Universiteit Leiden

Master Computer Science

[Natural Language Processing Methods for Supporting Indonesian Fact Checkers]

Name: [Brian Arnesto Sitorus]
Student ID: [S3277224]

Date: [07/01/2023]

Specialisation: [Computer Science:Data Science]

1st supervisor: [Dr. Suzan Verberne]

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

In this paper, we investigate the task of fake news classification and claim verification using a manually annotated dataset from an Indonesian fact-checking organization. We study how Natural Language Processing (NLP) methods can assist human fact-checkers in debunking the massive flow of misinformation in Indonesia. For fake news classification, we compare three term-based classifiers – Logistic Regression, Support Vector Machines (SVM), and Random Forest – to the IndoBERT model, a pre-trained BERT model for Indonesian that we fine-tune on our data. We compare a fine-tuned IndoBERT to GPT-3.5 with in-context learning (prompted with instructions and a few examples) for the claim verification task. Both tasks are quite challenging since the dataset is imbalanced. In the fake news classification task, SVM achieves the highest precision and IndoBERT the highest recall. Over- or undersampling to fix the class imbalance does not improve the results and even lowers the classifier's precision. In the claim verification task, the fine-tuned IndoBERT model performs substantially better than GPT 3.5, with a 94% accuracy. We think that Indonesian fact-checkers can be helped in their work by computational support, and we propose a workflow for supporting human fact-checkers with NLP tools. We will release the fine-tuned IndoBERT model and our code for future research.

Keywords: Factchecking, Claim verification, Indonesian

Contents

Chapter 1

Introduction

People are increasingly seeking and consuming news from social media platforms rather than traditional news organizations, as the amount of time spent online interacting via social media platforms continues to increase [42]. Social media use has spread throughout the world, particularly in Indonesia. This is shown by the fact that, as of January 2023, Indonesia has the third-highest number of Facebook users in the world [45]. Furthermore, another social media giant, Instagram, has 89 million active users in Indonesia, making Indonesia its fourth-largest market in the world in 2023. According to the data from Statista [46], there will be 210 million users of mobile devices, or smartphones, in 2021. This explains the dramatic increase in social media users as well since many people use their smartphones to use social media to connect with other people and also get information.

Furthermore, as social media and the market for smartphone devices grow, anyone can now post about anything without having their post validated. This is good because everyone has an equal right to express themselves, and information can be spread quickly. However, this could result in inaccurate information if neither a person nor any tools are available to validate it. People frequently publish or share the posts of others without verifying the source or assessing the information's validity or reliability. The majority of the time, a captivating headline is sufficient for an article to go viral, even if its body contains unsupported or completely fabricated claims [6]. For example, following the announcement of the results of the recent Indonesian presidential election in 2019, misinformation has been widely spread on social media, particularly in Indonesia's capital (Jakarta). A lot of information claims that the election is full of fraud and that many people will rally immediately. During the two days demonstration, at least 737 people were injured, and 50 died. Furthermore, the government must shut down social media and slow down the internet to contain this misinformation and prevent further

chaos.

There is a need for an automated tool that could detect and validate information, especially in situations like this. Situations like this could be avoided if there is automated tools that people could use to verify the information that they received to prevent an event that was previously mentioned. When identifying fake news, one critical component is categorizing the article's or post's stance into commonly used categories such as "favor", "against", and "neither". We call a post or tweet a rumor if no one has checked it out or if it has been checked out so quickly that it makes people doubt the post.

In Indonesia, there is already a manually annotated fact-checking application that we can use to verify information. There are numerous of them, such as Cekfakta from Tempo, Turnbackhoax and Hoax Buster from Mafindo, and so on. Furthermore, the Indonesian Ministry of Communication and Technology frequently publishes a list of fake news on its website. [1]

The main problem with this method is that they cannot check all the misinformation on Indonesian social media and other online platforms due to their limited resource. This study will focus on developing tools to automate specific tasks in this institution's manually annotated fact-checking task to speed up the process. This thesis could ultimately lead to the development of NLP tools to assist human fact-checkers in their work.

One of the most crucial parts of manually investigating fake news articles is retrieving relevant evidence and verifying the veracity of the claims made in relation to that evidence. This task is extremely time-consuming and tedious as fact-checkers must validate each retrieved evidence individually. However, this process can be automated using Natural Language Processing methods, which would significantly increase the efficiency of fact-checkers.

In general, the main research questions of this study are described as follows:

- To what extent can traditional machine learning techniques such as Support Vector Machines, Logistic Regression, Random Forests, and deep learning models such as the IndoBERT Model be utilized for fake news detection and classification into the respective categories?
- To what extent can BERT and GPT models, particularly IndoBERT for the Indonesian language, be utilized in claim verification as a preliminary step in manually fact-checking fake news, and how do these models perform when applied for claim verification tasks with Indonesian language articles?

[1] https://m.kominfo.go.id/content/all/laporan_isu_hoaks

Two approaches are used in the experiment to address the issue: manually annotated fake news and computational modeling for fake news detection. An interview is conducted with the institution that did the manual annotation to gain insight into how the work was manually annotated and the disadvantages of the method. Furthermore, after identifying the problem, NLP is used to optimize the human-manual annotated workflow.

The research methodology follows: Chapter 2 discusses numerous scientific works related to this investigation. The definition of fake news and the current trend are discussed. In addition, the background research on the multiple classifiers used in the experiment is described. The company visit and findings of the company visit are described in Chapter 3. Chapter 4 describes the data used in this study, how it was obtained, and the prepossessing step. Chapter 5 elaborates on the methodology for each experiment. The results of the experiment and a comparison of the performance of various classifiers used in this study are discussed in Chapter 6. Finally, Chapter 7 provides an overview of the conclusion, discussion, and potential future work.

Chapter 2

Background

In this chapter, we explore the definition of fake news and the trends, relevant dataset for fake news classification and claim verification, and existing method related to our task.

2.1 Definition of fake news and the trends

There are numerous definitions of fake news, making it difficult to precisely define the term since there are many types of disinformation. The purpose of fake news continues to evolve through the research's emphasis. Baptista and Gradim [4], for example, says that "fake news is a type of online disinformation with misleading and false statements that may or may not be related to actual events, intentionally made to mislead and manipulate a specific or imagined public through the appearance of a news format with an opportunistic structure (title, image, content) to attract the reader's attention, to get more clicks and shares and a wider reach".

In addition, Allcott and Gentzkow [3] define fake news as "intentionally and verifiable false news articles that may mislead readers". However, other research defines the term as a news article or message published and disseminated through the media that contains false information, regardless of the meaning and motivations [41, 31, 19]. Even though multiple studies have been conducted to define fake news, it is still rather complex. There is a significant disagreement about which types of content should be considered "fake news" and which should be excluded. This is especially true given that the term "fake news" has taken on a political connotation and is frequently used to attack the credibility of news organizations or to argue against commentary that differs from our [34, 50].

Furthermore, with the growth of social media, this platform has become the epicenter of fake news dissemination. While the social media platform developer still does some moderation, it is insufficient to moderate the massive flow of fake news on social media. One research study by Vosoughi et al., [54] discusses spreading true and false news on Twitter between 2006 and 2017. The author discovered that fake news spreads more rapidly than real news and is more novel than real news, suggesting that people are more inclined to share novel information. They also found that people's reactions to false stories were fear, disgust, and surprise. In contrast, true stories evoke anticipation, sadness, joy, and trust.

2.2 Dataset

We delve into the largest and most commonly used datasets for fake news classification and claim verification tasks, described as follows.

2.2.1 Fake news classification dataset

We summarise some of the existing fake news classification datasets as shown in table 2.2.1. This dataset contains various topics; there are two English-language datasets and one Indonesian-language dataset.

ISOT Fake News Dataset The ISOT Fake News Dataset comprises more than 12,600 articles from real-world sources, including both real and fake news. The truthful articles were obtained by crawling from Reuters.com, while the fake news articles were collected from unreliable websites flagged by Politifact and Wikipedia. With a focus on political and world news, the dataset encompasses various categories. The "True.csv" file contains over 21,400 articles, including "World-News" (10,145 articles) and "Politics-News" (11,272 articles). The "Fake.csv" file comprises more than 23,400 articles, categorized into "Government-News" (1,570 articles), "Middle-east" (778 articles), "US News" (783 articles), "left-news" (4,459 articles), "politics" (6,841 articles), and general "News" (9,050 articles). The dataset has been cleaned and processed, preserving the original punctuation and mistakes in the fake news text. It provides a valuable resource for studying and analyzing real and fake news characteristics and patterns, particularly in politics and world news [2, 1].

WELFake The WELFake dataset is a compilation of 72,134 news articles, including 35,028 real and 37,106 fake news articles. It was constructed by combining four prominent news datasets - Kaggle, McIntire, Reuters, and BuzzFeed Political. This merging was done to prevent classifier overfitting and provide more text data for improved

Dataset	Context	Label Space
ISOT Fake News	Article	Fake, Real
WELFake	Article	Fake, Real
The 2019 Indonesian Presidential Election Tweets	Tweet	Fake, Real, Misleading, Other

Table 2.1: Dataset for Fake News Classification

machine-learning training. The dataset comprises four columns: Serial number, which starts from 0; Title, representing the news headline; Text, providing the news content; and Label, which signifies the authenticity of the news, where 0 denotes “fake news” and 1 represents “real news” [53].

The 2019 Indonesian Presidential Election Tweets The dataset of tweets from the 2019 Indonesian Presidential Election, collected between September 23rd, 2018 (the beginning of the campaign) and May 28th, 2019 (a week after the results were declared), comprises 1,733 data points. The author has manually classified these data points into four categories: “True News” (896), “False News” (648), ”Misleading News” (189), and ”Other” (438) [47].

2.2.2 Claim verification dataset

We summarize some of the datasets for the claim verification task as shown in table 2.2.2. The ClaimBuster dataset is designed for the pre-preliminary step of claim verification where it ranked the data point from a scale of 0 to 1 and the closest score to 1 means that this data point is important to do further check. Meanwhile, both The SemEval-2016 Task 6 and the FEVER dataset were annotated directly for claim verification task purposes.

SemEval-2016 The SemEval-2016 Task 6, or Stance Detection dataset, provides a tool for claim verification on Twitter, where tweets are categorized into “favor”, “against”, or “neutral” towards certain topics. This dataset, with annotated tweets, enables machine learning models to be trained to recognize sentiments in social media content. Five diverse topics were selected - Atheism, Climate Change, the Feminist Movement, Hillary Clinton, and Abortion Legalization - to cover various perspectives and sentiments. The main aim is to advance models that not only verify claims but also understand public sentiment towards specific topics, aiding in a deeper analysis of online discourse [30].

FEVER The FEVER dataset, developed by Thorne et al. [51] is a benchmark dataset for claim verification task. It incorporates a claim with the associated truth label, such as “Supported”, “Refuted”, or “NotEnoughInfo”, and relevant evidence from Wikipedia, if available. The dataset aids in executing two operations: pinpointing documents carry-

Dataset	Context	Label Space
SemEval-2016 Task 6	Tweet	Favor, Against, Neutral
FEVER	Article	Supported, Refuted, NotEnoughInfo
ClaimBuster	Tweet	0 to 1

Table 2.2: Dataset for Claim Verification

ing pertinent evidence and assessing if the claim is corroborated or contradicted by the evidence. Though it poses some challenges, it has significantly advanced the field of automated fact-checking and claim verification. This dataset is beneficial in two main ways: it helps locate documents that contain relevant evidence, and it aids in determining if the evidence either supports or contradicts a given claim. Despite certain challenges, it has greatly enhanced the process of automatic claim verification.

ClaimBuster The ClaimBuster dataset is a valuable dataset designed to streamline and automate the process of fact-checking. This dataset consists of a large corpus of manually annotated sentences extracted from various domains, such as political debates, speeches, and television shows, and manually annotated based on their verifiability. Each sentence is assigned a numeric score between 0 and 1 indicating its verifiability. A score close to one indicates that the sentence will likely contain a factual claim that requires verification. As a result, the ClaimBuster dataset does not directly provide claim veracity but identifies potential claims that merit further investigation. It provides an essential first step in the claim verification process by highlighting sentences that are likely to contain factual information that can be verified [16].

We discovered that there is no extensive collection of multilabel datasets for both tasks in the Indonesian language and most of the available data is in the English language. The research conducted by Suhardika et al. [47] created manually annotated datasets for Indonesian language fake news classification. However, this dataset is limited to two main categories: real and fake news. The categorization of such data is more complex than simply dividing it into these two categories. Classifying the dataset into more distinct categories would be particularly useful in identifying the characteristics of fake news.

2.3 Text Representation using TF-IDF

TF-IDF (term frequency-inverse document frequency) is a standard method for showing text representation when working with text datasets. TF-IDF is a measure used in information retrieval (IR) and machine learning to quantify the importance or relevance of string representations (words, phrases, lemmas, etc.) in a document relative to a

collection of documents (also known as a corpus) [11]. TF-IDF can be divided into TF (Term Frequency) and IDF (Inverse Document Frequency).

A document's term frequency (TF) is calculated by dividing the total number of words in the document by the number of occurrences of a particular word.

There are several measurements or approaches to consider when defining frequency:

- The frequency of a particular word appearing in a text (raw count).
- The term frequency was modified based on the total document length (raw count of occurrences divided by several words in the document).
- The frequency expressed on a logarithmic scale, such as log(1 + raw count).
- The frequency of occurrence of the Boolean value (e.g., one of the terms occurs, or 0 if the term does not occur in the document) [37].

Term frequency can be calculated using the formula below.

$$\text{tf}(t, d) = \frac{\text{count of } t \text{ in } d}{\text{number of words in } d}$$

Each term is given the same weight as the others when calculating TF. However, certain words, for example "is", "of," and "are, "are frequently used in sentences but do not have a meaning because they are only used to help create a logical and coherent sentence by linking the words, phrases, and clauses together, indicating the relationship between them. Therefore, we should give less weight to terms that appear often in the document set and more weight to terms that appear less often. This is done by weighting the inverse document frequency factor (IDF), which gives less weight to words that often appear in the document set and more weight to terms that don't appear very often [25].

The following formula can be used to calculate IDF: N is the number of documents in the collection, and df is the number of documents containing the term t.

$$\text{idf}(t) = \frac{N}{\text{df}}$$

2.4 Automatic Fake News Classification

Numerous approaches are currently widely used on fake news classification tasks. This study uses the traditional machine learning classifier approach and the BERT model.

Several studies on these two approaches are described below.

2.4.1 Traditional Machine Learning Classifiers

Machine learning has been used for various tasks, including detecting fake news. Several machine learning-based strategies for automatically detecting fake news have been proposed. The study conducted by Shu et al. [42] suggested using linguistic-based features such as total words, characters per word, frequencies of significant words, frequencies of phrases, i.e., "n-grams," bag-of-words approaches [14], and parts-of-speech (POS) tagging to identify fake news. Mahir et al. [28] used multiple machine learning approaches to detect fake news on Twitter, including Support Vector Machines, the naive Bayes method, logistic regression, and recurrent neural network models, to demonstrate the efficiency of the classification performance on the dataset. Their experiment shows that SVM and Naive Bayes outperform the other classifiers. This demonstrated that even the most basic machine-learning algorithm can detect fake news and produce the desired result.

Furthermore, another approach to this problem is conducted by Zhao et al. [21], where they focus on identifying fake news through a two-stage examination: the characterization stage and the disclosure stage. In the first stage, the fundamental concepts and guidelines underlying fake news are brought to the forefront of social media. The second stage, the discovery stage, investigates the various supervised learning algorithms currently used to determine the most effective algorithm for detecting fake news.

We want to employ three traditional machine learning classifiers for the fake news classification task: Logistic Regression, Support Vector Machines (SVM), and Random forests.

1. **Logistic Regression classifier**
   Logistic regression is a method for calculating the probability of a discrete outcome given an input variable. The most important factor in this procedure is the input variable. The most common type of logistic regression model is a binary outcome, which can have only one of two possible values, such as true or false, yes or no, and so on. When a scenario has more than two discrete outcomes, multinomial logistic regression can be used to model the situation. Logistic regression is a useful analysis method for classification problems that require determining whether a new sample belongs to a specific category. This difficulty arises when analyzing large amounts of data [13].

2. **Random Forest**
   Random Forest is a type of classifier that improves the prediction accuracy of a

dataset by averaging the results of applying multiple decision trees to different subsets of the dataset. The random forest method does not rely on a single decision tree; instead, it collects the results of each tree and bases its prediction of the final output on the tree that received the most votes for its forecast.

The random forest algorithm works by dividing the procedure into two steps: the first step is combining the N decision tree with the construction of the random forest, and the second step is creating predictions based on each tree previously created in the first stage.

3. **Support Vector Machines (SVM)**
   Support Vector Machines (SVM) is a machine learning classifier that could be used for regression and classification problems in supervised learning. This algorithm is mainly used for classification problems. The main goal of SVM is to find the most optimal line or decision boundary for categorizing the n-dimensional space so that the new data point can be easily assigned to the correct category. The hyperplane is used as the decision boundary. SVM chooses the extreme ends and vectors that contribute to the formation of the hyperplane. These severe cases are known as support vectors, so this algorithm is called Support Vector Machines.

Imbalanced learning

Furthermore, since the dataset is imbalanced, we need to implement an imbalanced learning algorithm to tackle this data imbalance issue. When dealing with data imbalances and implementing them with traditional machine learning, two main approaches are mainly used: Oversampling and under-sampling, which will be further explained below.

1. **Oversampling with SMOTE**
   When implementing oversampling with the unbalanced class, the model will increase the number of observations where this was generated by duplicating the previous sample from the larger data pool. This process needs sufficient data samples to achieve a good result. Oversampling can be done by simply duplicating the existing elements of the majority class in the training set. Nevertheless, this technique is well known to be susceptible to over-fitting [9, 32]. To mitigate this issue, additional samples could be generated artificially by considering the minority class's distribution. The Synthetic Minority Oversampling Technique (SMOTE) is one approach to address this problem. SMOTE works by selecting close examples in the feature space, drawing a line between them, and finally drawing a

new sample at a point along the line [8].

2. **Down sampling** Down-sampling is selecting and discarding majority class observations at random to prevent the majority class from dominating the algorithm on the training phase [38].

   (a) **Random under sampling** To create a balanced data set, the Random Under Sampler function uses a random number generator to remove the majority of class occurrences [39]. The modified training dataset has fewer examples in the class containing most of the data. This is a factor of two reductions. This procedure can be repeated as many times as necessary to achieve the desired class distribution, such as having an equal number of samples in each of three different classes [27].

   (b) **NearMiss**

   This is known as a near-miss algorithm; one used to help balance an unbalanced dataset. This is a method for data balancing, which ensures that the data are evenly distributed. This is achieved by analyzing the distribution of the larger class and then randomly selecting samples from that distribution. When two points in the distribution belong to different types and are relatively close, this strategy removes the data point from the larger class to restore the distribution's balance. In contrast to Random Undersampling, which selects samples from the majority class at random and without regard for rules, NearMiss employs several heuristics to ensure that samples from the majority class are representative of the actual data [59]. The NearMiss algorithm has three versions, as shown in figure 2.1.

