture in Figure 7 shows the improvement in the percentage of true negatives in all algorithms, when compared to the classification using only news titles. So that the percentage of false negatives also decreases.

3.3.6 Experiment: Classification with title, summarized news content, and similarity score as the features

The last experiment in this research is to combine all the features to train the classification model. The news title, summarized news content, and the similarity score between them is combined. LSTM and IndoBERT also chosen as the classification algorithm.

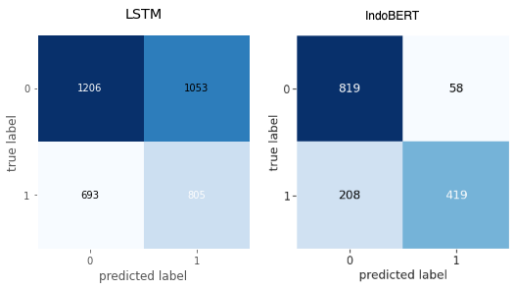

Figure 8. Confusion Matrix Of Classification With News Title, Summarized Content, And Similarity Score As Features

The classification algorithm with IndoBERT is still the best in this last experiment. IndoBERT outperformed LSTM with an F1-score of 0.76 (Table 5), 0.28 points higher than LSTM. It can be seen in the confusion matrix in Figure 8, with LSTM, the percentage of false positives tends to be higher when compared to the LSTM with a combination of features that have been tried in other experiments explained before. As for the IndoBERT algorithm, it still tends to be better in its true positive and true negative.

3.3.7 Final Clickbait Classification Results Discussion

The 6 experiments in the previous subchapter shows that different combination of features can produce model with various performance. The complete result of classification process is shown in Table 5:

Table 5. Classification Results

Features	Model	Accuracy	Precision	Recall	F1-Score
Similarity score	Decision Tree	0.45	0.45	0.45	0.45
	SVM	0.61	0.53	0.27	0.36
	Naïve Bayes	0.61	0.53	0.26	0.34
	Logistic Regression	0.61	0.53	0.26	0.34
	Random Forest	0.61	0.53	0.27	0.36
	LSTM	0.52	0.43	0.46	0.44
Title	Random Forest	0.72	0.70	0.52	0.60

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

	IndoBERT	0.82	0.83	0.70	0.76
	LSTM	0.52	0.43	0.46	0.44
Content Summary	Random Forest	0.71	0.68	0.53	0.59
	IndoBERT	0.73	0.71	0.60	0.65
	LSTM	0.51	0.40	0.44	0.42
Title + Content	Random Forest	0.74	0.73	0.56	0.63
	IndoBERT	**0.87**	**0.85**	**0.82**	**0.83**
Title + Content Summary	Random Forest	0.74	0.75	0.54	0.63
	IndoBERT	0.85	0.82	0.77	0.79
	LSTM	0.54	0.44	0.49	0.47
Title + Content summary + Similarity score	LSTM	0.82	0.88	0.67	0.76
	IndoBERT	0.54	0.43	0.54	0.48

The experiment shown that if trained with only the similarity score, the model only reached a maximum F1-score of 0.45. This maximum score is reached with the decision tree algorithm. This is worse if compared to another one feature classification experiment. The experiment which the model trained with title as the only feature, performed better with 0.76 F1-score. So are the model trained with content summary as the only feature, it was 19% better by scoring 0.65 of F1-score.

The classification model that uses one feature only produce maximum performance of 0.76 F1-score, with news title as the feature and IndoBERT as the classification algorithm. Compared to other classification experiments that involve two features like Title + Content and Title + Content Summary that reached 0.83 and 0.79 of F1-score respectively. This means the model performance is improved by adding the news content, whether summarized or not. From the F1-score perspective, there is 3% improvement when the news title is combined with the summarized news content, and 7% improvement when the news title is combined with the entire news content.

The process of summarization of the news content turns out lead to decreasing model performance by 4% of F1-score. This is mainly because the summarization method in this research is extractive. Summarizing the news content also means decreasing the number information the model can gain. That can cause the model to miss important information that contains in the last sentences of the paragraphs. So, the model was not thoroughly trained.

The IndoBERT model outperformed another model in almost all experiment. In the experiment with only title as the classification feature, IndoBERT improves the F1-score to 16% and 32% if compared to random forest and LSTM respectively. The same case also occurred in other experiment. IndoBERT model tends to perform better by a quite far margin. The exception is in the last experiment with 3 features (title + news summary + similarity score). The IndoBERT model is outperformed by LSTM model.

