Indonesian Hoax Identification on Tweets Using Doc2Vec

1st Titi Widaretna
School of Computing
Telkom University
Bandung, Indonesia
titiwidaretna@student.telkomuniversity.ac.id

2nd Jimmy Tirtawangsa
School of Computing
Telkom University
Bandung, Indonesia
jimmytirtawangsa@telkomuniversity.ac.id

3rd Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

***Abstract*—In this paper, we present our work on hoax detection on a collection of Tweets. We tackle the hoax detection as a text classification problem, with Doc2Vec as the text representation method and SVM as the classifier. We collected and annotated 5000 Tweets that consist of 2500 hoax Tweets and 2500 truth Tweets. The experimental results show that the accuracy of our proposed hoax detection on Tweets is 93.4%.**

***Index Terms*—Hoax Detection, Natural Language Preprocessing**

I. MOTIVATION

Nowadays, the ease of accessing news on various platforms makes some people abuse it for bad things. The Ministry of Communication and Information in 2017 stated that there were 132,7 million people in Indonesia who used the internet and social media [1]. One of the negative impacts of this is the number of hoaxes that are spread on various platforms such as social media, news articles, and broadcast messages such as Whatsapp and Telegram.

During this pandemic, the number of hoaxes has also increased. This is based on a statement from the Director of Information Management of the Ministry of Communication and Informatics [2] if from 23rd January 2020 to 15th June 2020 there were 850 hoaxes related to COVID19 circulating through various platforms. Until 23 September 2020, the number of hoaxes related to COVID19 had increased to reach 1984 [3].

With the support of technology and various kinds of media to spread information quickly, hoax can spread very quickly. As a result of this, year by year the level of public trust in the news platform has continuously declined. Data obtained from Edelman Trust Barometer Global Report in 2018 trust in every general news source and information on traditional media increased by 59% while on social media platforms and search engines it decreased to 51%. It has an impact on the public's trust in the social media platforms in Indonesia that decreased 4% from 71% in 2017 to 67% in 2018.

In research about "Wabah Hoax Nasional *National Hoax Outbreak*" [5], social media is the highest hoax news distribution channel, amounting to 87.5%, chat applications are in the second place with a percentage of 67%. While newspapers, websites, email, television, and radio have a percentage less than 30%. Research on hoax [6] with the title "Spread of Hoax in Social Media" concludes that Twitter is one of the more effective media to spread the news from person to person at a speed comparable to conventional mass media.

Twitter is one of the most widely used social media today. Twitter users share more information using text only. The amount of text that can be shared on Twitter is limited to 280 characters. Sharing only text can be a problem as a Twitter user shares information about an event without including photo or video evidence, and it will be difficult for other users to trust the validity of provided information.

The definition of hoax in this research is motivated by a statement in Asya [9] that states that hoax is a statement of news that contradicts its semantic truth. The truth of the news serves to guarantee the reader's trust because the truth is one of the factors in determining the credibility level of the news [10]. Besides, semantics is also related to errors in a language where news items must have corresponding meaning in their context. This is indicated by the existence of references [11].

According to research [12] on Language Style Context Analysis, especially in hoax news, it is stated that in the case of hoax the meaning or acceptance of each individual for information will be different. This depends on the environment, knowledge, psychology, or experience of each individual. Context is the part of a text or sentence that surrounds a particular word or passage and determines the meaning of the text, so it is an important part of being able to understand the meaning of information [13]. Context is also related to semantics because to find the context, the process of deriving the semantic content of the words in a sentence is carried out and arranging them according to the syntactic structure of the sentence [14].

Doc2Vec computes a feature vector for every document in the corpus and can reconstruct the semantic from an incomplete paragraph [15]. This vector representation has the advantage of capturing the semantics, namely the meaning, of the input text. This means that texts that have the same meaning or context will be closer to each other in vector space.