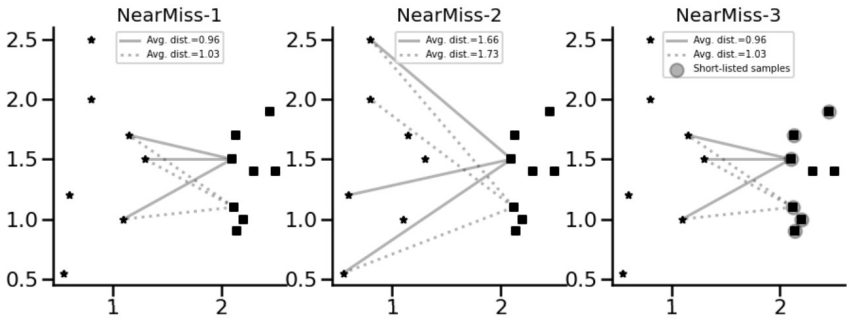

Figure 2.1: Three version of NearMiss [59]

These are how the NearMiss algorithm works: First, calculate and divide the distance between each point in the larger and smaller class of the dataset. This process makes the undersampling method technique easier. Then, the instances of the more significant type most similar to those of the smaller

class are chosen. The collection of the n class must be saved before then gets deleted. The larger class will return m $\times$ n instances of itself if the smaller type has only m illustrations.

i. NearMiss-1
NearMiss-1 chooses the majority class instance with the shortest mean distance to the next N samples. NearMiss-1 keeps track of majority-class locations with the shortest mean distance from the N nearest minority-class locations. In other words, it will retain all characteristics of the majority class that are similar to those of the minority class [59].

ii. NearMiss-2
In contrast to NearMiss-1, NearMiss-2 keeps the majority class points with the shortest mean distance to the minority class's N, the farthest points. In other words, characteristics of the dominant type as opposed to those of the minority class will be maintained [59].

iii. NearMiss-3
The NearMiss-3 algorithm is a two-stage algorithm that combines the best features of the NearMiss-1 and NearMiss-2 algorithms. It begins by sampling each majority class's N nearest neighbors. The majority-class examples are the furthest away from their N nearest neighbors on average [59].

2.4.2 BERT models

Google AI Language researchers published a paper titled "Bidirectional Encoder Representations from Transformers," or "BERT" [12]. It demonstrated cutting-edge results in various NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (NLI), and others, causing a commotion in the machine learning research community.

BERT's most significant technical advancement is applying the popular attention model Transformer's bidirectional training to language modeling. Previous studies, however, focused on text sequences that moved from left to suitable or combined right-to-left and left-to-right instruction. BERT's most significant technological advance is the application of bidirectional training to language modeling, which is based on the widely used attention model Transformer. Previously, research concentrated on teaching text sequences from left to right or from right to left and left to right simultaneously [40].

Several studies have used deep learning models for fake detection, particularly the

BERT model. Kumar et al., [20] propose a deep learning technique called FakeBERT that is based on BERT (Bidirectional Encoder Representations from Transformers) and is made up of parallel blocks of a single-layer deep convolutional neural network (CNN) with different kernel sizes and filters. This is an excellent way to deal with ambiguity, which is the hardest part of understanding natural language. Based on classification results, their proposed model (FakeBERT) performs better than existing models with 98.90% accuracy.

Numerous studies have shown that the BERT model performs better than other deep learning models for fake news classification tasks. For instance, Wani et al. [55] evaluated the performance of various deep learning models using the Constraint@AAAI 2021 COVID-19 artificial news dataset. The models used by the author are Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), and the Basic Extended Recurrent Transducer (BERT). Regarding accuracy, BERT, a pre-trained transformer model, outperformed other deep learning models by a margin of 3%

Furthermore, Szczepaski et al. [48] propose a method for improving the explainability of the BERT model by analyzing the BERT-based model's attention mechanism and calculating the relevance of the input token. This relevance score is then used to determine the most significant ticket in the input text and explain the model's decision-making process. This model was then tested on the fake news dataset to show that it can be used to identify the most relevant token and improve the BERT model's explainability, allowing the user to understand the model's decision-making process.

IndoBERT

In detecting fake news in the Indonesian language using the BERT model, there is already a pre-trained BERT model called IndoBERT developed by Koto et al. [24]. This pre-trained BERT model was trained using 220 million words aggregated from three primary sources: Indonesian Wikipedia (74 comments), new articles from Indonesian media such as Kompas, tempo [49], and Liputan6 (55 million words total), and the Indonesian Web Corpus [29] (90 million words). Isa et al. [17] use IndoBert on a fake news classification task and classify the article as fake or real. In addition, the authors use a labeled Indonesian news article that originated from multiple sources. Their experiment showed that the BERT model has a great result in detecting and classifying fake news in the Indonesian language and can be used to combat the spreading of fake news in Indonesia. The dataset for their research originated from Mafindo, the same Institution where the dataset for this research also originated. Instead of using all nine dataset categories, they are generalizing the dataset only into two sets: "real" and "fake news." In addition, their experiment was done with a limited dataset of only 4031,

consisting of 3.4465 fake news stories and 766 “real news stories.” Besides that, their source code or their pre-trained model is also not publicly available.

After researching defining fake news and multiple approaches to automatic detection, the BERT model has shown incredible promise, as it performs better than other deep learning methods. In this research, the performance of the BERT model will be compared to the traditional machine learning approach to observe which approach performs better, especially when tested using an Indonesian news article dataset.

2.5 Claim verification

Claim verification is examining a claim or statement to determine the veracity of a given claim, which is essential for various downstream applications [26]. In general, the input for a claim verification task is a claim or statement that needs to be evaluated to verify the information. This claim or statement may cover various topics or subjects, such as scientific findings, historical events, and news events. The supporting evidence or fact is required to verify this claim’s veracity. The output or label is generally “support”, “reject”, or “neutral”, indicating that the evidence is either supporting or rejecting the claim and, in some cases, is neutral towards the given evidence. When there is insufficient evidence to make a conclusive determination, an additional label such as “unknown” or “unverifiable” is used. Based on available evidence or data, this task’s result assesses the claim’s veracity.

There is various research that has been conducted related to claim verification tasks. When retrieving documents, Thorne et al. [52] used the TF-IDF approach, which is also used when extracting evidence sentences. Yoneda et al. [58] create a logistic regression model by employing a variety of heuristic features in its construction. In [10, 15, 33], the Enhanced Sequential Inference Model (ESIM)was utilized. ESIM utilizes the co-attention mechanism and two BiLSTMs to organize hypotheses by the premises. Thorne et al. [52] use decomposable attention [36], which compares and aggregates words in soft-aligned sentences for claim verification. The most recent work is done by Soleimani et al. [43], where they proposed BERT for evidence-based claim verification. The authors experimented using two BERT models: one for retrieving evidence sentences that support or reject claims and another for verifying claims based on those sentences. The researchers used pointwise and pairwise loss functions to train a system called BERT for information retrieval, focusing on the technique of harmful hard mining. The system’s performance was evaluated on a dataset known as FEVER. The results showed that it ranked second with a FEVER score of 69.7. It also set a new record, achieving an 87.1%

Furthermore, some researchers leverage the generative pre-trained transformer (GPT) on claim verification-related tasks. GPT-3 is an autoregressive language model that was trained with 175 billion parameters before being evaluated in “few-shot learning settings” (situations where a new language task can be completed using only a few examples) [7]. Based on previous natural language texts, autoregressive language models can predict the next element in a text, usually a word [22]. Jeong et al. [18] use the GPT-3 model to generate additional evidence for data augmentation purposes. The authors extracted relevant sentences from the FEVER dataset to generate additional evidence sentences with the GPT-3 model. Moreover, a new dataset is created by combining the newly created augmented dataset with the primary dataset. Their experiment shows that the augmented model outperforms the baseline model. Their research shows how GPT-3 can improve a fact-verification system.

Chapter 3

Company Visit & Interview

In this section, we will discuss the company visit and the questions prepared in advance for interviewing the fact-checkers. The purpose is to delve deeper into their workflow, the challenges they face during their work, and explore how NLP could be utilized to improve their efficiency. We will be visiting two institutions located in Jakarta, Indonesia. The first institution is Mafindo [1] (Masyarakat Anti Fitnah Indonesia), which translates to The Indonesian Anti-Hoax Citizen. Mafindo is a non-profit organization that focuses on combating the massive flow of disinformation in Indonesia. The second institution is Kominfo [2] (Kementerian Komunikasi dan Informasi), the Ministry of Communication and Information Technology in Indonesia. Kominfo is a government organization in Indonesia that oversees the communication and information sector.

3.1 Manual fake news detection

The primary goal of this study is to compare human-annotated and NLP-based methods and propose a workflow that combines human and automated methods to increase the fact-checkers work efficiency. This section will be written following the company's visit to this institution to learn its methodology for annotating and classifying articles.

To identify their method, some important aspects need to be explored.

1. Analysing and describing how the human fact-checkers work
2. Analysing which of their tasks could be supported by NLP methods
3. Evaluating existing methods for these tasks

[1] https://www.mafindo.or.id/
[2] https://www.kominfo.go.id/

4. Proposing a workflow for combining human and automated methods

For the company visit, some questions will be asked in the interview session to learn how the institution has done its manually annotated fake news classification. The respondent from each institution will consist of one fact-checkers supervisor and two fact-checkers.

Questions to ask for fact-checkers

1. Are there any well-prepared guidelines for this task that have already been created?
2. How fact-checkers find information to verify a news article or post. Is it sufficient to use an online news article?
3. What is, in your opinion, the most reliable news/post-verification source?
4. Internet-based news article from a reliable source: does it contain any bias?
5. Could you describe how you determine whether an article is biased or how you detect a bias? Does this require prior knowledge, or do you sometimes rely on intuition?
6. How the necessary information is gathered to validate the information. Do you think automating the data scraping process would be advantageous if it is not already?
7. What should be done when there is insufficient information to validate the news/post?
8. Does this institution have experts to validate unknown or unconfirmed information? If available, on what topic that this organization has?

Questions to ask for the fact-checkers supervisor

1. How do you ensure no bias with the fact-checkers? Is there only one person in charge of validating a post or article, or are there multiple people in charge so we can compare their judgments on a topic?
2. Does this institution have a subject-matter expert whose knowledge can be used to validate the information or for unknown information directly?
3. What qualifications are necessary for someone to work as a fake news fact-checkers?
4. If a workflow already exists, do you find it efficient? If not, what improvements can be made?
5. Is there a task that can be automated to improve efficiency and facilitate work?

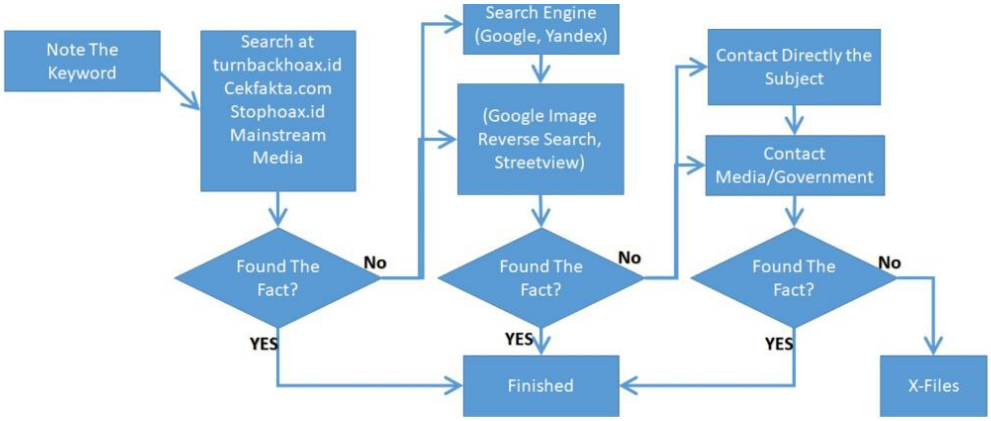

Figure 3.1: Mafindo Workflow

Following the visit, the companies generally operated in the same manner. However, their workflow differs significantly because Kominfo is a government organization with many layers of bureaucracy involved before making a decision. In contrast, Mafindo is a private institution with fewer layers, which will be explained later.

3.2 Mafindo

The Mafindo fact-checker team consists of 11 fact-checkers and one supervisor who validates and gives feedback on the fact-checkers work. They communicate through a telegram group, sending their work and directly getting feedback.

As shown in the Fig. 3.1, Mafindo works in this order:

1. fact-checkers find a topic and then make sure that they haven't covered it yet on their platform
2. The fact-checkers propose a topic to be validated to the supervisor through the telegram group
3. After the proposed topic has been approved, they try to find the counter-article using search engines. They are utilizing Google for searching articles and Yandex for image reverse search.
4. If they find a counter article that can give a clear answer, they write their feedback article and send their investigation article to the group to be reviewed by the supervisor. The online article that they are using as their reference must be certified and acknowledged by the Indonesian Press Council. After getting the review from the supervisor and finishing the final investigation, the article will be published on their platform.

5. If the fact has not been found after a Google search and Yandex image reverse. Then, they will either try to trace and find the source or contact news media or government organizations related to the issue. The investigation is written and published on their platform if the fact is found after directly contacting the source. Otherwise, it will be kept in their repository until an article or press release addresses the issue.

3.3 Kominfo

Kominfo has a dedicated team called AIS that works under the supervision of the Aptika Department. They have two work shifts from 9 am to 4 pm and 4 pm to 10 pm that operate from Monday to Friday. This team consists of 5 people for each work shift to ensure no bias since many work on the same topic. In contrast to Mafindo, they only work with a topic that has importance to the general population, has a national interest, and might pose a security threat if not handled as soon as possible. So they will not cover topics such as celebrity gossip, as this is not a threat to national security.

They are working as explained below:

1. The fact-checkers find a topic that might pose a national security threat and check if it has been investigated before.
2. If the topic has not been investigated before and poses a national security threat. Then, they will ask the approval from their supervisor to investigate this topic
3. The fact-checkers then collaborate to find the counter article using search engines. After retrieving the results, the search results are filtered to include only articles from media companies approved and acknowledged by the Indonesian Press Council.
4. If they find the counter article and have enough evidence to make their judgment, they write an investigation article and submit it to their supervisor. Their supervisor then gives feedback after the article has been finalized. The investigation article is then sent to the General Director of the Aptika Department to be reviewed. This ensures they did not make any errors since they are a government agency, and a small mistake can lead to fatal consequences. And this investigation is then reviewed by the General Director, and after getting approval, this article is published on their platform.
5. If no Indonesian-approved media has already written an article regarding the topic. The fact-checkers then attempt to contact the primary source directly, or

they can look for the press release or any other published publication if the topic is related to a specific institution, company, or public figure. Some organizations, institutions, or public figures respond quickly to rumors because it will affect their credibility. After the necessary evidence has been found, they write the investigation article and send it again to their supervisor. After reviewing the article, the supervisor forwards it to the general director for final approval.

6. If, after doing all of that, they still cannot find the counter article or reliable source, they will hold this topic until they find a reliable article or press release from the direct source that could be used to address the issue.

Automatic classification is utilized to expedite the classification process compared to manual classification. However, this method is not without its drawbacks. When identifying and publishing the results of the classification of false news, the institution must ensure that it is publishing the correct classification; otherwise, it could be prosecuted under Indonesian law. Specifically, Mafindo, the institution from which the dataset was collected. The Indonesian press law does not protect them because they are not a press organization. Unlike any other registered press company, they can revise their published article if they make a mistake. Automatic classification may not be the optimal solution for detecting fake news in real-world study cases, given that no machine learning technique is capable of making 100% accurate predictions. However, this is an excellent comparative tool for the fact-checker's evaluation of the annotated article.

Chapter 4

Method

This research consists of two tasks: fake news classification and claim verification. The fake news classification was conducted with three traditional machine learning classifiers and a pre-trained BERT model for Indonesian called IndoBERT. At the same time, the claim verification was performed with the same pre-trained BERT model for Indonesia called IndoBERT and the Generative Pretrain Transformer (GPT), described below.

4.1 Fake news classification

When doing automatic fake news classification, two approaches were used in this research: using the fine-tuned IndoBert Model, a customized pre-trained BERT model that was trained using the Indonesian language, and then comparing the performance of the BERT model with traditional machine learning using Linear Regressor, Random Forest and Support Vector Machines Classifier.

4.1.1 Traditional machine learning classifiers

TF-IDF is used for data analysis to find the most critical word in each class to get more insight into the dataset. In addition, three machine learning classifiers are being utilized for the fake news classification task.

Text representation using TF-IDF

In the experiment, TF-IDF (term frequency-inverse document frequency) is used to examine the textual representation of each classification. We are using TF-IDF to find the most critical word in each class to gain insight into the characteristics of each class

for both tasks. For the fake news classification, we are examining the “content” column based on the classification in the “classification” column. While for the claim verification label, the “fact” column is examined for each class where the label is located in the “claim” column.

We combine TF-IDF and Logistic Regression to identify the top 10 most significant words for each class in both fake news classification and claim verification tasks. The model calculates the importance score using this approach.

1. The TfidfVectorizer transforms raw text into a numerical matrix that represents the significance of each word using TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF measures the importance of a word in a document relative to the entire collection. The logistic regression model then works in this numerical matrix representation of TF-IDF.
2. During the training process, the Logistic Regression model assigns weights to each word based on its interaction with the target variable. These weights indicate how much each word influences the probability of a particular class. Positive weights increase the log odds (and hence the probability) of the response, while negative weights decrease it. These weights are then sorted in descending order, revealing the top 10 words with the highest absolute weights for each class. These influential words play a significant role in predicting each class.

Classifier

There are several traditional machine learning classifiers that we are experimenting with in the dataset. In our experiment, this is how we are doing the fake news classification with the Indonesian language dataset using traditional machine learning.

1. **Dataset Extraction**
   We first extract the dataset using the API provided by Mafindo and then store the extracted dataset in the CSV format. The dataset is then loaded into the panda’s data frame.
2. **Data Preprocessing**
   We then preprocess the necessary data as specified in section 5.3.1.
3. **Feature Extraction**
   In this step, we select which data features should be used and which are irrelevant to the model. Insignificant data may reduce classifier efficiency and model accuracy. As a result, removing unnecessary data from the dataset is necessary. For selecting the features that are significant with the dataset, we are combining Term

Frequency (TF) and Term Frequency Inverse Document Frequency (TF-IDF) to find the most important word in each class as mentioned in section 5.2.1.

4. **Model Selection**
   This step involves selecting the optimal machine-learning model for the given task. The selection depends on the dataset's characteristics and the problem's type, either classification, regression, clustering, etc. Upon doing an extensive research study, as mentioned in section 2.4, we found that SVM, Random Forest, and Logistic Regression are the most common and effective traditional machine learning models for the fake news classification task, and therefore we are utilizing these models.

5. **Model Training**
   We then trained the selected model with the preprocessed data. We use the "content" column as input to the classifiers, normalized for spelling errors and slang as described in section 5.3.1 on the fifth step of the data preprocessing. In the training phase, the model adjusts its internal parameters to find the underlying relationship and pattern between the training dataset. Generally, this step involves a process that tries to find the optimal value for the model's internal parameter that minimizes a specified loss or error metric; in this case, we are using accuracy, precision, recall, and f1-score as the error metrics.

6. **Model Evaluation**
   After training the model using the training dataset, we evaluate the model's performance using a separate test set, a set of datasets to which the model had not been exposed during the training process. Evaluation metrics such as accuracy, precision, recall, and F1-score are calculated to determine how well the model generalizes with the test data. This step aims to improve the understanding of the model performance and to identify any underlying problems, such as overfitting or underfitting.

As explained before, three traditional machine learning classifiers are being utilized in the experiment, such as logistic regression, random forest, and support vector machines. Some hyperparameters were added, and GridSearchCV from scikit-learn is being utilized to analyze all possible combinations of the hyperparameters and find the best accuracy score during the training process [1].

1. **Logistic Regression Classifier**
   We are using the Stochastic Average Gradient Descent (SAGA) solver; this algorithm supports large datasets and the non-smooth penalty term, such as the

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

L1 penalty option. Since we are working with a large dataset, the "max_iter" parameter is set to 10000 to ensure the model can converge to an optimal solution. In addition, we are optimizing two parameters, regularization parameter $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ and the $penalty \in \{l1, l2, elasticnet\}$. The C parameter controls the regularization strength, and the regularization technique helps to prevent overfitting by preventing the model from becoming too complex. A smaller C value indicates a smaller regularisation. While the penalty parameter is for choosing the type of regularization that will be implemented, three types of regularization are tested in the experiment (l1, l2, and elasticnet); elasticnet itself is the combination of l1 and l2 regularization.

2. **Random Forest**
   We are optimizing two hyperparameters, the number of trees n and the maximum tree depth m using the following grids: $n \in \{50, 100, 150, 200$ and $m \in \{10, 20, 30, 40, None\}$. The n parameter represents the number of trees, while the m parameter represents the depth of the tree.