From all the experiment, the IndoBERT model with title + news content as the features for the classification model is the best performer, with precision of 0.85 and recall 0.82, with F1-Score of 0.83. This proves by adding the news content as the classification feature, can improve the model performance by 7% if compared to models that only use news title to classify clickbait.

4. CONCLUSION

With this research, we know that the addition of news content improves the performance of clickbait classification model. Adding the news content alongside its title, produces a model with precision of 0.85 and recall 0.82, with F1-Score of 0.83. This is 7% better compared to model with only news title as the classification feature. Also, from this experiment we know that the summarization process of the news content is not necessary. As the model can perform well even with including the entire content. This will solve modern clickbait news that is not enough to classified with only considering the news title. While this research increases the model performance, this model is currently cannot trained or tested by news with Indonesian slang words. So, this can be another research for building model with dataset consisting of slang words.

REFERENCES

[1] G. C. Foundation, “What is clickbait?,” 24 June 2021. [Online]. Available: https://edu.gcfglobal.org/en/thenow/what-is-clickbait/1/.
[2] D. Y. Hadiyat, “Clickbait on Indonesia Online Media,” Pekommas, vol. 4, p. 4, 2019.
[3] P. Biyani, K. Tsioutsiouliklis dan J. Blackmer, “"8 Amazing Secrets for Getting More Clicks": Detecting Clickbaits in News Streams Using Article Informality,” 2016.
[4] A. Pujahari dan D. S. Sisodia, “Clickbait Detection using Multiple Categorization Techniques,” 2020.
[5] A. Anand, T. Chakraborty dan N. Park, “We used Neural Networks to Detect Clickbaits: You won’t believe what happened Next!,” 2019.
[6] N. Kaothanthong, S. Kongyoung dan T. Theeramunkong, “Headline2Vec: A CNN-based Feature for Thai Clickbait Headlines Classification,” INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY, vol. 5, 2021.
[7] H.-T. Zheng, J.-Y. Chen, X. Yao, A. K. Sangaiah, Y. Jiang dan C.-Z. Zhao, “Clickbait Convolutional Neural Network,” 2018.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 4, Maret 2023 Page: 1986−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i4.3030

[8] O. Johnson, B. Lou, J. Zhong dan A. Kurenkov, "Saved You A Click: Automatically Answering Clickbait Titles," arXiv:2212.08196, 2022.
[9] S. Manjesh, T. Kanakagiri, V. P, V. Chettiar dan S. G, "Clickbait Pattern Detection and Classification of News Headlines using Natural Language Processing," 2017.
[10] A. Agrawal, "Clickbait Detection using Deep Learning," 2016.
[11] K. Shu, S. Wang, T. Le, D. Lee dan H. Liu, "Deep Headline Generation for Clickbait Detection," 2018.
[12] V. Indurthi, B. Syed, M. Gupta dan V. Varma, "Predicting Clickbait Strength in Online Social Media," Proceedings of the 28th International Conference on Computational Linguistics, p. 4835–4846, 2020.
[13] P. Xu, C.-S. Wu, A. Madotto dan P. Fung, "Clickbait? Sensational Headline Generation with Auto-tuned Reinforcement Learning," Center for Artificial Intelligence Research (CAiRE), 2019.
[14] R. Gothankar, F. D. Troia dan M. Stamp, "Clickbait Detection in YouTube Videos," 2021.
[15] D. Varshney dan D. K. Vishwakarma, "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences," 2021.
[16] T. Xie, T. Le dan D. Lee, "CHECKER: Detecting Clickbait Thumbnails with Weak Supervision and Co-teaching," 2021.
[17] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li dan J. P. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv:1910.10683, vol. 3, 2020.
[18] A. William dan Y. Sari, "CLICK-ID: A novel dataset for Indonesian clickbait headlines," Data in Brief, vol. 32, 2020.
[19] Kata.ai, "GitHub - kata-ai/indosum," 2018. [Online]. Available: https://github.com/kata-ai/indosum. [Diakses 9 August 2022].
[20] Cahya, "Hugging Face - cahya/t5-base-indonesian-summarization-cased," [Online]. Available: https://huggingface.co/cahya/t5-base-indonesian-summarization-cased. [Diakses 9 August 2022].
[21] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar dan A. Purwarianti, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, p. 843–857, 2020.