In this work, we perform hoax detection as a text classification task. We represent the input text using Doc2Vec method and perform the classification using SVM method. Doc2Vec produces a semantic vector representation from the tweet resulting in some meaningful similarities between tweets. Semantic vectors created by Doc2Vec should be sufficient to determine the context from the tweet. If the resulting context matches other tweets in the semantic vector model, SVM should be sufficient to classify the original tweet as a hoax or truth.

II. RELATED WORK

Several previous research on hoax has been conducted, including Rasywir and Purwarianti [16] detected hoax in news articles using 220 articles of data. The method used is feature selection such as Information Gain, Mutual Information, Chi Square, Term Frequency, and TF IDF and classification using SVM, Naive Bayes, and C 4.5. The best results were obtained from union operation of Information Gain and Mutual Information with Naive Bayes with an accuracy of 91.36%.

Prasetijo [17] in his research on hoax detection using 200 website news data proposed TF IDF combined with SVM, SGD, and SGD Modified Huber. Classification using Modified-Huber SGD combined with TF IDF has better accuracy than using other classification methods, namely 86%, where this accuracy is 4% better than SVM.

Research from Afriza [18] conducted hoax detection using 600 data on website news, broadcast messages, and social media proposed feature selection using TF IDF and combined with Rocchio and Multinomial Naive Bayes. The result showed that Rocchio has better accuracy than Multinomial Naive Bayes with 85.3%.

Wardani [19], proposed an analysis of hoax news by checking the characteristics of news writing. The data used is news content obtained from the turnbackhoax.id site. The factors used to determine hoaxes are the manipulation of speech acts such as the use of words that describe positive or negative emotions and the use of capital letters in certain words to highlight the essence of the news, and manipulation of punctuation such as excessive use of punctuation and letters (e.g. i ' I'm sooooo haappyy !!!!). The results of this study indicate that the above factors can be used to detect hoax.

From all previous researches, the data used is different from the data we use in this research. Afriza uses data from social media Facebook. Meanwhile, we focus only on hoax detection on Twitter. The selection feature used is frequency-based using TF IDF. This method is not suitable for Twitter data where the number of texts is relatively small.

Because hoax is related to differences in the meaning of information, this study proposed to change the method from the frequency-based approach [16][17][18] and writing style-based approach [19] to a method with a semantic approach to be able to capture the context of information from a tweet, so the system will check the meaning of each information in the tweet and can determine the tweet as hoax or truth.

III. PROPOSED RESEARCH

The purpose of the proposed method is to detect hoax by checking the context of the tweet. Checking the context of the tweet will increase accuracy in detecting hoax. So we used Doc2Vec for feature selection which this process serves to construct a semantic vector representation that captures context. From this process, the program will get the context of the tweet and the semantic vector generated will be used for the classification process. In the classification process, we use the Support Vector Machine (SVM) as a classifier so, from the method we propose, we can determine whether a tweet is a hoax or truth. We also do some initial processes to clean the dataset through the Preprocessing stage. Lowercase dataset, remove punctuation marks, remove links, and remove stopwords. The purpose of this process is to make the data the same. Figure 1 is a diagram process of the proposed method.

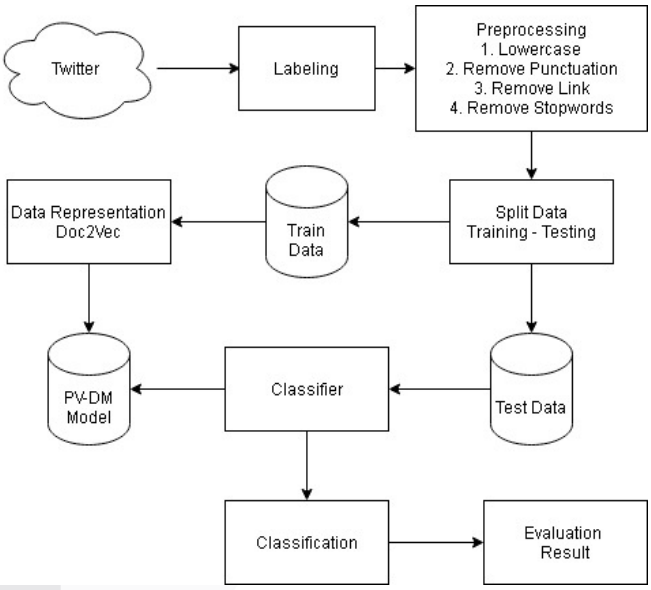

Fig 1. Diagram Process of Proposed Method

Explanation of the diagram process above:

3.1. Preprocessing

This section consists of:

3.2.1 Lowercase

This is the technique to convert all text from a dataset to lowercase (e.g. *jutaan produksi rokok yang terpapar covid-19 beredar luas di masyarakat* **millions of cigarette production exposed to covid-19 are circulating widely in society **)

3.2.2 Punctuation Removal

This process is the second stage in preprocessing. After going through the lowercase stage, the program will remove punctuations that are contained in the text (e.g.

jutaan produksi rokok yang terpapar covid19 beredar luas di masyarakat *millions of cigarette production exposed to covid19 are circulating widely in society*)

3.2.3 Stopwords Removal

The corpus used in the stopwords removal stage consists of 2 corpora, namely corpus from Sastrawi and Afriza [18]. In Afriza's previous research, she combined the stop words from Tala's research [20] with the words she collected. The difference between these two corpora is that the Sastrawi corpus only consists of 127 words and contains common words such as sudah, dan, untuk *already, and, for*. while the Afriza corpus has 750 words and contains more detailed common words. For example in the Sastrawi corpus there is the word sudah *already*, while in the Afriza corpus there is the word sudah, sudahlah *already, never mind*.

3.2. Vector Semantic using Doc2Vec

The idea of vector semantics is to represent a word as a point in some multidimensional semantic space. Vectors for representing words are generally called embedding because the word is embedded in a particular vector space. Vector semantic models are also extremely practical because they can be learned automatically from text without any complex labeling or supervision. It can be used to represent the meaning of words, by associating each word with a vector. As a result of these advantages, vector models of meaning are now the standard way to represent the meaning of words in NLP [21]. Doc2Vec is general and applicable to texts of any length: sentences, paragraphs, and documents. The input is not a word but a document token. It is capable of constructing representations of input sequences of variable length. Doc2Vec has two models that are Distributed Memory Model (PV-DM) and Distributed Bag of Words (PV-DBOW) [15]. The output of Doc2Vec is a vector representation containing context in the form of a matrix. In this research, every word in the tweet including each tweet itself is mapped to vectors.

3.3. Support Vector Machines

Support Vector Machines (SVM) was introduced by Vapnik as a kernel-based machine learning model for classification and regression tasks. The extraordinary generalization capability of SVM, along with its optimal solution and its discriminative power, has attracted the attention of data mining, pattern recognition, and machine learning communities in the last years [22].

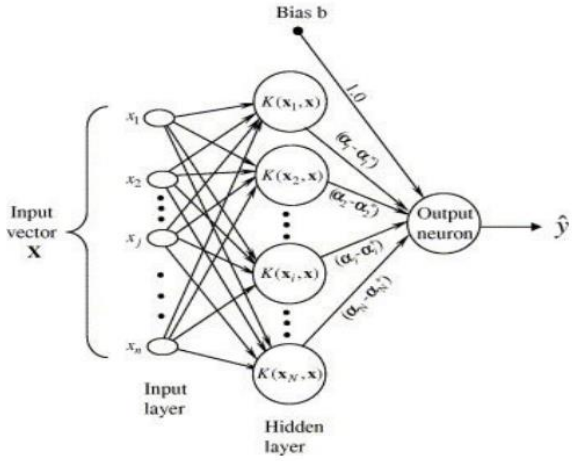

Fig.2. The Architecture of SVM [23]

SVM is highly accurate, owing to its ability to model complex nonlinear decision boundaries. They are much less prone to overfitting than other methods. The support vectors found also provide a compact description of the learned model. SVM can be used for numeric prediction as well as classification [24].