3. **Support Vector Machines (SVM)**
   We use the linear kernel to optimize the C parameter using the grid $C \in \{0.1, 1, 10, 100, 1000\}$. In SVM, the C parameter is a regularisation parameter determining the penalty amount applied to a misclassified point. Lower C values mean that the penalty will be low and the margin-based boundary will be higher, while high C values mean otherwise. The C parameter affects the trade-off between model complexity and classification accuracy [5].

Imbalanced learning

The dataset is imbalanced, as shown in table 5.3.1. There are two imbalanced learning methods that we are experimenting with, such as Over-sampling and Down-sampling, to tackle the data imbalance issue that is explained below.

1. **Oversampling with SMOTE** Over-sampling with SMOTE will increase the size of the minority class to the size of the largest class. The largest class in the data set is the Misleading Content" class, with 2629 instances. SMOTE will balance the class distribution by creating synthetic samples for the minority class. This SMOTE algorithm will generate samples until the number of samples in each class matches the size of the largest class.

2. **Down Sampling**
   Down-sampling will reduce the size of all the majority classes to that of the smallest minority class. This method will compare the size of each class to the num-

ber of instances with the most minor instances, resulting in a balanced dataset. As shown in table 5.3.1, the “satire” class has the lower instances with 162 instances/articles. The down-sampling method will reduce all other classes accordingly to the size of the “satire” class; 162 instances per article will be selected randomly for each class. One crucial thing about downsampling is that this process may lead to the loss of potentially important information since we reduced the size of each class as some instances or articles were removed.

(a) **RandomUnderSampler** We are performing RandomUnderSampler, a method for under-sampling the majority classes to match the size of the minority class. The main difference between this method and NearMiss is that this method removes instances from the majority class. In contrast, NearMiss selects instances closest to the minority class based on a predefined proximity measure.

(b) **NearMiss**
We are using all three NearMiss variants, such as NearMiss-1, NearMiss-2, and NearMiss-3.

4.1.2 IndoBERT model

Therefore, fine-tuning is needed since the IndoBERT model does not yet train for the fake news classification task. Then we fine-tuned the pre-trained IndoBERT model on our training set. We are modifying this implementation of the fine-tuning BERT model for fake news classification from Skillcate AI to fit with our dataset. [2] We trained the model for 30 epochs with the Adam-W optimizer.

The fine-tuning of IndoBERT for fake news classification is done in this manner:

1. Data Loading: The content and the classification label are extracted from the dataset. The content originated in the content” column, and the label was taken from the classification” column.
2. Tokenization: We used the default IndoBERT tokenizer for tokenizing the content. Tokenization is breaking down the text into tokens and converting those tokens into a numerical representation that the model can understand.
3. Encoded Claim and Evidence Tensors: Since only one input exists, the model uses a cross-encoder to encode the “content” column. A unique token like [CLS] at the beginning of the sentence and [SEP] at the end was added. The [CLS]

[2] https://medium.com/@skillcate/detecting-fake-news-with-a-bert-model-9c666e3cdd9b

token is specifically designed to hold sentence-level representation and is used in the final classification task.

4. Fine-Tuning the Model: A sequence classification layer is added on top of the pre-trained IndoBERT model. This classifier operates on the [CLS] token's embedding from the final hidden state of the IndoBERT model. The entire model, including all layers of IndoBERT along with the sequence classification layer, is trained during the fine-tuning process.
5. Model Logits for Labels: The model returns the logits for each classification label for each batch during the training phase. These logits represent the predicted classification class. The higher the class logits towards a class, the more likely the input is predicted towards that class.
6. Predicted Label: In this step, the label is determined based on the predicted class of the given input with the highest logits score. This predicted label was compared to the existing label in the label list index. A softmax function is applied to the output of the model to convert these logits into probabilities for each class, providing the final class prediction.
7. Model Performance Using Classification Metrics: The fine-tuned model was evaluated using the unseen data from the test set. The performance of the fine-tuned IndoBERT model was then evaluated using multiple metrics such as precision, recall, f1-score, and accuracy.

4.2 Claim verification

Claim verification is the process of determining the veracity of a claim by examining the supporting evidence. We performed the claim verification task using the IndoBERT and Generative Pre-train Transformer models.

4.2.1 IndoBERT

When experimenting using the IndoBERT model, the IndoBERT model originally has not been fine-tuned for the claim verification task. The dataset is then annotated with the assumption that for every true or real news article or post, the fact must support the evidence; therefore, all the true or real news will be annotated as "support," and all the fake news will be annotated as "reject," as the evidence in the fake news will be against and not supporting the claim. As mentioned in section 5.1.1, there are nine classes in the dataset, and two belong to real news, such as the "true" and "clarification" labels,

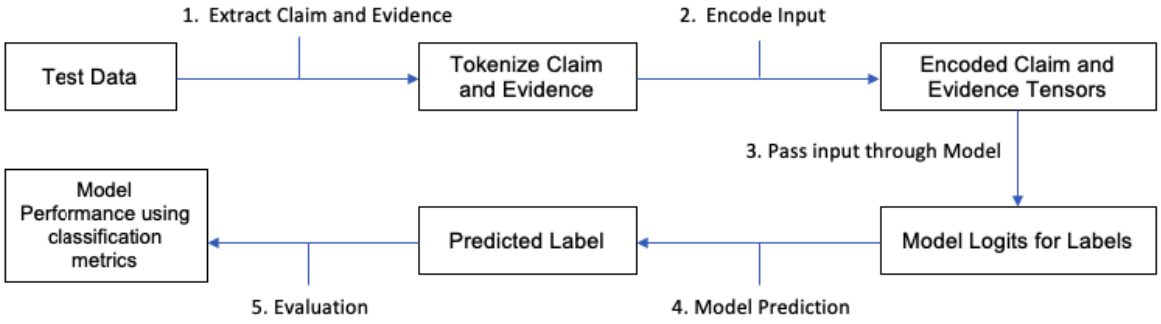

Figure 4.1: Claim Verification Pipeline

while the rest are classified as fake news. The clarification class consists of articles usually written by the person or institution that address the claim or rumor and clarify the issue.

We followed the cross-encoder approach of Soleimani et al. [43] as shown in figure 4.1:

First, tokenize the claim and evidence with the IndoBERT tokenizer, then concatenate them using the [SEP] token and encode them as a dense embedding vector.

1. Extract claim and evidence: the claim and evidence extracted from the dataset, with the claim extracted from the “title” column and the evidence extracted from the “facts” column.
2. Tokenize the claim and evidence: The IndoBERT tokenizer is used to tokenize the claim and evidence separately. This tokenization process splits each text into the corresponding tokens, resulting in separate token sequences for the claim and evidence.
3. Encode claim and evidence tensors: The tokenized claim and evidence are encoded separately using the IndoBERT model. This encoding step generates dense embedding vectors for each token in the claim and evidence token sequences.
4. Model logits for the labels: The encoded claim and evidence tensors are combined by inserting the [SEP] token between them, following the common LM style approach for Natural Language Inference tasks. The combined encoding is then passed through the IndoBERT model, which produces label logits representing the model’s predictions for each class label.
5. Predicted label: A softmax activation function is applied to the logits to obtain a probability distribution over the labels. The label with the highest probability is chosen as the predicted label.
6. Model performance using classification metrics: The model’s performance can

be evaluated using classification metrics such as accuracy, precision, recall, and F1-score. These metrics assess how well the model performs on the claim verification task using unseen data.

4.2.2 Generative pre-trained transformer (GPT) 3

GPT-3 is an autoregressive language model trained with 175 billion parameters. It was highly successful in few-shot learning settings (where a new language task can be completed using only a few examples) [7]. We experimented with GPT-3.5 (text-davinci-003) for claim verification. [3]

We experimented with the GPT-3.5 for task claim verification that works in this manner.

1. Data Loading: we first loaded the dataset from the test set collection, which consists of 2232 test items, into the panda's data frame. This test set consists of a list of the claim and the evidence.
2. Prompt Formulation: the prompt contains four examples of support and rejects (claim and evidence) and one test item (claim and evidence) to be classified. [4] The goal of this example is to provide more context for the GPT model to provide a better understanding of the task or problem.
3. Generating the response: the model then generates a response based on the given claim and evidence pair.
4. Claim Verification Extraction: we parse the generated response to find the predicted label: *didukung* (support) or *ditolak* (reject). If one of these labels is found, the classification is extracted; otherwise, the default value "unknown" is assigned.
5. Translation of the classification: since the prompt is in the Indonesian language and the prompt also asked to output the classification label in Indonesian. This classification result was then translated into English for enhanced comprehension of the classification's outcome. This classification is translated in English as *didukung* as (support) and *ditolak* as (reject).
6. Model Evaluation: The model's performance was then evaluated using a classification report, which included metrics such as precision, recall, and f-1 score. This metric provides a comprehensive overview of how well the model predicts each class while considering false positives and negatives.

[3] We approached GPT-3.5 through the OpenAI API.
[4] The full prompt is shown in the appendix A.2.

Chapter 5

Data

5.1 Data overview & analysis

5.1.1 Data description

The dataset in this research originated from Mafindo and consisted of 10,756 fact-checked Indonesian-language articles. The following is a list of all the dataset's features, along with the data types for each feature and a brief description given below:

- **Authors**: the author name of the investigation article's author. Some authors decide to remain anonymous to hide their identities since some topics are so sensitive and could lead the investigator into serious problems.
- **Status**
- **Classification**: represent the classification of the fake news based on the nine classifications mentioned below.
- **Title**: represent the article title or the fake news claim.
- **content**: contain the content or the fake news article.
- **Facts**: contain a list of the evidence based on the given title or claim and manually searched and retrieved by the human fact-checker.
- **Claim**: represent the classification of the given claim (title column) to the evidence (facts column), either the given claim supporting or rejecting the evidence.
- **References**: represent the list of references where the fact-check found the evidence.

- **source_issue**: explain where the claim originated; the claim usually originated from social media such as Facebook, Instagram, or Twitter, as well as text messaging applications such as WhatsApp, Telegram, etc.
- **source_link**: represent the URL of the fake news. If it comes from a site for fake news that originated from a messaging app, the name of the messaging app is used.
- **picture1**: since some of the content is in an image form, it is followed by a text caption that emphasizes the picture context. This column represents that image source URL.
- **picture2**: additional image URL if the post/article contains more than one image.
- **tanggal/date**: represent the posting date of the post/article.
- **tags**: containing the tag they use to categorize the article. The tag can be the article platform source, where they posted the reviewed article since Mafindo oversees multiple fact-checking apps and the category of the article (e.g., politics, healthcare, etc.).
- **conclusion**: represent the investigation article written by the human fact-checking; in this part, they will interpret their finding and make a conclusion based on the evidence they found.

The dataset is annotated for various purposes, and this study focuses on fake news classification and claim verification labels. For the classification label, the dataset also has nine distinct classifications, including misleading content, fabricated content, satire, manipulated content, impostor content, false context, false connection, clarification, satire, and accuracy. In addition, the 1829 articles do not have a classification.

Classification	Total
Misleading Content	2629
False Context	2234
-	1829
Fabricated Content	1152
Manipulated Content	1152
Impostor Content	503
TRUE	420
Clarification	396
False Connection	279
Satire	162

Table 5.1: Total count of each classification category in the Mafindo for fake news classification task dataset before data preprocessing

The dataset consists of 15 columns, including id, authors, status, classification, title, content, fact, references, source_issue, source_link, picture1, picture2, tanggal, tags, and conclusion. Tanggal is the Indonesian translation for a date. The "-" class represents an unclassified or unlabelled dataset. There are 1829 samples or articles that have not yet been labeled.

There are nine classifications in the dataset.

1. **Misleading Content**
   In this context, "misleading content" refers to using information to misrepresent an issue or a person. As previously stated, this artwork was used to attack and frame a person.

2. **Fabricated Content**
   The most dangerous type of content is considered to be content that has been fabricated. This type of content created with 100%

3. **Manipulated Content**
   This type typically comprises edited content from a reputable and large news media company. Manipulated content is created by repurposing an existing news article designed to mislead the public's perception of a particular topic.

4. **Impostor Content**
   A piece of content could be classified as an impostor if it pretends to be someone else and usually mimics a high-influence figure. Impostor content targets an individual and an institution by leveraging the institution's fame. This type of misinformation is typically spread in a video or an article, where the cut takes only a specific part of the video or article and misleads the viewer about the context since it only takes part of the content.

5. **False Context**
   False context can occur when an event, depicted through a statement, photograph, or video, is inaccurately narrated or placed in an incorrect setting. Often, this involves presenting an event as if it's happening in a particular location when it's not. This type of misinformation is usually spread via social media, accompanied by captions that misrepresent the actual context.

6. **False Connection**
   The most common characteristic of this type of content is a mismatch between the title and description. This type of content is also referred to as clickbait. Where the actual content sometimes does not even discuss what the title claims. This type of content is typically created to increase the content's financial value or

engagement.

7. **Clarification**
   Clarification is content created to clarify a specific topic or misinformation spreading to the public. A government official, an institution, or a credible news media outlet usually creates this content. This type of content is created to prevent the spread of fake news and raise awareness about the problem.

8. **True**
   This category contains the actual content. All content is created with the correct context between the article's title and its entire body of text. This content is intended to educate, inform, or raise awareness about a particular topic. The 'TRUE' classification in this data point refers to the fact-checked article that turned out to be real news.

9. **Satire**
   This type of content has no evil intentions but only to mislead. Satire is a type of content created to satirize a particular party. The content has elements of parody, irony, or even sarcasm. In general, satire was created as a criticism of an individual or group for addressing an issue that is currently happening. Satire is not harmful content, but some people take the information too seriously and take it as the truth.

Classification	Total
Reject	8111
Support	816

Table 5.2: Total count of each classification category in the Mafindo for claim verification dataset

In addition, for claim verification, we transposed the dataset as follows: the title is the claim; the facts are the evidence, and we added one of the labels 'support' and 'reject' to each pair of a claim and evidence in the data. All the faithful and clarification news categories are labeled as 'support', and all the fake news categories are labeled as 'reject' The resulting dataset is imbalanced: 8111 items have 'reject' and 816 'support' as shown in table 5.1.1.

5.2 Data analysis

Two labels used in the experiment from the same dataset are labeled for different purposes, such as Fake News Classification and claim verification. The dataset is then

explored and visualized to get more insight and a better understanding of the characteristics of the data. The most common and essential word is data or text distribution since the dataset is a text dataset, and the word cloud gives a better visual representation of the data.

5.2.1 Fake news dataset

To begin with, we are identifying the top 10 most common words based on the count of occurrences in the "content" column, which consists of the fake news article. From figure 5.1, we can see that Indonesia is the most common word, followed by "account" and "Facebook". This is an indication that the fake news mainly originate from Facebook accounts and are mainly spreading on Facebook. Another exciting finding is another keyword sentence, "video," which indicates that much of this fake news are in the form of a video where the curator of the fake news edited the video to mislead the viewer. In addition, "COVID" is also one of the primary topics being discussed. At the beginning of the COVID pandemic, many people talked about it on social media and shared fake or misleading information regarding COVID. We are also using word clouds

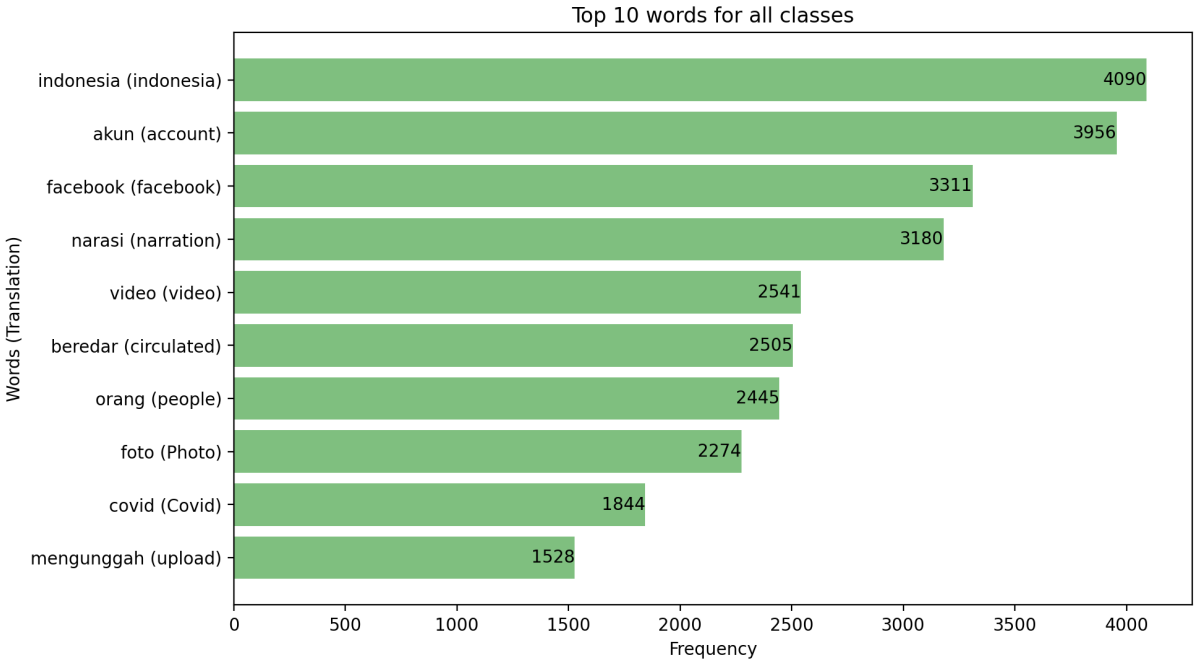

Figure 5.1: Top 10 most occurring words in all classes

to give a better visual representation of the fake news dataset. As shown in figure 5.2, the "Indonesia" word appears bigger than the rest since this is the most common word in the dataset. Still, the word does not have a significant size difference compared to "akun/account" and "Facebook" since there is not a significant difference in the num-

ber of occurrences of these three sentences. The size of the word in the word cloud indicates how many times that word appears in the dataset, so the larger the text, the more frequently that word appears in the dataset, and vice versa.

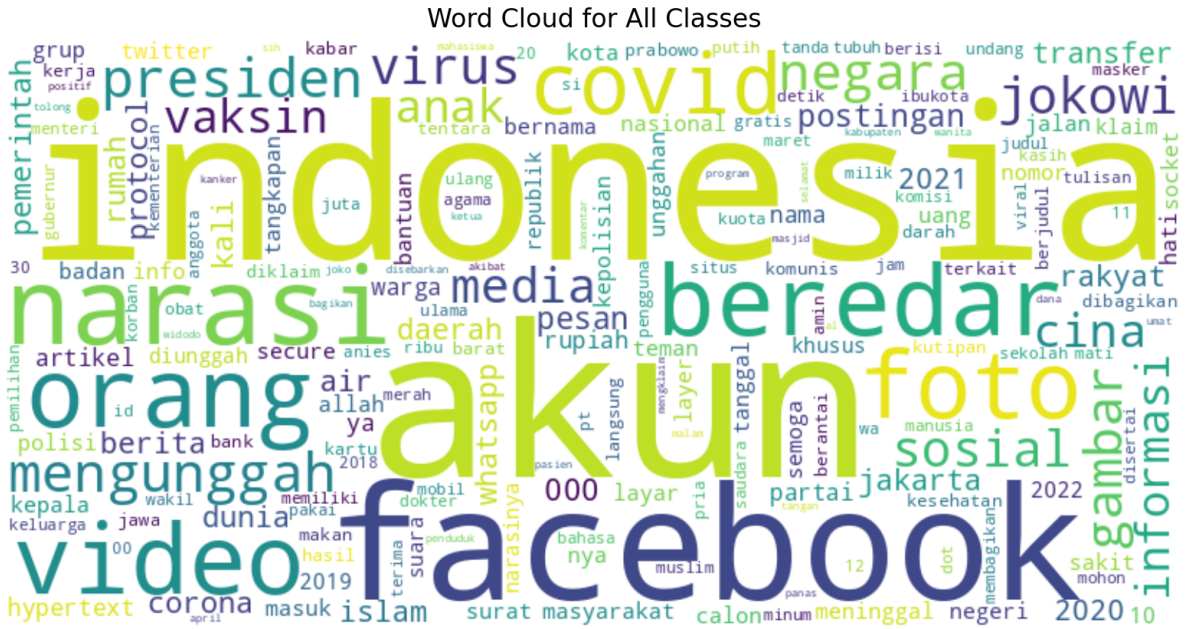

Figure 5.2: Word Cloud for All Classes with the Fake News Dataset

Furthermore, we also explore the text distribution of the entire dataset across all classes to give a better insight into how long the text is. The metrics used here are the character size instead of the word size since some words are shorter than others, so there will be some cases where the article is shorter if counted based on the number of words instead of the character. The number of characters gives a better understanding and is more precise and consistent regarding the actual size of the article. As shown in figure 5.3, most articles have less than 1000 characters since fake news usually come in a short form since it is mainly targeted at people who do not have a habit of reading. For this type of audience, short and controversial articles will gain their attention more efficiently, and they usually will not try to fact-check and read a long, scientifically proven article to validate the information they consume and take that fake news article bluntly.

Moreover, we delve into the details of some classes from the fake news label to gain insight into the dataset's characteristics in each class. The insights that will be gathered from some of the students will give an overview of the characteristics of the rest of the class. In addition to other metrics previously used when examining the overview of the entire dataset, another metric is used to understand better each class's characteristics, such as the most important word, using TF-IDF (Term Frequency and Inverse Document Frequency). The TF-IDF method assigns a score to a word by multiplying

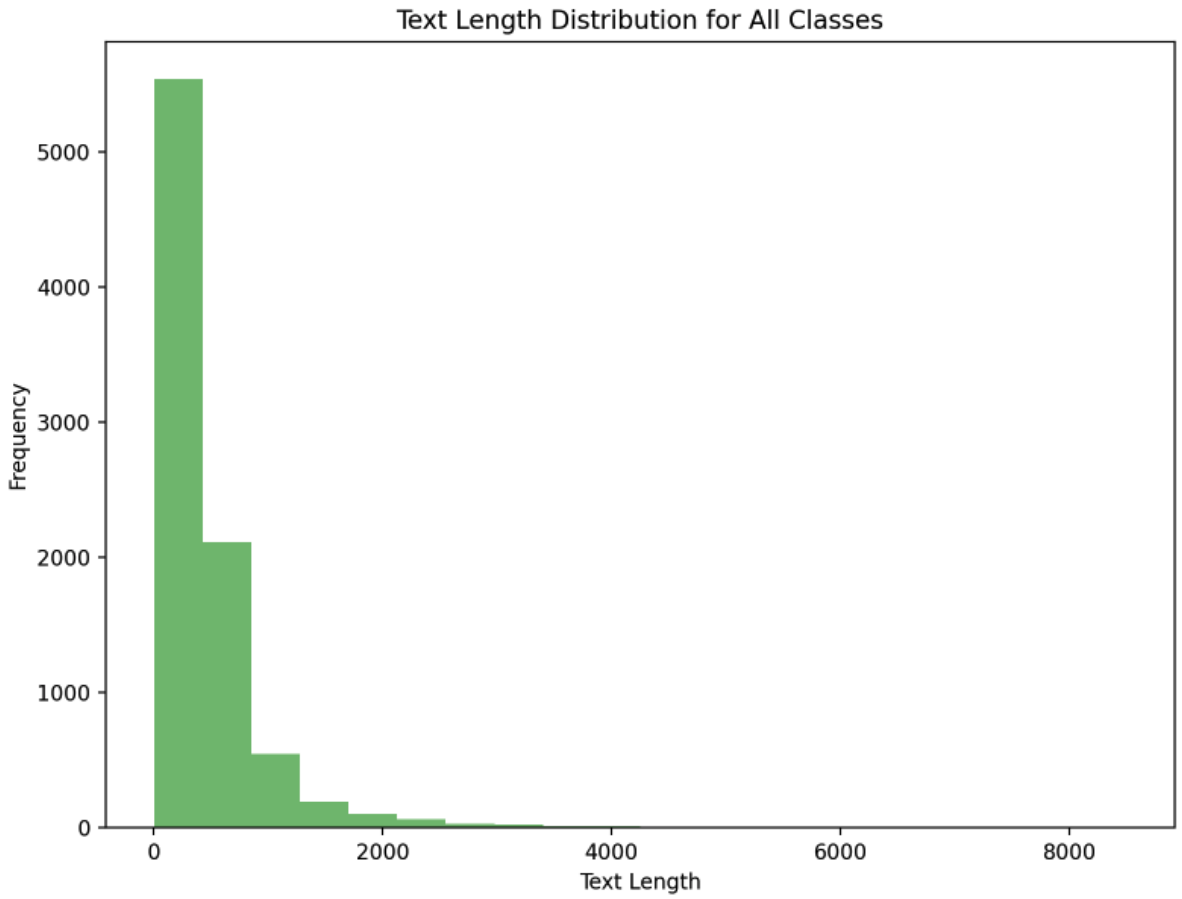

Figure 5.3: Text Length Distribution for All Classes with the Fake News Dataset

its TF and IDF scores. The more scores a word has, the greater its significance. The TF-IDF score was then sorted for all words in each class, and the ten words with the highest TD-IDF scores were shown. In text mining and information retrieval tasks, TF-IDF is one of the most effective measures for evaluating the relevance of words and extracting meaningful insights from text datasets.

1. **Misleading Content**
   In this context, “misleading content” refers to using information to misrepresent an issue or a person. As previously stated, this content was used to attack and frame a person. This is supported by the most common word in this category, ”isu,” which translates to ”issue” or ”rumor” and is a common word used to frame someone. This sentence was used to spread an unverified rumor about someone or something and frame the story being told as true. This type of content is created to influence public opinion in favor of the author’s point of view. Misleading content is created using real information, such as an image, an official statement, or statistical data. The information is then edited to be irrelevant to the actual context. Table 5.3 displays the top ten most important sentences.

Word		Score
English	**Indonesian**	
issue	isu	3.55
clarification	klarifikasi	3.51
hoax	hoaks	2.55
police	polisi	1.67
corona	corona	1.50
appear	muncul	1.49
2019	2019	1.36
voice	suara	1.27
council	dewan	1.22
country	negara	1.19

Table 5.3: Top 10 Most Important Words for the Misleading Content Class

When looking into the most common word as shown in figure 5.4, we found that the top 4 most important words in the misleading content class are similar to the overall data point, with a slight difference where “narration” ranked fourth, followed by “Facebook”. In contrast, in the entire data point, the word “Facebook” ranked higher and occurred more frequently. The word such as “COVID”, “virus”, and “vaccine” also occurred more frequently in this class, indicating that most of the article is related to the COVID pandemic and the vaccine.

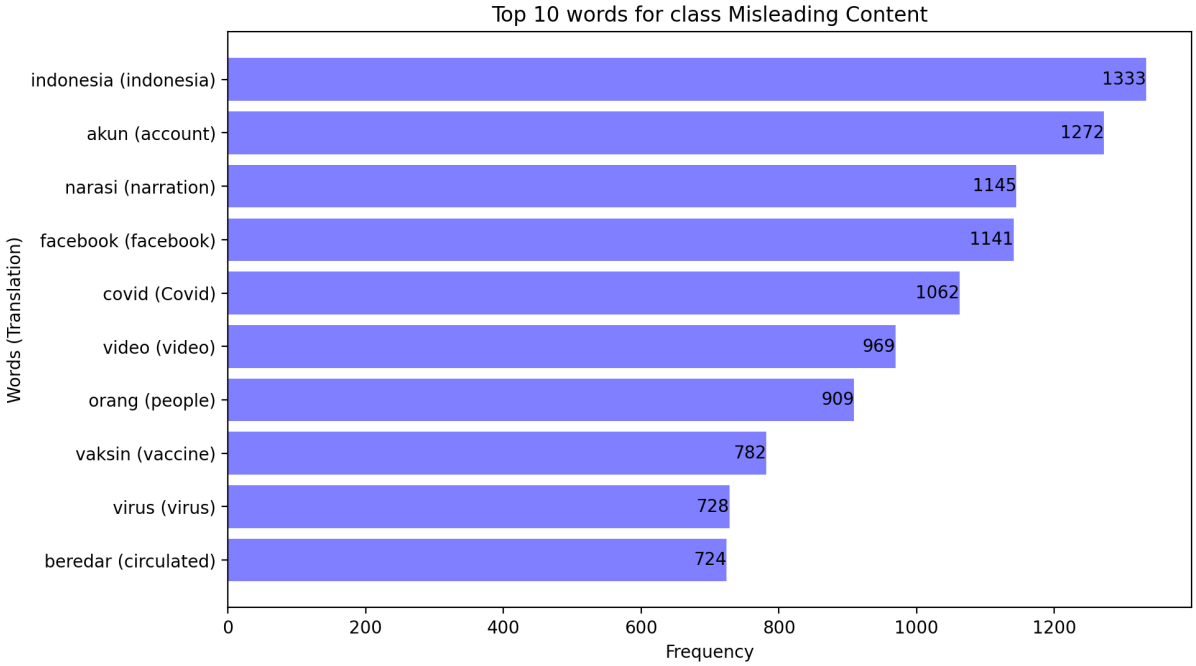

Figure 5.4: Top 10 most occurring words in Misleading Content Class

In terms of the data distribution, as shown in figure 5.5, it is evident that most of the data point is lower than 500 characters, indicating that most of the mislead-

ing content class is short-form text content. In general, most of the data point is shorter than 2000 characters. There is some noticeable outlier data point with a length of around 5000 characters.

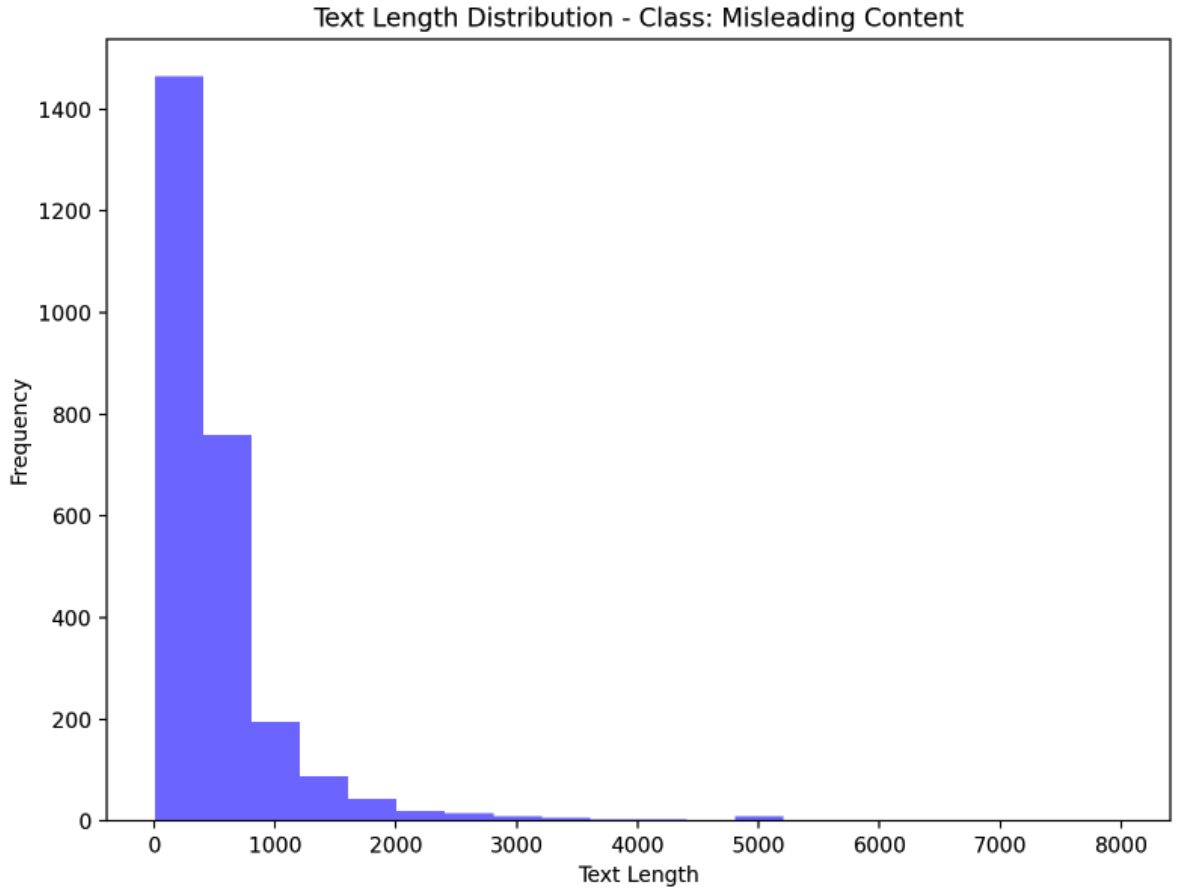

Figure 5.5: Text Length Distribution for Misleading Content Class

2. **Fabricated Content**
   Fabricated content, particularly in the context of fake news, refers to information that is entirely fabricated and intentionally deceptive. This isn't just about biased or misleading information; it's about content that is completely false, has no basis in reality, and can range from manipulated images and videos to wholly made-up articles, statements, or events. The danger of such fabricated content is its potential to mimic legitimate news, making it difficult for people to distinguish between what's real and what's not. This not only leads to misinformed public opinions and skewed perspectives but can also, in some cases, result in real-world harm or unrest.

   This type of content typically takes the form of false job advertisements, lottery winners, and so on. This is evident from the top word mentioned in this category in table 5.4. For instance, ”hadiah” refers to a gift or ”lowong” which translates

Word		Score
English	**Indonesian**	
message	pesan	2.43
gift	hadiah	2.35
whatsapp	WhatsApp	2.25
info	info	2.19
vacancy	lowong	2.07
pt	pt	1.88
information	informasi	1.81
time	jam	1.79
dot	dot	1.71
2022	2022	1.70

Table 5.4: Top 10 Most Important Words for the Fabricated Content Class

to ”vacancy”. This type of misinformation is used to defraud people by posing as a job offer or informing them that they have won the lottery. The most common tactic is to ask the victim to transfer money, which they usually refer to as a lottery tax payment or transportation cost in the case of a job offer. This type of misinformation is usually intended to defraud people.

When examining the top 10 most frequent words based on occurrences in the class of fabricated content as shown in figure 5.6, some unique words do not appear in the list of the top 10 most common words in the entire dataset, including transfer, rupiah, and information. This is a common word in fraudulent job advertisements and fake lottery winner announcements. This confirms the finding with the top most important word result in table 5.4 that this type of false information is intended to defraud individuals for financial gain.

The data distribution is quite similar to the full dataset as shown in figure 5.7, where most of the article is shorter than 500 characters. The datasets in this class range up to around 4000 datasets.

3. **Manipulated Content**
   This type typically comprises edited content from a reputable and large news media company. Manipulated content is created by repurposing an existing news article designed to mislead the public's perception of a particular topic.

   The most commonly manipulated news articles in Indonesia are about Papua and Israel, as shown in table 5.5. Indonesians have potent feelings about these two topics. Indonesians are very concerned about the human rights violations in Israel, especially since Indonesia is a Muslim-majority country. The conflict between Israel and Palestine is always prevalent in Indonesia. There is some

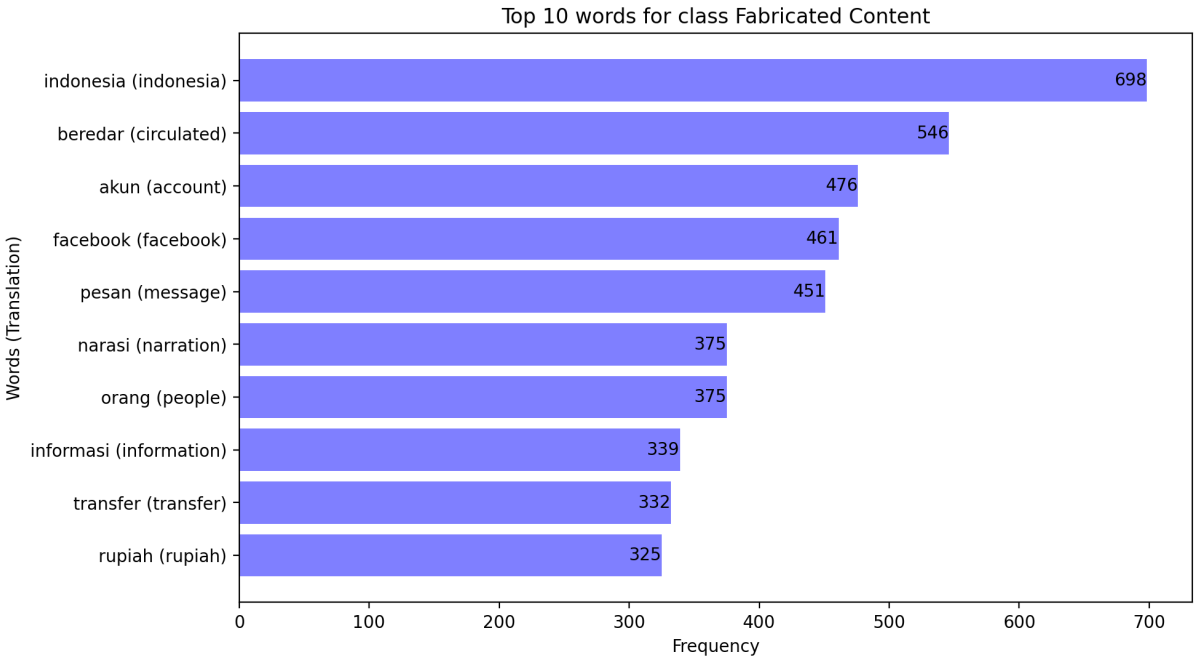

Figure 5.6: Top 10 most occurring words in Fabricated Content Class

Word		Score
English	**Indonesian**	
papua	papua	1.52
title	judul	1.42
ikel	ikel	1.34
posting	postingan	1.33
israel	israel	1.27
amen	amin	1.22
rohingya	rohingya	1.18
animal	binatang	1.18
children	anak	1.14
infrastructure	infrastruktur	1.13

Table 5.5: Top 10 Most Important Words for the Manipulated Content Class

manipulated content that was created to increase sympathy for Palestinians. The same thing is happening with the Papua region, which has been requesting independence from the Indonesian government. Most Indonesians believe Papua will always be a part of Indonesia, and much-manipulated content has been created to gain more support for the Papua rebel group.