3.4. Data

The data that is used in this research is Indonesia hoax from Twitter collected manually from the period 2012 to 2020 as many as 5000 tweets consisting of 2500 tweets with a hoax class, and 2500 tweets with a truth class. To label the data, this is done by verifying the tweet data and comparing it with news articles from trusted sources such as turnbackhoax.id and liputan6.com. We double-check this labeling process to make sure there are no wrong labels.

Before entering the experimental stage, we try to investigate the dataset so we can determine the value for each parameter that we will use in this research. As already mentioned, Twitter limits the number of characters in each tweet, which is only 280 characters. From a total of 5000 tweets that we collected, in general, tweets consist of 20 words, both for hoax tweets and not. The patterns of writing hoax tweets and truth tweets that we found in this observation process are:

1. There are missing words from hoax tweets when compared to actual tweets.
2. The use of capital letters in certain words in hoax tweets.
3. Excessive use of punctuation in hoax tweets

An example can be seen in Table I. Regarding the closure of tourism objects. In the example of the first tweet which was a hoax, they only mentioned about Pangandaran beach tourism objects will be closed, and in the second tweet, it mentioned that information regarding the closure of Pangandaran tourist objects is incorrect. Even though the first tweet included a link as a reference regarding the news, it did not contain information about the truth of the news, so the tweet was considered a hoax.

In the third tweet, the main information of the tweet is about Jackie Chan converting to Islam. This is a hoax. The tweet wrote the headline in the form of a hoax using capital

letters while the other text did not. There he also asked the word *Benarkah???* Really??? and used 3 question marks at once. The fourth tweet explains the existence of a banner with the words "Larangan Bersholawat" where the user thinks that Muslims should not pray (Larangan in Indonesian means prohibition). The word Larangan referred to in the banner has the meaning of an area in Banten, Indonesia. Where the context of the banner is to inform and invite Muslims in the Larangan area to pray together. If other users who read this tweet do not know the Larangan area, then that user will conclude that the context of the tweet is that Muslims should not perform prayers.
Below are examples of data.

TABLE I.
EXAMPLE OF DATA

Tweet	Class
CEK FAKTA: Objek Wisata Pantai Pangandaran akan Ditutup https://goo.gl/fb/QhmFKb #MDK *CHECK THE FACTS: Pangandaran Beach Tourism Objects will be Closed https://goo.gl/fb/QhmFKb #MDK*	Hoax
Kabar Objek Wisata Akan Ditutup, Pjs Bupati Pangandaran: Tidak Benar https://ruber.id/berita-mengenai-akan-ditutupnya-objek-wisata-pangandaran-adalah-tidak-benar/ lewat @ruber.id #pangandaran #jabar #news #Senin *News Tourism Objects Will Be Closed, Acting Regent Pangandaran: Not True https://ruber.id/berita-mengenai-ditutupnya-objek-wisata-pangandaran-adalah-tidak-benar/ via @ ruber.id #pangandaran #jabar # news #Senin*	Truth
"JACKIE CHAN" MASUK ISLAM, Benarkah??? Beredarnya kabar aktor film laga dan aksi terkenal, Jackie Chan, memeluk... http://fb.me/5nMNUboiF *""JACKIE CHAN" ENTERS ISLAM, Really ??? Rumor has it that famous action and action film actor Jackie Chan hugs ... http://fb.me/5nMNUboiF"*	Hoax
"Udah mulai terang2an larangan bersholawat" Di gambar itu, terdapat tulisan "LARANGAN BERSHOLAWAT" dengan warna kuning. "Have started to LARANGAN BERSHOLAWAT" In the picture, there is the inscription of "LARANGAN BERSHOLAWAT" in yellow color.	Hoax

3.5. Experimental Setting

In this research, we performed the classification under different several settings to analyze the effect of each process and parameter on classification results. The process and parameters are defined as follows:

3.5.1 Doc2Vec Parameter

The following are the parameters we use in this research.