The most common word that could be seen in figure 5.8 shows an interesting finding regarding the characteristics of this class. The name of Indonesia's current president (2019–2024), "Jokowi," appears in the top 10 most. This finding indicates substantial fake news related to the Indonesian president. Many people

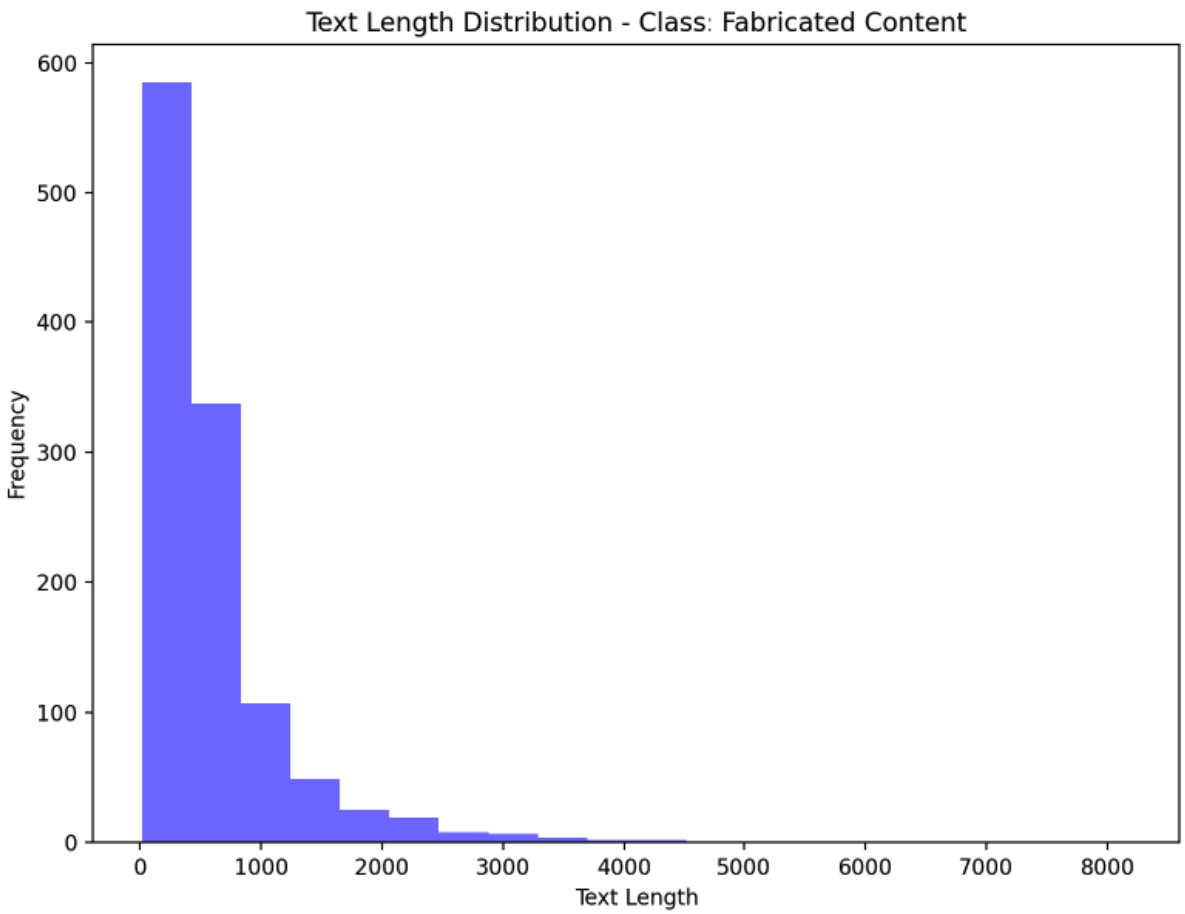

Figure 5.7: Text Length Distribution for Fabricated Content Class

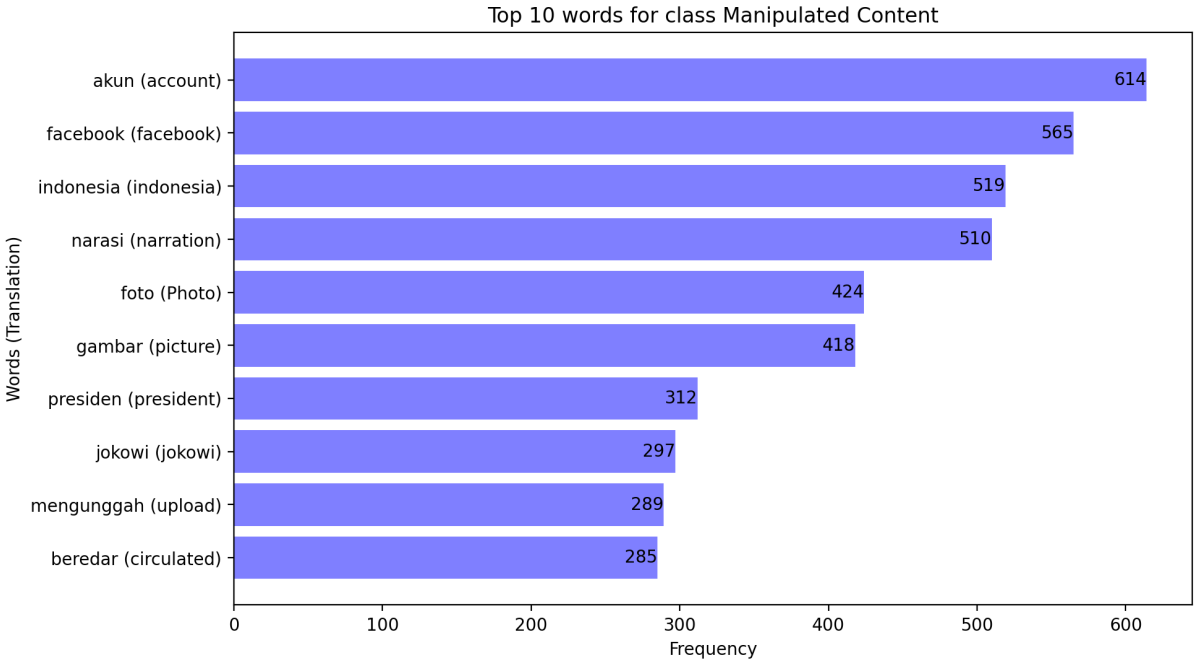

Figure 5.8: Top 10 most occurring words in Manipulated Content Class

try to spread fake news about the president and spread hate towards him.

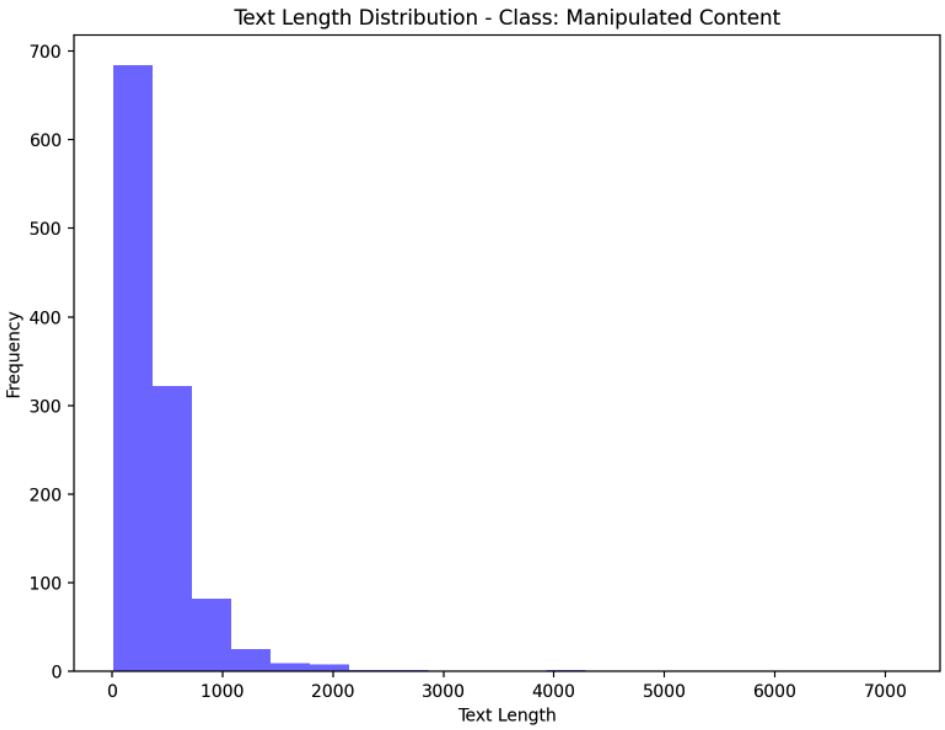

Figure 5.9: Text Length Distribution for Manipulated Content Class

The text length distribution shows that most of the data point is shorter than 500 characters, like the rest of the other classes and the full data point. Unlike other classes, the manipulated content class ranges only up to 2000 characters, and this class has no noticeable outliers.

5.2.2 Claim verification dataset

The same analysis is used when exploring the claim verification label dataset. The class analysis started by looking for the top 10 most occurring words for the combined support" and "reject" classes. As shown in figure 5.10, the word "Indonesia is still on top as previously with the fake news classification label, followed by "video" and "photo". This last sentence is an indication that the data point consists of these two types of content: "video" and photo. There are some keywords that indicate manual evidence retrieval, such as "search", "based on"," and result, which indicate that most of the evidence was gathered by looking for related articles using a search engine.

The word cloud better represents the dataset, as shown in figure 5.11. The word "Indonesia" appears as the biggest as it is the most occurring word in the dataset, followed by video and foto/photos.

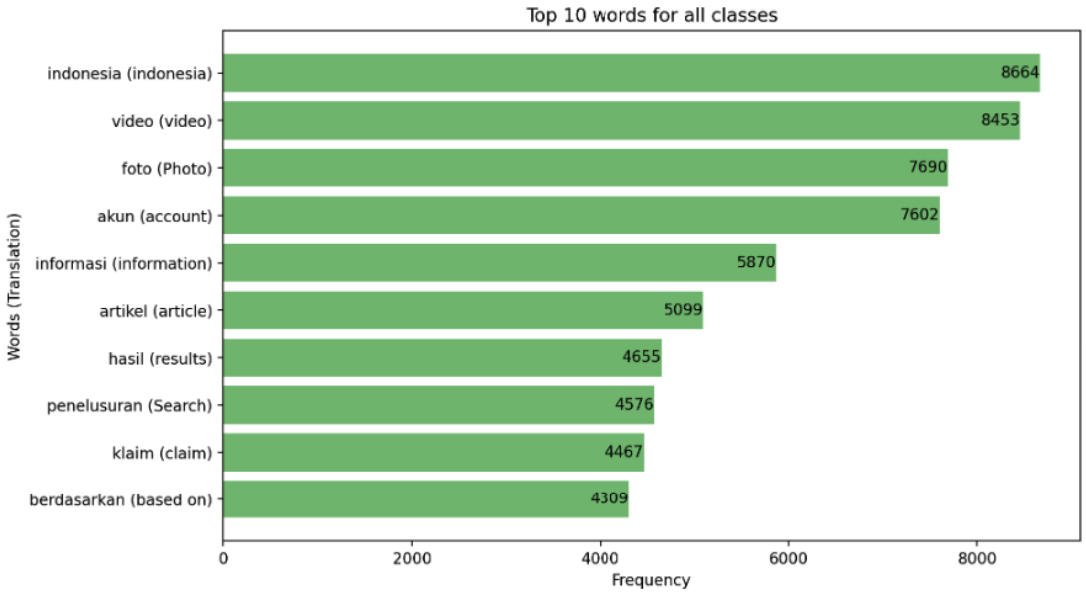

Figure 5.10: Top 10 most occurring words in all classes

Figure 5.11: Word Cloud for All Classes with Claim Verification Label

The words “Covid” and “2020 are also noticeable in the word cloud; this refers to the COVID pandemic that happened in 2020. This finding indicates substantial material related to the previous COVID pandemic.

The text length distribution shown in figure 5.12 gives an overview of how the dataset length varies across all the datasets. Most of the dataset is lower, around 300 to 1000 characters. There is a notable outlier where the text has a length of around 8000 characters. Furthermore, we look deeper into each class to see how the characteristics differ for both the “support” and “reject” classes.

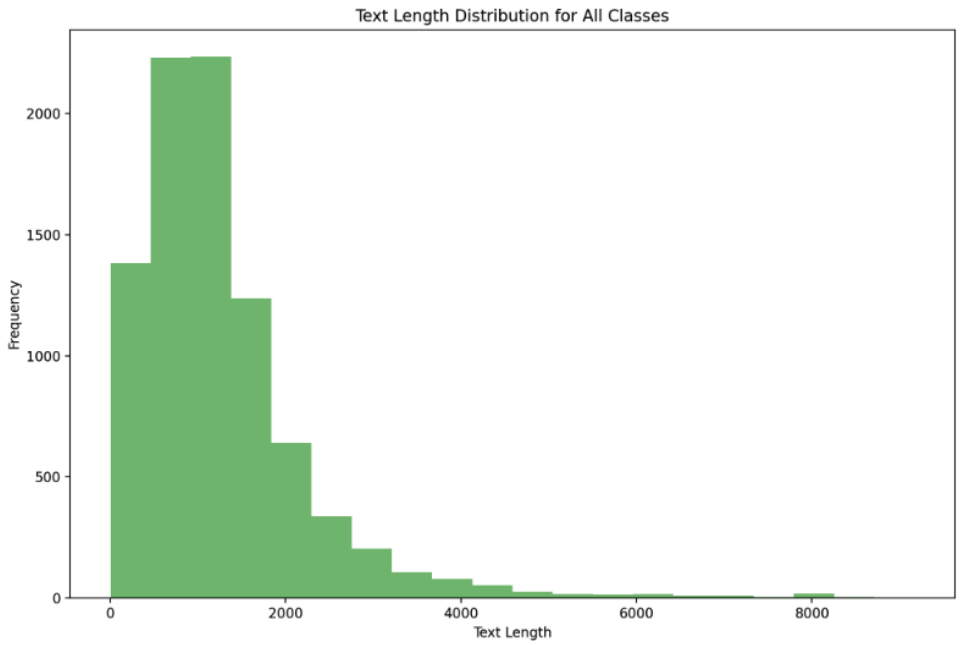

Figure 5.12: Text Length Distribution for All Classes with Claim Verification Label

1. **Support**

   The support class represents the class where the given evidence supports the claim. The evidence refers to the fact column, while the claim is the title column in the data point. The support class consists of the “true” and “clarification” classes from the fake news classification label. When looking for the most important word using TF-IDF, as shown in table 5.6, the term ”social” appears on top, indicating that social issues emerge as the most significant topic or characteristic in the support class. Other sentences, such as “sanitation,” suggest that many supported claims revolve around the topic of sanitation or hygiene. In addition, the other sentences, such as “video facilitate”, “help”, etc. indicate the wide range of topics that are being discussed in this class. However, this term should be interpreted within the context, as the meaning of this word in the sentence can vary depending on the sentence structure or the word surrounding it.

   When looking deeper into the data point for the most frequently occurring sentences, as shown in figure 5.13, some similar sentences that occurred previously in the fake news label also appear, such as “Indonesia” and “information”. Another interesting finding is that the word “video” is not shown in the top 10 most frequently occurring sentences, indicating that video is not a primary source to validate information since the fact column consists of a list of evidence based on the given claim in the title column. Instead, “photo” occurs in this list to indicate the use of photos as supporting information for the fact-checking process.

Word		Score
English	**Indonesian**	
social	sosial	0.53
sanitation	sanitasi	0.28
due to	akibat	0.26
video	video	0.23
facilitate	memudahkan	0.23
disappear	menghilang	0.22
help	pertolongan	0.21
paste	paste	0.20
write	tuliskan	0.18
urban village	kelurahan	0.18

Table 5.6: Top 10 Most Important Words for the Support Class for Claim Verification Task

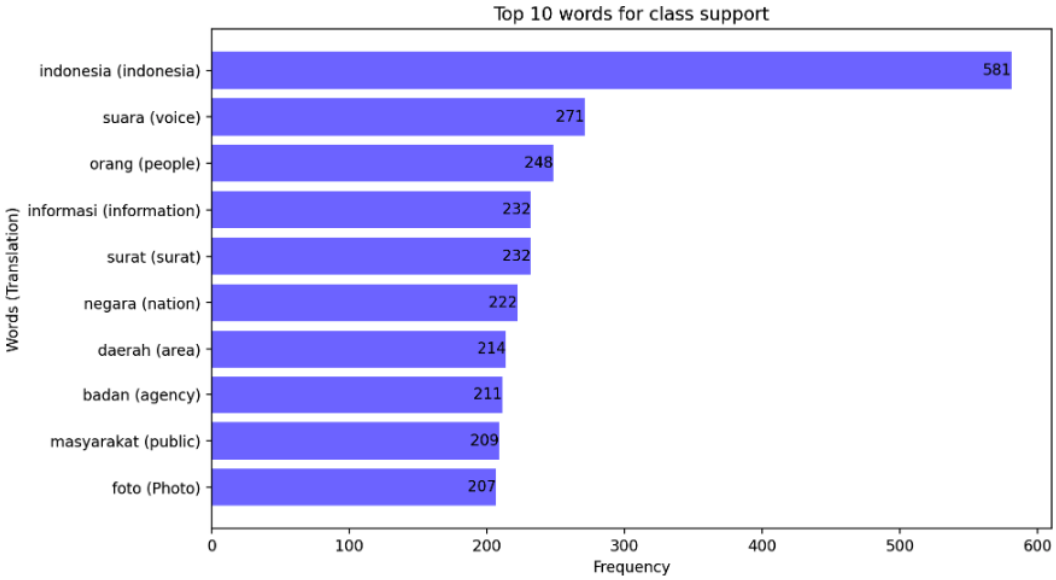

Figure 5.13: Top 10 most occurring words in Support Class

The majority of the text data is less than 1000 characters, with a few outliers exceeding 8000 characters, as shown in figure 5.14. This indicates that there is no need for much evidence to refute a claim. A few short pieces of evidence are sufficient to refute and evaluate the veracity of a claim.

2. **Reject**

   The reject class represents where the given evidence rejects the claim. The reject class consists of all other seven categories of fake news. When TF-IDF is utilized to find the essential word, as shown in table 5.7, we found that the word “authentic” is mentioned on the top, indicating that discussion regarding the authenticity of something is shared in rejected claims. This is due to the nature of fake news.

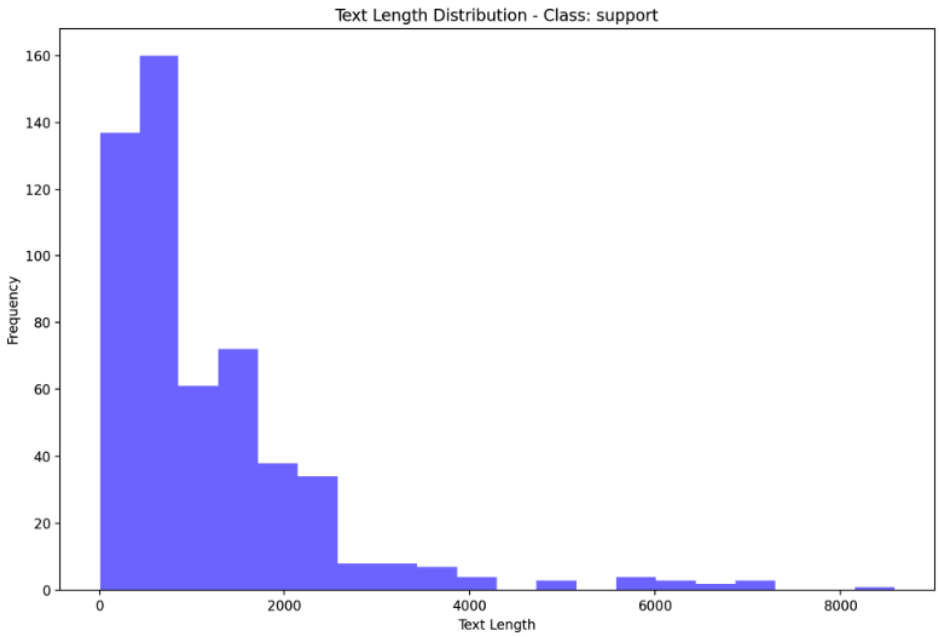

Figure 5.14: Text Length Distribution for Support with Claim Verification Label Dataset

Word		**Score**
English	**Indonesian**	
authentic	asli	0.31
in the form of	berupa	0.29
Kediri	kediri	0.25
information	information	0.24
appear	muncul	0.24
reported	dilansir	0.22
Instagram	Instagram	0.21
titled	berjudul	0.18
place	tempat	0.15
offer	menawarkan	0.15

Table 5.7: Top 10 Most Important Words for the Reject Class for Claim Verification Task

Most of it is a modified version of the original ‘post where the authenticity of this post is being questioned since some of the information is being manipulated on purpose. Key terms, such as “information, “appear”, “Instagram”, etc., are crucial within the “reject” class.

When looking into the most frequently occurring word, the result looks similar to the support class as shown in figure 5.15. Some keywords such as “results”, “searching”, and “based on” show that most of the evidence is gathered by searching on a search engine. The word “account,” which refers to a social media account, indicates that most of the article or post was sourced from social media.

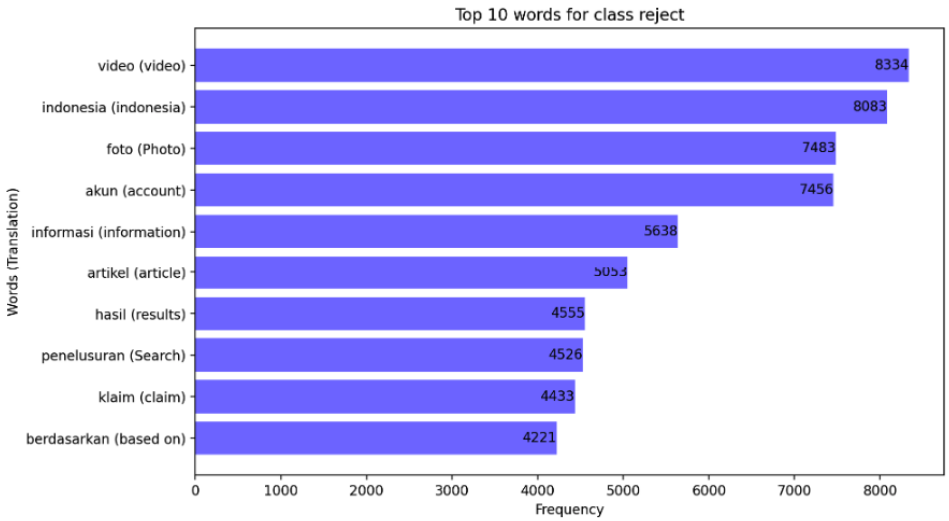

Figure 5.15: Top 10 most occurring words in reject class

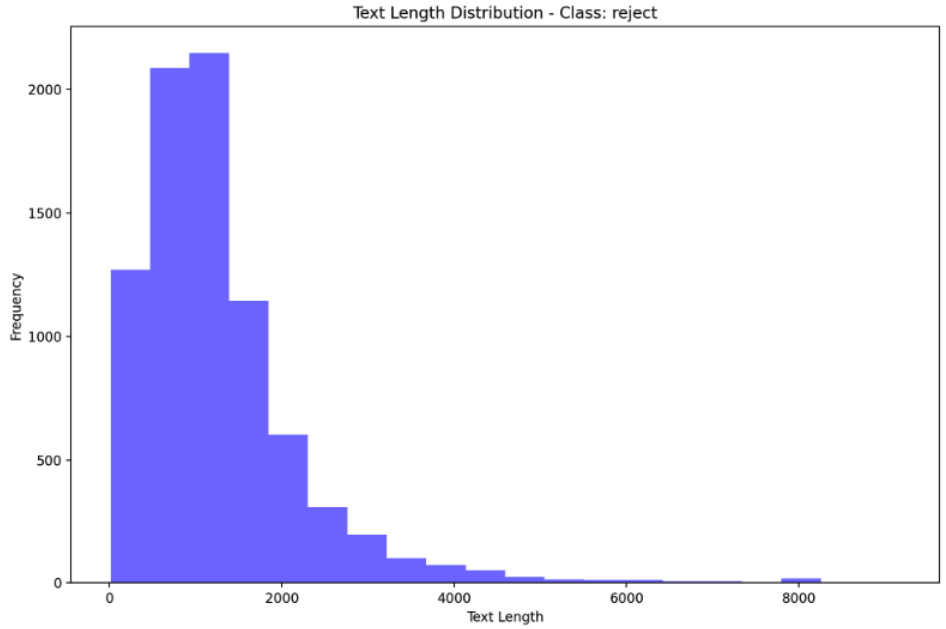

Figure 5.16: Text Length Distribution for Reject Class with Claim Verification Label

When looking into the text length distribution as shown in figure 5.16, we found that most of the articles in this class range around 500–200 thousand words, which is comparatively more extended than the “support” class. This is an indication that more evidence is needed to debunk the fake news since the “reject” class only consists of the fake news class. Some outliers are also shown, where this article has a size of around 8000 characters.

5.3 Data preprocessing

5.3.1 Word-based classifiers

To clean text data for the word-based classifiers, we perform the following steps:

1. **Removing the Unnecessary Column**: the unnecessary column is removed, leaving the dataset with only six columns: title, content, fact, source issue, tanggal (which translates to date), and classification.

2. **Text Normalization**: we normalized the use of Indonesian slang language, known as *kamus alay* ('alay dictionary'). This type of word is usually made up of a sentence. It is sometimes derived from a viral post in which an influencer or the Indonesian media makes up a new word for an existing word, and it becomes viral and is adapted by Indonesians. One example of this Indonesian slang is 'mager', derived from the phrase 'malas gerak', which translates to 'lazy' or 'unwilling to move or do something'. We used an existing list of *alay*words [1] and used it to perform normalization of the texts.

3. **Text Normalization with Stemming**: the dataset obtained from Mafindo is still raw and requires extensive preprocessing; one data cleaning method is stemming. For instance, the words "likes," "likely," and "liked" are all derived from the same root word, "like," and "like" can be used as a synonym for all three. For the context of the Indonesian language, let's use the verb 'berlari" which means "to run" as an example. This word's origin is "lari," which means "to run"; related words include "berlarian" (to be running around), "terlari-lari" (to run aimlessly), "larian" (running track), etc. An NLP model can determine that all three words are related in some way and are used in related contexts. Stemming allows us to standardize words to their base stem regardless of inflection, which is helpful for a wide range of applications such as clustering and text classification. The word can be standardized as well. The dataset was stemmed using a Sastrawi stemmer in the data processing step. This Python library allows us to reduce inflected Indonesian (Bahasa Indonesia) words to their root form (stem) [44].

4. **Removing Unnecessary Information From the Title Column**: in the dataset, the annotator occasionally included the classification category in the title of their reviewed article. This should be removed because it is redundant, unrelated to the title, and a data leak from the labels to the content, which will be used as the claim in the claim verification task.

[1] https://github.com/nasalsabila/kamus-alay

Classification	Total
Misleading Content	2629
False Context	2234
Fabricated Content	1152
Manipulated Content	1152
Impostor Content	503
TRUE	402
Clarification	396
False Connection	279
Satire	162

Table 5.8: Total count of each classification category in the Mafindo dataset after data preprocessing

5. **Spell Correction**: the dataset also contains many typos, which are addressed using a spelling checker based on Peter Norvig's Spell Checker algorithm [35] and the Indo4DB dataset [56]. This spell checker works by gathering a large corpus of cleaned text, in this case, using the Indo4DB dataset to build the language model. When a word is misspelled, the algorithm will generate a possible corrected word by performing operations such as deletion, transposition, and character replacement or insertion. Each correction is scored based on the number of occurrences in the corpus, and the one with the highest score is chosen as the corrected word.

After the necessary data preprocessing, the final dataset consists of 2827 articles. The detailed and updated count for each class is shown in table 5.3.1. Besides the fact that the 'unclassified' class is removed, there are no changes in all other classes since the null value in the dataset was filled in manually or taken from another related column.

5.3.2 Transformer-based models

Since the dataset will be trained with IndoBERT Model, which only accepts numeric values as labels, the claim verification label will be converted to numeric values and stored in a new column named **label**.

We summarize text since the BERT model can only accept 512 input tokens. There are 351 texts longer than 512 tokens in the "content" column and 2279 texts longer than 512 tokens in the "facts" column. This dataset will be truncated when fed into the BERT model, which could result in information loss. Therefore, we summarize all datasets that are longer than 512 tokens, while the rest that are less than 512 are kept.

We are summarising using the Indonesian BERT2BERT Summarization Model, available in HuggingFace [57]. This is a fine-tuned encoder-decoder model for Indonesian text summarization trained on Liputan6.com news articles, one of Indonesia's most

prominent news media companies [23]. Cahya, the Indonesian BERT2BERT Summarization Model developer, provided a code implementation in his HuggingFace on how to use the train model, which we adapted for our dataset. [2] Both the model and tokenizer of the pre-trained BERT2BERT summarization model are utilized during the summarization process. The tokenizer converts the unprocessed input into a sequence of tokens for input to the model using the model's pre-trained weight. No datasets are longer than 512 tokens after the summarization process, which prevents information loss when fed into the BERT model for later experimentation.

5.4 Dataset splitting

We divided the dataset into two sets for fake news classification and claim verification tasks, each containing a training set (75% of the dataset) and a test set (25% of the dataset). In addition, when splitting the dataset, we used stratified sampling to ensure that the proportion of classes in the training and test sets was the same as in the original dataset. This method is beneficial when datasets are imbalanced, such as the original dataset. Without stratification, a random dataset split may not accurately reflect the distribution of the original dataset, which could result in potential performance issues. We are stratified sampling based on the "classification" column for the fake news classification task and the "claim" column for the claim verification task.

[2] https://huggingface.co/cahya/bert2bert-indonesian-summarization

Chapter 6

Results

This experiment investigates the performance of traditional machine learning algorithms (SVM, Random Forest, and Linear Regression) and transformer-based IndoBERT for automatic fake news classification. In addition, this research also utilizes IndoBERT and GPT-3 for claim verification tasks. The result of this experiment gives substantial insight into how machine learning and natural language processing can be used to combat the growing problem of misinformation, especially in Indonesia. Some metrics, such as precision, recall, f1-score, and accuracy, evaluate the model's performance. This experiment's setup and hardware are as below. The code for the experiment is available on GitHub. [1] In addition, we also provide the pre-trained IndoBERT model for the fake news classification [2] and claim verification tasks [3] on Hugging Face.

6.1 Automatic classification

The experiment used a traditional machine learning strategy with multiple classifiers and the BERT model. Furthermore, multiple imbalanced learning approaches were used because the data was unbalanced, and the performance was compared to unbalanced data. The model is then trained using IndoBERT, a pre-trained Indonesian BERT model.

[1] https://github.com/BrianArnesto/Fake-news-classification-and-claim-verification-for-supporting-Indonesian-fact-checkers

[2] Pre-trained IndoBERT for fake news classification: https://huggingface.co/brianarnesto/IndoBERT-Fake-News-Classification

[3] Pre-trained IndoBERT for claim verification: https://huggingface.co/brianarnesto/IndoBERT-Claim-Verification

6.1.1 Traditional machine learning

We use the "content" column as input to the classifiers, normalized for spelling errors and slang as described in section 5.3. We transformed the text content to a bag-of-words representation with tf-idf weights, and we experimented with three classification models: Logistic Regression, Random Forest, and Linear SVM. Using the grid search strategy, many combinations of hyperparameters, as mentioned in Section 4.1.1, are being tested to find the best parameter for each classifier.

1. **Logistic Regression**
   We are experimenting with this classifier with different combinations of hyperparameters and we found that the best parameters combination for Logistic Regression is: 'C' : 1, '$penalty'$: '$l2'$. As shown in table 6.1, the model performs exceptionally well for classifying the "impostor content" class" with a precision of 0.86; this indicates a high true positive rate and fewer false positives. For the recall, the "misleading content" class has the highest score, with 0.63, but for this class, the precision score is quite low at 0.44, indicating many false positives. In addition, impostor content" has the highest F1-score of 0.65, making this class the overall highest performing class. With F1-scores of 0.50 and 0.45, respectively, "fabricated content" and "false context" also perform well. In contrast, categories like False Connection' and satire perform poorly, with a 0 score in all evaluation metrics, indicating that the classifier failed to learn in these two classes. Furthermore, This classifier has an accuracy of 46%.

Classification	**precision**	**recall**	**f1-score**	**number of test item**
Clarification	0.33	0.04	0.07	99
Fabricated Content	0.58	0.44	0.50	288
False Connection	0.00	0.00	0.00	70
False Context	0.39	0.53	0.45	559
Impostor Content	**0.86**	0.52	**0.65**	126
Manipulated Content	0.51	0.28	0.37	288
Misleading Content	0.44	**0.63**	0.52	657
Satire	0.00	0.00	0.00	40
TRUE	0.77	0.32	0.46	105
macro avg	0.43	0.31	0.33	2232

Table 6.1: Logistic Regression for Fake News Classification with the hyperparameter setting: 'C' : 1, '$kernel'$: '$linear'$.

2. **Support Vector Machines (SVM)**
   When we applied the SVM algorithm to this problem, we found that the best parameter for this classifier is: 'C' : 1, '$kernel'$: '$linear'$. When we trained the SVM

classifier with the given hyperparameter, the result was quite inconsistent among all classes, as shown in table 6.2, as also happened with the logistic regression classifier. When we trained the SVM classifier with the given hyperparameter, the result was quite inconsistent among all classes, as with the logistic regression classifier. Regarding recall, the "false connection" class has a score of 1, which is entirely consistent with the low recall and f1-score in this class of 0.001 and 0.03, respectively. This is an unusual result having a perfect recall score but a very low recall and f1-score happens since the classifier only makes a few correct predictions. This could lead to a high precision score since there are no false positives but a low recall score since there are numerous false negatives. This is because the dataset is imbalanced, and particularly for the "false connection" class, there are only 70 data points in the test set that were used to evaluate the classifier.

Classification	**precision**	**recall**	**f1-score**	**number of test item**
Clarification	0.30	0.09	0.14	99
Fabricated Content	0.58	0.46	0.51	288
False Connection	**1.00**	0.01	0.03	70
False Context	0.40	0.55	0.46	559
Impostor Content	0.84	0.60	**0.70**	126
Manipulated Content	0.52	0.32	0.39	288
Misleading Content	0.44	**0.60**	0.51	657
Satire	0.00	0.00	0.00	40
TRUE	0.74	0.40	0.52	105
macro avg	0.54	0.34	0.36	2232

Table 6.2: Support Vector Machines for Fake News Classification with the hyperparameter setting: 'C' : 1, '$kernel'$: '$linear'$

Furthermore, for the "impostor content" class, compared to the logistic regression classifier, the recall dropped slightly to 0.60 with a high precision score of 0.84 and a representative f1-score of 0.70. This is an indication that the "impostor content" class has a high rate of true positives and a low rate of false positives. The fact that the recall score of the "misleading content" class has increased to 0.60 shows that the classifier is now able to recognize this class better. The classifier's overall accuracy is 47%, performing slightly better than the Logistic Regression classifier.

3. **Random Forest**
   We found the best hyperparameter for this Random Forest with the parameter: 'max_depth' : $None$, '$n_estimators'$: 200'. The random forest classifier still shows an inconsistent result, as shown in table 6.3, as also happened to the previous

two classifiers. The "impostor content" class has the highest precision score of 0.81, showing that the classifier was able to correctly predict 81% of the instances in this class to be in the correct class. The "satire," on the other hand, had the lowest precision of 0.01, showing that only 8% of the instances classified as this class indeed belong to the "satire" class. Furthermore, the "misleading content" class has the highest recall among the rest with 0.73. In contrast, the "false connection" and "clarification" have the lowest score with 0.01. Regarding the f1-score, the "impostor content" class has the highest score, with 0.66. The classifier's macro-average precision, recall, and f1-score were 0.48, 0.29, and 0.31, respectively. The low values of these macro average scores, particularly in recall and f1-score, indicate that while the model performs better in some classes, it generally struggles to maintain this performance across all classes.

Classification	**precision**	**recall**	**f1-score**	**number of test item**
Clarification	0.50	0.01	0.02	99
Fabricated Content	0.69	0.30	0.41	288
False Connection	0.25	0.01	0.03	70
False Context	0.39	0.45	0.42	559
Impostor Content	**0.81**	0.56	**0.66**	126
Manipulated Content	0.60	0.11	0.19	288
Misleading Content	0.39	**0.73**	0.51	657
Satire	0.08	0.05	0.06	40
TRUE	0.65	0.39	0.49	105
macro avg	0.48	0.29	0.31	2232

Table 6.3: Random Forest for Fake News Classification with the hyperparameter setting: $'max\_depth' : None$, $'n_estimators' : 200'$

6.1.2 Imbalanced learning

As previously mentioned, the dataset is imbalanced among all classes. We are employing the oversampling and undersampling methods to handle the issues from the experiment we are conducting for the three classifiers mentioned before, such as Logistic Regression, Random Forest, and Linear SVM. We found that Linear SVM, with the hyperparameters $'C' : 1$, $'penalty' : 'l2'$ performed the best among the rest. Therefore, we use this classifier with the specified hyperparameter for imbalanced learning for both oversampling and undersampling methods.

Classification	precision	recall	f1-score	number of test item
Clarification	0.22	0.15	0.18	99
Fabricated Content	0.47	0.49	0.48	288
False Connection	0.11	0.09	0.10	70
False Context	0.40	0.45	0.42	559
Impostor Content	**0.88**	**0.52**	**0.66**	126
Manipulated Content	0.34	0.40	0.37	288
Misleading Content	0.45	0.48	0.47	657
Satire	0.08	0.03	0.04	40
TRUE	0.67	0.47	0.55	105
macro avg	0.40	0.34	0.36	2232

Table 6.4: Oversampling with SMOTE combine with Linear SVM for fake news classification

Oversampling with SMOTE

We perform imbalanced learning by combining SMOTE and Linear SVM with the specified hyperparameter mentioned before; the result is shown in table 6.4. We found that the "impostor content" had the highest precision, recall, and f1-score with respective values of 0.88, 0.52, and 0.66. In contrast, the "satire" category performed better than the rest for all three metrics (precision, recall, and f1-score), with respective values of 0.07, 0.23, and 0.11. This indicates that the combination of Linear SVM and SMOTE is having difficulties correctly predicting this class. Further, the macro average score for this model for precision, recall, and f1-score was 0.40, 0.34, and 0.36, respectively. These low values indicate the model's performance is inconsistent among each class. Despite the help of SMOTE in balancing the training data, as observed in table 6.4, the model still struggles to predict most of the class, resulting in poor overall performance. In addition, the model has 43% overall accuracy.

Down sampling

We are performing the under-sampling method with Random Under Sampling and NearMiss with all three variants of NearMiss.

- **Random Under Sampling**

  The performance of the random undersampling method when combined with Linear SVM is quite varied, as shown in table 6.5. However, it shows the same result as the oversampling method with SMOTE, where the "impostor content" class yields the best result regarding the precision, recall, and f-1 score with 0.74,

Classification	precision	recall	f1-score	number of test item
Clarification	0.10	0.28	0.15	99
Fabricated Content	0.52	0.40	0.45	288
False Connection	0.08	0.36	0.13	70
False Context	0.37	0.23	0.28	559
Impostor Content	**0.74**	**0.64**	**0.69**	126
Manipulated Content	0.23	0.30	0.26	288
Misleading Content	0.52	0.24	0.32	657
Satire	0.07	0.23	0.11	40
TRUE	0.36	0.53	0.43	105
macro avg	0.33	0.36	0.31	2232

Table 6.5: Random Under Sampler combine with Linear SVM for fake news classification

0.64, and 0.69. The precision in this class is lower than the previous oversampling method, but the recall score increased significantly, resulting in a higher f1-score. Meanwhile, the "satire" class has the lowest score across all metrics for recall and f1-score, at 0.07, 0.23, and 0.11. The macro average score's precision, recall, and f1-score were 0.33, 0.36, and 0.31, respectively. This method has an overall accuracy of 31%.

- **NearMiss**
  We are experimenting with the NearMiss undersampling technique to address the issue of data imbalance. We are conducting experiments with three NearMiss variants, including NearMiss-1, NearMiss-2, and NearMiss-3.

  1. NearMiss-1
     With the NearMiss-1 variant, the "impostor content" showed the best performance compared to other classes in all metrics (precision, recall, and f1-score) with 0.49, 0.61, and 0.54 as shown in table 6.6. There is a significant decrease in performance in this class compared to the Random undersampling method. The "false connection" class had the lowest f1-score of 0.10, indicating that this method has issues predicting this class. The macro average scores for the precision, recall, and f1-score were respectively 0.29, 0.33, and 0.21. The accuracy of this model is quite low, with an overall accuracy of 18%.

  2. NearMiss-2
     When we experiment with NearMiss-2, as shown in table 6.7, there is an increase in overall performance compared to the previous NearMiss-1. The highest-performing class is still the "impostor content" class for all the metrics, with 0.61 precision, 0.72 recall, and 0.66 f1-score. The "satire" and

Classification	precision	recall	f1-score	number of test item
Clarification	0.11	0.47	0.18	99
Fabricated Content	0.42	0.28	0.34	288
False Connection	0.06	0.51	0.10	70
False Context	0.40	0.04	0.08	559
Impostor Content	**0.49**	**0.61**	**0.54**	126
Manipulated Content	0.35	0.15	0.21	288
Misleading Content	0.58	0.05	0.10	657
Satire	0.05	0.30	0.08	40
TRUE	0.17	0.51	0.26	105
macro avg	0.29	0.33	0.21	2232

Table 6.6: Near-Miss1 combine with Linear SVM for fake news classification

"false connection" both have the same result on precision and f1-score and are also the lowest among other classes with 0.06 and 0.10, respectively. The macro average for precision was 0.30, the recall was 0.36, and the f1-score was 0.29. The accuracy also increased to 28%.

Classification	precision	recall	f1-score	number of test item
Clarification	0.14	0.42	0.21	99
Fabricated Content	0.35	0.37	0.36	288
False Connection	0.06	0.31	0.10	70
False Context	0.41	0.13	0.20	559
Impostor Content	**0.61**	**0.72**	**0.66**	126
Manipulated Content	0.31	0.27	0.29	288
Misleading Content	0.47	0.21	0.29	657
Satire	0.06	0.28	0.10	40
TRUE	0.30	0.57	0.39	105
macro avg	0.30	0.36	0.29	2232

Table 6.7: NearMiss-2 combine with Linear SVM for fake news classification

3. NearMiss-3

   In this experiment, the accuracy further increased to 32%. The "impostor content" class is still the highest performing, with a precision score of 0.51, a recall score of 0.62, and an f1-score of 0.56. In contrast, the "satire" remained with the lowest performance, with precision, recall, and an f1-score of 0.05, 0.15, and 0.07, respectively. The macro average for the precision score was 0.29, the recall was 0.35, and the f1-score was 0.29.

Classification	precision	recall	f1-score	number of test item
Clarification	0.13	0.33	0.19	99
Fabricated Content	0.36	0.26	0.30	288
False Connection	0.12	0.21	0.16	70
False Context	0.39	0.38	0.39	559
Impostor Content	**0.51**	**0.62**	0.**56**	126
Manipulated Content	0.28	0.28	0.28	288
Misleading Content	0.53	0.24	0.33	657
Satire	0.05	0.15	0.07	40
TRUE	0.27	0.65	0.38	105
macro avg	0.29	0.35	0.29	2232

Table 6.8: NearMiss-3 combine with Linear SVM for fake news classification

6.1.3 IndoBert model

We are fine-tuning the IndoBERT model with 30 epochs, and we observe the loss value for each epoch and plot it. As shown in Figure 6.2, the model had an average loss of 1.64 in the first epoch. Since the model parameters are randomly initialized, it is common for the model to have a high loss value in the early iterations of training, resulting in significant discrepancies between the prediction model and the data. The loss value keeps decreasing, indicating that the model's prediction ability has improved. The model's loss value increases slightly after the 20th epoch to the 22nd epoch, indicating that the model tends to overfit, where the model learns from noise in the data or outliers in the data, limiting the model's ability to generalize. The loss value gradually decreased until the 30th epoch as, the final epoch, where the model got a 0.1 average loss.

The fine-tuned IndoBERT model is evaluated using multiple metrics, such as precision, recall, f1-score, and overall model accuracy. The overall model performance is quite diverse since the dataset is imbalanced. Some classes have a small size, making it challenging for the model to predict those classes.

In terms of the precision score, the "impostor content" class had the highest precision of 0.72, which indicates that the model correctly recognizes this class and has a low rate of false positives. On the other hand, "false connection" had the lowest precision of 0.09, which means that the model misclassifies the majority of the data point in this class, which leads to a lot of false positives. Furthermore, on the recall score, the "impostor content" class has the highest score with 0.67. For every 100 articles in the "impostor content" class, the model has 67 true positives and 33 false negatives. In contrast, the "satire" class had the lowest recall of 0.05, which means this class has a lot of false negatives.

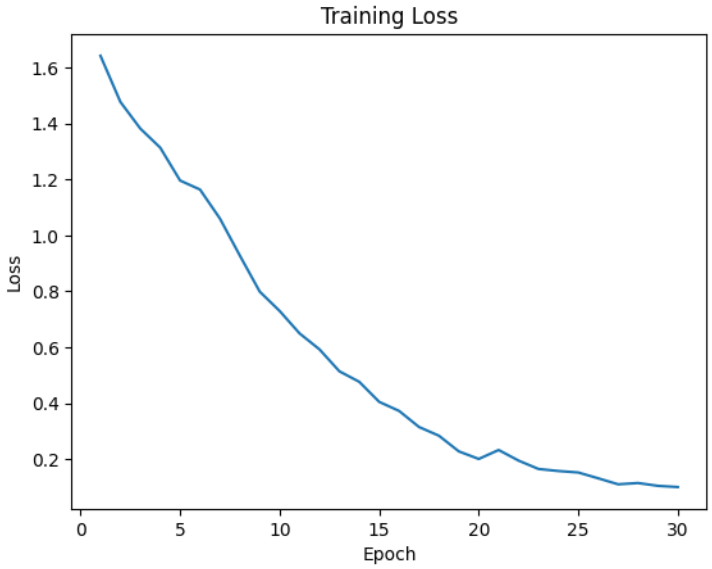

Figure 6.1: Training Loss during Fine-tuning the IndoBERT Model for Fake News Classification Task

Classification	**precision**	**recall**	**f1-score**	**number of test item**
Misleading Content	0.45	0.51	0.48	657
False Context	0.46	0.38	0.41	559
Fabricated Content	0.45	0.52	0.49	288
Manipulated Content	0.34	0.35	0.35	288
Impostor Content	**0.72**	**0.67**	**0.69**	126
Clarification	0.32	0.26	0.29	99
TRUE	0.47	0.60	0.53	105
False Connection	0.09	0.10	0.10	70
Satire	0.15	0.05	0.08	40
macro avg	0.39	0.38	0.38	2232

Table 6.9: Results per category for the IndoBERT model

Moreover, for the f1-score, the score for every class is quite diverse, with the “impostor content” class having the highest score of 0.69, which means that this category has better precision and recall than others. On the other hand, the model has a very low F1-score for both ”False Connection” and “Satire”. as shown by its F1-scores of 0.10 and 0.08, respectively. This also indicates that these two classes have low accuracy and recall scores.

Table 6.10 shows that the term-based classifiers have higher precision, while IndoBERT has a higher recall. SVM performs best with the highest precision of the three term-based classifiers. Since SMOTE is unsuitable for embedding-based models like BERT, we evaluated the data balancing methods only on the best-performing term-

based classifier (SVM).

	Precision	**Recall**	**F1-Score**
Logistic Regression	0.43	0.31	0.33
SVM	**0.54**	0.34	0.36
Random Forest	0.48	0.29	0.31
IndoBERT	0.39	**0.38**	**0.38**

Table 6.10: Results for the original imbalanced data for the fake news classification task. Precision, Recall, and F1 are Macro averages over all classes

The results of the data balancing in Table 6.11 show that oversampling with SMOTE gives the best performance in terms of F1, while random undersampling leads to the highest recall but at the cost of lower precision.

6.2 Claim verification

We address the claim verification task with the IndoBERT model and GPT-3.5. [4]

6.2.1 IndoBERT

We are pretraining the IndoBERT for ten epochs, and the model iteratively refines its predictions to minimize the loss in each epoch. The average loss is computed after each epoch to track the model's learning progress. We then monitored the loss value progression for each epoch during the training process, as shown in figure 6.2. The loss value represents the comparison between the actual label, or ground truth, and the model's prediction on the training data. Lower loss values indicate that the model could predict the actual label. The cross-entropy loss method is used to evaluate the model's loss value. This method is intended to increase the penalty for confident but incorrect predictions in cases where the model is confident about a prediction that turns out to be incorrect, making it a reliable performance metric for classification models. The model got an average loss of 0.0052 in the final epoch, the 10th epoch; this relatively low value indicates that the model learned adequately. The test set is then fed into the pre-trained model for evaluation to evaluate the performance of the fine-tuned model.

The results for claim verification are shown in Table 6.12. IndoBERT's overall accuracy is 94%, and the Macro averaged F1 over the two classes is 0.81. Because the dataset is imbalanced, the model performs differently for both classes. However, the model

[4]Note that term-based classifiers are unsuitable because they do not allow a cross-encoded input of two texts (claim and evidence).

	Precision	Recall	F1
No balancing	**0.54**	0.34	0.36
Oversampling with SMOTE	0.40	0.34	**0.36**
Random Undersampling	0.33	**0.36**	0.31
NearMiss-1	0.29	0.33	0.21
NearMiss-2	0.30	0.36	0.29
NearMiss-3	0.29	0.35	0.29

Table 6.11: Results with the balanced data using the SVM classifier for the Fake News Classification Task. Precision, Recall, and F1 are Macro averages over all classes

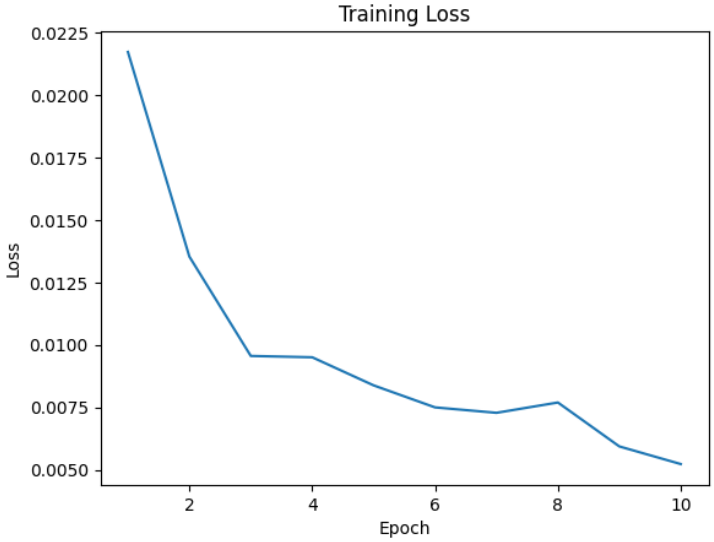

Figure 6.2: Training Loss during Fine-tuning the IndoBERT Model for Claim Verification Task

performs well when evaluated using multiple metrics such as precision, recall, and f1-score. The model generally has a high precision score, indicating that it generalizes well to unseen data. The model performed exceptionally well for the “support” class, especially with the precision score, while still having a low score for the recall and f1-score, with 0.74 for the precision score, 0.58 for the recall score, and 0.65 for the f1-score. This indicates that 74% of the dataset predicted by the model as ”support” was indeed ”support,” and there are 58% of the actual ”support” dataset was correctly identified by the model.

Furthermore, the model’s precision, recall, and f1-score for the ”reject” class are 0.96 and 0.78, respectively. This demonstrates that the model accurately identifies and predicts instances of ”reject.” Despite this, the performance gap between the ”support” and ”reject” classes suggests that the model predicts the ”reject” class more accurately than the ”support” class.

Classification	precision	recall	f1-score	number of test item
support	0.74	0.58	0.65	204
reject	0.96	0.78	0.81	2028
MACRO AVG	0.85	0.78	0.81	2232

Table 6.12: Claim Verification with IndoBERT model

Classification	precision	recall	f1-score	number of test item
support	0.24	0.34	0.28	152
reject	0.96	0.81	0.88	2019
MACRO AVG	0.60	0.57	0.58	2171

Table 6.13: Claim Verification with GPT-3 model

6.2.2 Generative pre-trained transformer (GPT) 3

We perform claim verification using the GPT model, and the result is shown in table 6.13. The table also shows the results for the GPT model, which can provide a reasonable result even without fine-tuning, albeit much lower than IndoBERT. The overall accuracy of the GPT model is 80%

6.3 Workflow of human and automated method

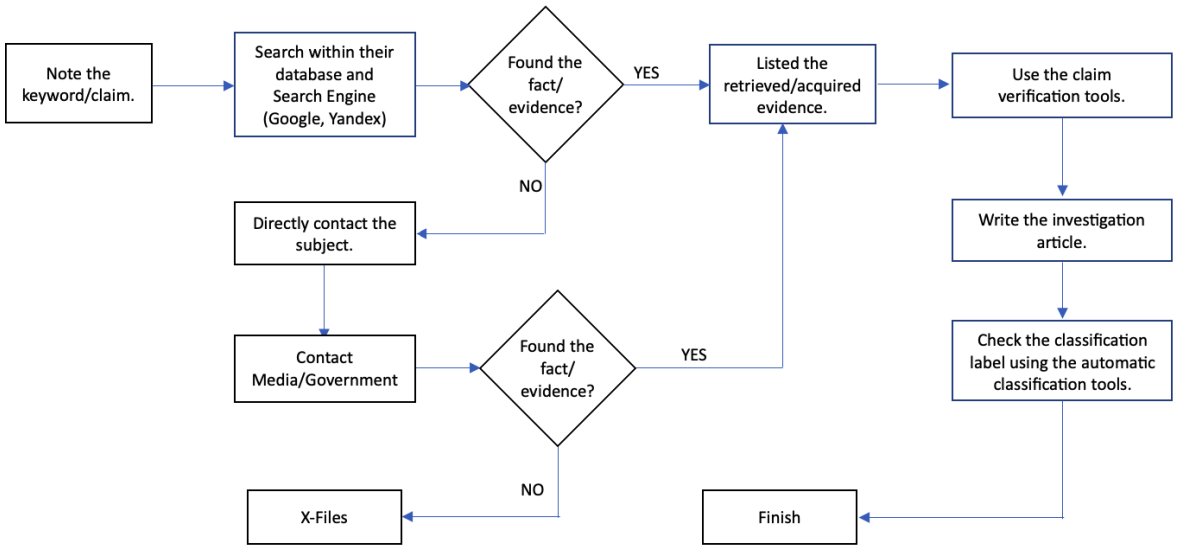

Figure 6.3: Combined Workflow of Human and Automated Method

The combined workflow of human and automated method work in this manner as shown in figure 6.3:

1. The human annotator takes notes of the keyword or the claim.
2. The annotator checks whether the claim has already been verified by them or other mainstream media by looking at their internal database and using a search engine to find relevant information.
3. If the fact or evidence is found, this fact and evidence are then collected and stored. If not, the procedure for handling this issue will be further explained in procedure 7 and onwards.
4. The annotator then used the claim verification tools to determine the claim's veracity.
5. Upon finding and collecting the supporting evidence, the annotator then writes an investigation article based on the collected evidence or facts.
6. The automatic classification tools are then used to determine the classification label of the given claim. After the classification label has been defined, this task is considered finished.
7. Continuing from procedure 3 if the fact is not found, the annotator will try to find and track down the source of the original post, as well as contact the author of the post or the person that is mentioned in the claim. Since the origin of the post in cases of fake news does not usually come from the person or institution mentioned in the claim, this type of post is created to give a negative impression of the person or institution or to frame them as something they are not.
8. if none of this is reachable if they did not have enough or sufficient information, or if the claim is about a specific topic rather than a person. The annotator then contacts news organizations or government agencies to find out if they have experts or research that could address the issue.
9. If sufficient evidence has been gathered after attempting to contact numerous sources related to the claim or issue, the next step is to proceed as described in procedure 4, where all collected evidence is fed into the claim verification tool and an investigation article is written based on the evidence. The automatic classification tool will be used again to determine the claim's classification label. The task is considered complete once the classification label is obtained.

Chapter 7

Discussion

We discovered that there is a lack of publicly available large annotated datasets with multi labels for both fake news classification and claim verification tasks. Most publicly available datasets are binary classes and have limited sample sizes. Therefore, we approached Mafindo, one of Indonesia's largest fact-checking agencies, and obtained API access to their annotated dataset collection.

To gain insight into the dataset, we experimented with multiple exploratory data analysis methods, and the results showed that in terms of the top 10 most occurring words, each dataset had a different label across each class. The words "Indonesia video, *akun* (account), and *foto*(photo) show that the dataset, in general, has similar characteristics across different classes. In addition, we are using TF-IDF to see the most important word in each class. The result of this experiment allows us to deeply unravel the characteristics of each class, especially with the fake news classification label. For example, we found that the "misleading content" class mostly discussed the previous COVID-19 pandemic and the vaccine. In contrast, the "fabricated content" class was full of fraudulent content that the author created to scam the reader.

This classification task is challenging because the fake news classification dataset is imbalanced, and some classes only have a few training items. The three term-based classifiers produced F1 scores between 0.31 and 0.36, with Linear SVM performing slightly better than the others. Oversampling or undersampling the data did not improve the results and significantly lowered the precision. A fine-tuned IndoBERT model outperforms the term-based classifiers on Recall and F1 but not on precision.

For the claim verification task, the fine-tuned model only has to distinguish two labels (support and reject), which is a much easier classification task. This dataset is still imbalanced, with most of the dataset labeled reject. This is reflected in the 0.97 F1

score for the ‘reject’ class and the 0.65 for the ‘support’ class achieved by the IndoBERT model. Our experiments with GPT-3.5 for claim verification in Indonesian indicate that it cannot reach the quality level of the fine-tuned IndoBERT model. Since GPT has not seen any task-specific training (fine-tuning) data, the misclassifications are not caused by an imbalance in our data. We can only speculate that the model was more inclined to the reject class because of either the content of the pre-training data underlying GPT or the writing style of the evidence in the Mafindo data. Given the fact-checking purpose, the content might be more about debunking claims than supporting them. Future work on open-source generative large language models could shed more light on our findings.

When implemented to determine the most important word for each class, feature extraction using TF-IDF and logistic regression is a very effective method for gaining a deeper understanding of the characteristics of each class within a dataset. The primary keyword in each category of misinformation identifies the type of information spread, the method of dissemination, and the target recipient.

Imbalanced learning cannot address the issue of unbalanced data in the fake news classification task. This is due to a massive gap between classes, with some classes having a significantly smaller dataset than others. A pre-trained IndoBERT model specified for the Indonesian language should perform better than the other model. Still, due to the small size of the dataset and the high imbalance among the datasets, the IndoBERT model does not perform as well as expected.

Despite the imbalanced dataset issue, the model performed exceptionally well in the claim verification task because there are only two classes in the dataset. However, the IndoBERT model outperforms the GPT-3 model in the claim verification task. In addition, more labeled datasets are needed, particularly for the ”support” class, to improve this model’s performance in this class. Combining a human and automatic workflow with Natural Language Processing (NLP) hopes to speed up the work of a human annotator. Because there is a massive amount of misinformation created and spread every day, there is a need for tools or a workflow that can improve the annotator’s work so that they can work on more claims or issues and work more quickly and efficiently.

Chapter 8

Conclusion, Limmitation & Future Work

8.1 Conclusion

This research was conducted to develop tools that can automate specific tasks in the manually annotated fact-checking task using NLP and machine learning and also provide a workflow that combines these tools with the existing workflow that this institution currently has. The proposed research questions, as well as how we approach and address the research question, are as follows:

To what extent can traditional machine learning techniques such as Support Vector Machines, Logistic Regression, Random Forests, and deep learning models such as the IndoBERT Model be utilized for fake news detection and classification into the respective categories?
In this study, we are utilizing traditional machine learning classifiers such as logistic regression, SVM, and Random Forest alongside the deep learning model IndoBERT. The experiment shows that SVM achieves the highest precision score of 0.54, surpassing all other models, while the deep learning model has the highest recall and f1-score. The dataset is imbalanced, which makes the fake news classification task quite challenging. We are utilizing various resampling techniques to address this data imbalance issue, including SMOTE for oversampling, Random undersampling, and three variants of NearMiss for oversampling. As shown in 6.11, while these methods improved some performance metrics (Random Under Sampling, NearMiss-2, and NearMiss3 improved the recall slightly), none uniformly improved all the performance metrics. Imbalanced learning cannot address the issue of unbalanced data in the fake news classification task. This is due to a massive gap between classes, with some classes having a sig-

nificantly smaller dataset than others. Using a pre-trained IndoBERT model specified for the Indonesian language and fine-tuning the model with the Indonesian fake news dataset should perform better than the other model. Still, due to the small size of the dataset and the high imbalance among the datasets, the IndoBERT model does not perform as well as expected.

We conclude that while traditional machine learning and deep learning models such as IndoBERT still show promising prospects for being used on fake news classification tasks, the issue of class imbalance in the dataset remains a significant challenge for this task.

To what extent can BERT and GPT models, particularly IndoBERT for the Indonesian language, be utilized in claim verification as a preliminary step in manually fact-checking fake news, and how do these models perform when applied for claim verification tasks with Indonesian language articles?

Despite the imbalanced dataset issue, the model performed exceptionally well in the claim verification task since there are only two classes in the dataset, sufficient for the IndoBERT to learn properly. However, despite being trained on a larger dataset and with more parameters, the IndoBERT model outperforms the GPT-3.5 model. We assume this is because the GPT 3.5 model is not well-trained with the Indonesian dataset, especially for the claim verification task dataset. Moreover, term-based classifiers such as Support Vector Machines, Logistic Regression, and Random Forest are unsuitable because they do not allow a cross-encoded input of two texts (claim and evidence). Furthermore, more labeled datasets are needed, particularly for the ”support” class, to improve this model's performance in this class.

Furthermore, since there is a massive amount of misinformation created and spread every day (especially in Indonesia), there is a need for tools or a workflow that can improve the fact-checkers work. We argue that the support of human fact-checkers by NLP tools for classification and verification can help them better analyze and classify the claims and evidence they collect. Our proposed workflow in Section 6.3 is a suggestion for the practical implementation of our work. This will hopefully speed up and improve the quality of the fact-checker's work.

The novelty of this study is the fine-tuning of the IndoBERT model using a dataset with multiple class labels. Some research has been conducted on the fine-tuning of the IndoBERT model for the classification of fake news, but the vast majority of this research employs binary classification with a significantly smaller dataset. In addition, relatively little research has been conducted on claim verification in the Indonesian language using the BERT model; this research will serve as a solid foundation for future

studies on claim verification tasks in the Indonesian language. In addition, there is a lack of research on applying the GPT model to the claim verification task, particularly with the Indonesian language dataset. This study may pave the way for future research on utilizing the GPT model for claim verification and other NLP-related tasks in the Indonesian language.

8.2 Limmitation

There are limitations to this research, including:

1. Token Limitation: The BERT model can only accept a maximum of 512 tokens, which may restrict the length of the input sequences.
2. High Cost: The usage of the GPT 3.5 model can be expensive, which could pose financial constraints for extensive experimentation.
3. Dataset Availability: There is a scarcity of large-sized and balanced datasets in the Indonesian language, specifically for multi-class fake news classification and claim verification tasks.
4. Lack of Neutral Class: The dataset lacks a neutral class, which prevents the model from being trained to identify neutral claims in the claim verification task.
5. Lack of Model Information: The GPT model does not provide detailed information regarding the specific datasets used for training, especially in the context of both fake news classification and claim verification tasks with the Indonesian language.

8.3 Future Work

The future work of this research will focus more on gathering more data, particularly from classes with very small sizes, and ensuring that we have a balanced data set with sufficient size to train. This will significantly improve the model's performance, and we will have a reliable model to use as a comparison tool to supplement the work of the human annotator. To overcome the token limitation, an alternative approach could be considered such as Longformer which could handle more than 512 tokens. By experimenting with a model such as Longformer, it becomes possible to incorporate longer sequences without extensive summarization. This approach mitigates the risk of data loss and its potential impact on the overall performance of the model.

Bibliography

[1] H. Ahmed, I. Traore, and S. Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In I. Traore, I. Woungang, and A. Awad, editors, *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, volume 10618 of *Lecture Notes in Computer Science*, pages 127–138, Cham, 2017. Springer.

[2] H. Ahmed, I. Traore, and S. Saad. Detecting opinion spams and fake news using text classification. *Journal of Security and Privacy*, 1(1), January/February 2018.

[3] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[4] João Pedro Baptista and Anabela Gradim. A working definition of fake news. *Encyclopedia*, 2(1), 2022.

[5] Bhavesh Bhatt. svm-c-gamma-hyperparameter. *Deepnote*, 2021. Accessed on July 4th, 2023.

[6] Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, 2019.

[7] Tom B Brown, Benjamin Mann, and Nick Ryder. Language models are few-shot learners, 2020.

[8] Jason Brownlee. *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery, 2020.

[9] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.

[10] Qian Chen, Xiaodan Zhu, Z Ling, Si Wei, and Hui Jiang. Enhancing and combining sequential and tree lstm for natural language inference. corr abs/1609.06038 (2016). *arXiv preprint arXiv:1609.06038*, 2016.

[11] Jiayi Clien, Yutao Wei, and Yukang Zou. Analysis of the relevance between title of product and search term. In *2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*, pages 172–176. IEEE, 2022.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Thomas Edgar and David Manz. *Research methods for cyber security*. Syngress, 2017.

[14] Johannes Fürnkranz. A study using n-gram features for text categorization. *Austrian Research Institute for Artifical Intelligence*, 3(1998):1–10, 1998.

[15] Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv preprint arXiv:1809.01479*, 2018.

[16] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anand Nayak, et al. Claimbuster: the first-ever end-to-end fact-checking system. In *Proceedings of the VLDB Endowment*, volume 10, pages 1945–1948, 2017.

[17] Sani Muhamad Isa, Gary Nico, and Mikhael Permana. Indobert for indonesian fake news detection. *ICIC Express Letters*, 16(3):289–297, 2022.

[18] Jaehwan Jeong. Data augmentation method for fact verification using gpt-3. *Stanford CS224N Custom Project*, 2023.

[19] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[20] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications*, 80(8):11765–11788, 2021.

[21] Z Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering*, volume 1099, page 012040. IOP Publishing, 2021.

[22] Diane M Korngiebel and Sean D Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *NPJ Digital Medicine*, 4(1):93, 2021.

[23] Fajri Koto, Jey Han Lau, and Timothy Baldwin. Liputan6: A large-scale indonesian dataset for text summarization. *arXiv preprint arXiv:2011.00679*, 2020.

[24] Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. *arXiv preprint arXiv:2011.00677*, 2020.

[25] K Latha. *Experiment and Evaluation in Information Retrieval Models*. Chapman and Hall/CRC, 2017.

[26] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics, 2019.

[27] Yunqian Ma and Haibo He. Imbalanced learning: foundations, algorithms, and applications. 2013.

[28] Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq, et al. Detecting fake news using machine learning and deep learning algorithms. In *2019 7th international conference on smart computing & communications (ICSCC)*, pages 1–5. IEEE, 2019.

[29] Marek Medved and Vıt Suchomel. Indonesian web corpus (idwac). *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (FAL), Faculty of Mathematics and Physics, Charles University*, 2017.

[30] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, 2016.

[31] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: was it preventable? In *Proceedings of the 2017 ACM on web science conference*, pages 235–239, 2017.

[32] Iman Nekooeimehr and Susana K Lai-Yuen. Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets. *Expert Systems with Applications*, 46:405–416, 2016.

[33] Yixin Nie, Haonan Chen, and Mohit Bansal. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866, 2019.

[34] Rasmus Kleis Nielsen and Lucas Graves. " news you don't believe": Audience perspectives on fake news. 2017.

[35] Peter Norvig. How to write a spelling corrector. http://norvig.com/spell-correct.html, 2011.

[36] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016.

[37] Aseem Patil. Word significance analysis in documents for information retrieval by lsa and tf-idf using kubeflow. In *Expert Clouds and Applications: Proceedings of ICOECA 2021*, pages 335–348. Springer, 2022.

[38] Alireza Rahnama, Sam Clark, and Seetharaman Sridhar. Machine learning for predicting occurrence of interphase precipitation in hsla steels. *Computational Materials Science*, 154:169–177, 2018.

[39] Farshid Rayhan, Sajid Ahmed, Asif Mahbub, Rafsan Jani, Swakkhar Shatabda, and Dewan Md Farid. Cusboost: Cluster-based under-sampling with boosting for imbalanced classification. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pages 1–5. IEEE, 2017.

[40] Harshita Sharma and Tinkle Jain. Neural network with nlp. 2021.

[41] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[42] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[43] Amir Soleimani, Christof Monz, and Marcel Worring. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 359–366. Springer, 2020.

[44] Dewi Soyusiawaty and Yahya Zakaria. Book data content similarity detector with cosine similarity (case study on digilib. uad. ac. id). In *2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, pages 1–6. IEEE, 2018.

[45] Statista. Countries with the most instagram users, 2023. Accessed on June 17, 2023.

[46] Statista. Smartphone users in indonesia, 2023. Accessed on June 17, 2023.

[47] Rayan Suryadikara, Suzan Verberne, Frank W Takes, S Stefan Conrad, and I Tiddi. False news classification and dissemination: the case of the 2019 indonesian presidential election. In *Proceedings of the CIKM 2020 Workshops, co-located with 29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 1–9. CEUR-WS. org, 2020.

[48] Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705, 2021.

[49] FZ Tala. The impact of stemming on information retrieval in bahasa indonesia. *Proc. CLIN, the Netherlands, 2003*, 2003.

[50] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining "fake news" a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.

[51] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[52] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[53] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893, 2021.

[54] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[55] Apurva Wani, Isha Joshi, Snehal Khandve, Vedangi Wagh, and Raviraj Joshi. Evaluating deep learning approaches for covid19 fake news detection. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1*, pages 153–163. Springer, 2021.

[56] Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. Purwarianti. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *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*, 2020.

[57] Cahya Wirawan. bert2gpt-indonesian-summarization. Hugging Face Model Hub, 2021.

[58] Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, 2018.

[59] 김애니. *Optimal Selection of Resampling Methods for Imbalanced Data with High Complexity*. PhD thesis, 연세대학교, 2021.

Appendix A

Appendix

A.1 Cont. Text Representation using TF-IDF

A.1.1 Impostor Content

Content that could be classified as an impostor is that information profiteer, a statement of an important and high-influence figure. Impostor content targeted not only an individual but also an institution by leveraging the institution's fame. This misinformation is typically spread through a video or an article summarizing a video and misleads the interpretation. Some articles also mention ”cina,” the Indonesian word for China. While there is a lockdown due to the pandemic in 2020, but there are still some Chinese workers who could come from mainland China. Some people then write a false article about it, implying that the government is allying with China and is more concerned with the Chinese people. This content refers to the government being unjust to the locals. While most Indonesians could not work due to the lockdown, Chinese workers continued to arrive from the Chinese mainland. The truth is that this worker is part of an Indonesian infrastructure project with the Chinese government. The Chinese government will provide funding and expertise to build infrastructure, but some workers must be imported from China. So the government is not in favor of Chinese workers, but this is part of the Indonesian infrastructure project agreement with the Chinese government. This problem explained why the words ”cina” and ”2020” appeared on one of the most influential words in this category, as shown in table A.1.

A.1.2 False Context

False context is content that is presented with incorrect narration and context. Typically, false context consists of a statement, photograph, or video depicting an event that is

Word	Score
video	2.798572182043894
cina	2.0551808872742097
2020	1.9007893086267738
demo	1.8565440175513954
lihat	1.7261227757211712
banjir	1.6527833964189238
narasi	1.6520345551026063
india	1.5754882087686881
peristiwa	1.4285737352513324
bakar	1.4238318494218223

Table A.1: Top 10 Most Important Words for the Impostor Content Class

taking place in a specific location, but contextually, this article is not based on actual facts. This type of false information is typically disseminated through social media with a misleading caption to the actual context. The most influential word in this category is shown in table A.2. This type of false information is frequently used to attack a ”bupati” or regent head. Another discovery revealed the word ”WhatsApp,” which refers to the messaging app WhatsApp, where this type of false information is also primarily spread through group chat. In addition, there are attempts to defraud individuals by requesting donations (”donasi”) through the fabricated content they have created.

Word	Score
akun	5.6500467458580586
bupati	4.094666467868563
profil	3.4667302656107557
whatsapp	2.8516304533333727
nomor	2.558209200773188
investasi	2.5477217832774683
edar	2.2719343842579334
kirim	2.148468313743546
mengatasnamakan	2.076140413825091
donasi	2.0749891410739427

Table A.2: Top 10 Most Important Words for the False Context Class

A.1.3 False Connection

The most common characteristic of this type of content is a mismatch between the title and description. This type of content is also referred to as clickbait. The content sometimes does not even discuss or discuss what the title claims. This type of content

Word	Score
gambar	3.063240501594392
sunting	3.059078898650748
olah	2.5012601035202198
judul	2.0435845953792264
rizieq	1.8308177921948645
layar	1.663441318441208
megawati	1.4441799360553345
habib	1.4169742903966651
asli	1.382880953209079
deh	1.3596440772802472

Table A.3: Top 10 Most Important Words for the False Connection Class

is typically created to increase the content's financial value or to increase its engagement. The most influential word in this type of fake news can be shown in table A.3.

A.1.4 Clarification

Word	Score
vaksin	2.7310934962671136
covid	2.3022402747202015
unggah	2.141521877218089
narasi	2.0294397373981603
klaim	1.8961893047398317
kandung	1.8803770231826868
buah	1.7747741267441444
ambil	1.6238465187766618
kali	1.519627997985014
jantung	1.5042954526035746

Table A.4: Top 10 Most Important Words for the Clarification Class

Clarification is a type of content that is created to clarify a certain topic or misinformation that spreading in the public. This type of content is usually created by a government official or institution or a credible news media. This type of content is created to prevent the spread of fake news and to raise awareness about the problem. As shown in table A.4, most data point relates to "vaksin," which translates to vaccine and covid. The Indonesian public is very concerned about these two issues at the outbreak's onset. There is a great deal of misinformation circulating on the internet that covid is a hoax or that the vaccine is ineffective and will harm our bodies rather than protect us from covid. Some individuals became so skeptical of covid and the vaccine due to this infor-

mation that some even became anti-covid vaccines. Through government officials and health institutions, the government strives to educate the public on this matter. Therefore, more people will be willing to get vaccinated, allowing this pandemic to be ended sooner.

A.1.5 True

This category contains the actual content. All content is created with the correct context between the article's title and its entire body of text. This content is intended to educate, inform, or raise awareness about a particular topic. As shown in table A.5, the feature importance assigned to this category's dataset does not accurately reflect the nature of the content. "foto" means photographs, "unggah" means to upload, "gambar" means picture, and "lihat" means to view. One of the reasons this does not reveal much about the nature of the category is that the dataset in this category is quite small, and more articles in this category could be scraped to gain more insight into the category of articles.

Word	Score
foto	1.9777252488308208
unggah	1.8859720456885618
gambar	1.7731836495325413
lihat	1.7304444318762129
cuit	1.4585446537144369
facebook	1.434970850334249
anies	1.2576789931000378
jomlo	1.1819383037201445
buah	1.1022878414455262
2022	1.087082975666492

Table A.5: Top 10 Most Important Words for the True Class

A.2 Prompts

Indonesian:

```
Anda akan diberikan klaim dan fakta yang terkait dengan berbagai
topik. Anda diminta untuk menganalisis klaim tersebut
berdasarkan fakta yang disediakan dan klasifikasikan sebagai
"didukung" atau "ditolak" saja.

Sebuah klaim diklasifikasikan sebagai "didukung" jika ada bukti
atau pernyataan dalam fakta yang secara langsung mendukung klaim
tersebut atau membuat klaim tersebut menjadi benar. Ini bisa
berupa penelitian, pernyataan dari orang-orang yang berwenang,
atau fakta yang diakui secara umum yang mendukung klaim.

Sebaliknya, klaim diklasifikasikan sebagai "ditolak" jika fakta
yang disediakan menunjukkan bahwa klaim tersebut tidak benar,
salah, atau tidak memiliki dasar yang valid. Fakta ini mungkin
juga termasuk pernyataan dari ahli, penelitian, atau informasi
umum yang menyangkal klaim tersebut.

Dalam kasus di mana fakta tidak langsung mendukung atau menolak
klaim, lakukan analisis yang mendalam dan logis berdasarkan
informasi yang disediakan dan pengetahuan yang sudah Anda miliki
untuk membuat keputusan. Jika tidak yakin, berikan penilaian
terbaik Anda.

Berikut adalah beberapa contoh:
Contoh 1:
Klaim: "mengonsumsi kulit pohon jambu mete bisa menetralisir
racun akibat gigitan ular"
Fakta: "Liputan6.com: Kulit pohon jambu mete tak menetralisir
racun ular. WHO sarankan hindari obat herbal, percayai antibisa.
Johan Marais juga menyatakan hal serupa. Gigitan ular berbisa
harus ke rumah sakit dengan stok antibisa."
Klasifikasi Klaim: "ditolak"

Contoh 2:
```

Klaim: "rusia keluar dari pbb"
Fakta: "Akun TikTok 'amrika' memposting video menyesatkan tentang Rusia keluar dari PBB. Rusia hanya keluar dari Dewan HAM PBB, bukan PBB secara keseluruhan"
Klasifikasi klaim: "ditolak"

Contoh 3:
Klaim: "pandemi corona menteri pendidikan dan kebudayaan dana bos bisa untuk beli kuota internet siswa"
Fakta: "Menteri Pendidikan dan Kebudayaan, Nadiem Makarim, katakan dana bantuan operasional sekolah (BOS) bisa beli kuota internet guru dan siswa."
Klasifikasi Klaim: "didukung"

Contoh 4:
Klaim: "Vaksin COVID-19 efektif mencegah penyebaran virus"
Fakta: "Berbagai studi dan data menunjukkan bahwa vaksin COVID-19 telah berkontribusi besar dalam menurunkan penyebaran virus dan keparahan gejala bagi mereka yang terinfeksi."
Klasifikasi Klaim: "didukung"

Sekarang, tolong klasifikasikan klaim berikut ini berdasarkan fakta yang ada:
Klaim: "{}"
Fakta: "{}"

Pertimbangkan pertanyaan berikut dalam analisis Anda:
1. Apakah klaim ini secara langsung didukung atau ditolak oleh fakta yang disediakan?
2. Jika tidak, bagaimana Anda dapat menganalisis klaim dan fakta ini secara kritis dan logis untuk membuat keputusan?

Pada akhirnya, klasifikasikan klaim ini sebagai "didukung" atau "ditolak".

Klasifikasi Klaim (didukung atau ditolak):

English

Translated to English for the purpose of this paper.

```
You will be presented with claims and facts related to various
topics. You are asked to analyze these claims based on the provided
facts and classify them as either "supported" or "rejected".

A claim is classified as "supported" if there is evidence or
statements in the facts that directly support the claim or make the
claim true. This could be research, statements from authoritative
figures, or generally acknowledged facts supporting the claim.

Conversely, a claim is classified as "rejected" if the provided
facts indicate that the claim is not true, incorrect, or lacks a
valid basis. These facts may also include statements from experts,
research, or general information that contradicts the claim.

In cases where the facts do not directly support or reject the
claim, conduct a thorough and logical analysis based on the provided
information and the knowledge you already have to make a decision.
If unsure, provide your best judgement.

Here are a few examples:
Example 1:
Claim: "Consuming cashew tree bark can neutralize poison from snake
bites."
Fact: "Liputan6.com: Cashew tree bark does not neutralize snake
venom. WHO suggests avoiding herbal remedies, trust antivenom. Johan
Marais also stated the same. Venomous snake bites need to be taken
to the hospital with antivenom stocks."
Claim Classification: "rejected"

Example 2:
Claim: "Russia has withdrawn from the United Nations"
Fact: "The 'amrika' TikTok account posted a misleading video about
Russia withdrawing from the UN. Russia only withdrew from the UN
Human Rights Council, not the UN as a whole."
Claim Classification: "rejected"
```

Example 3:
Claim: "During the corona pandemic, the Minister of Education and Culture said school operational assistance funds can be used to buy internet quotas for students."
Fact: "The Minister of Education and Culture, Nadiem Makarim, said that school operational assistance funds (BOS) can be used to buy internet quotas for teachers and students."
Claim Classification: "supported"

Example 4:
Claim: "COVID-19 vaccine is effective in preventing virus transmission."
Fact: "Various studies and data show that the COVID-19 vaccine has contributed greatly in reducing virus transmission and the severity of symptoms for those infected."
Claim Classification: "supported"

Now, please classify the following claim based on the available facts:
Claim: "{}"
Fact: "{}"

Consider the following questions in your analysis:

Is this claim directly supported or rejected by the provided facts? If not, how can you critically and logically analyze this claim and these facts to make a decision?
In the end, classify this claim as "supported" or "rejected".

Claim Classification (supported or rejected):