1. DM : PV-DM (Distributed Model) is a model that makes the token paragraph act as a memory that remembers what is missing from the current context - or the topic of the paragraph [15]. Another way is PV-DBOW (Distributed Bag of Words) that ignores the context words in the input but forces the model to predict words randomly sampled from the paragraph in the output [15]. Because of the observation on the dataset, the results are that if one of the differences between hoax and truth tweets is that there are some missing words, then we use PV-DM to create a model so the model can check for missing words tweet and generate context that matches that tweet.
2. Min_count: 1. It means that Doc2Vec will ignore all words with a total frequency lower than 1[25]. Because the results of the observation show if in general, the tweet consists of only 20 words, we set a value of 1 for the min_count so not many words are deleted. So the original arrangement of the tweets can be preserved.
3. Window: The function of the window is to set the maximum distance between the current and predicted word within a sentence [25]. A good guess of window size in many applications is between 5 and 12. In IMDB, varying the window sizes between 5 and 12 causes the error rate to fluctuate 0.7% [15]. We use a range of 1 to 10 because the number of words is only 20 and the possibility of those 20 words has a reference link as shown in Table I. Because what we need at this stage is the information conveyed in the tweet, we don't need a referral link. This is done because we do not check the contents of the reference link, we only focus on the content of the tweet. So to find the right window value to generate relevant tweets, we tried several window values with values smaller than a total of 20 words.
4. Vector_size: The function of this parameter is for the dimensionality of the feature vectors [25]. Because the data we use is relatively short, around 20 words, we try to experiment to find the best vector_space value to help the search for relevant tweets. Here we try the vector_size values 5, 10, and 20.
5. Epochs: 45. This parameter is used to determine the number of iterations. The default is 10 for Doc2Vec. The number of times to train the new document. Larger values take more time but may improve quality [25]. In this research, we tried to use 45 epochs.

3.5.2 Support Vector Machine Parameter

The following parameters that used in this research.

1. Kernel: RBF (Radial Basis Function). The function of the kernel is to take data as input and transform it into the required form.
2. C value: we tried to use the values 1 to 20 to see the impact that the value C had on the classification results.

3.6. Experimental Result

A confusion matrix summarizes the classification performance of a classifier for some test data. It is a two-dimensional matrix, indexed in one dimension by the true class of an object and in the other by the class that the classifier assigns [26]. To calculate the confusion matrix, the formula is defined as follows:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = 2\frac{Precision\ x\ Recall}{Precision\ x\ Recall} \quad (4)$$

IV. TESTING AND RESULTS ANALYSIS

4.1. Implementation with Frequency-Based Approach Method

Before proposing a semantic-based approach to get the context, we tried to implement the method from Afriza's research [18]. This study was chosen because the data used is a combination of various news sources, so it must contain informal and formal words like the Twitter data we use. Meanwhile, other studies [16][17] only use 1 type of data, namely news articles. Below is a table showing the results of implementing the Afriza [18] method using our dataset with the same amount of data.

TABLE II.
FREQUENCY-BASED APPROACH WITH TWITTER DATA

Accuracy	Precision	Recall
65%	65.62%	67.74%

When implemented on Twitter data, the accuracy decreased because from the analysis results there were several dominant words. A dominant word is a word that has a higher frequency than other words in a set of documents. If dominant words exist mostly in hoax documents, the probability to be classified as a hoax. Some original tweets that must be classified as truth but contain the dominant word in the hoax document cause the original tweet to be classified as a hoax, and vice versa.

4.2. Vector Size Parameter in Doc2Vec

In [27] the determination of the value of the vector size parameter is based on the dataset used which consists of 2 to 3 paragraphs. So it is preferable to use a small value and the size vector value used is 10. Besides, in [28] also uses less data so it is appropriate to test the value of the smaller size vector. So an experiment was carried out on the vector size value, namely the vector size $\in$ {5, 10, 15, 20, 25, 30, 40, 50, 70, 100, 130, 180, 210, 240, 270, 300}. Determining the vector value for the word dimension is very important during the training process. The determination of the vector size is determined based on the corpus statistics and determines the lower bound for the vector size to be used. Based on the results of research by Moradi [27] we will use the value of 10 as the lower bound of the vector size and perform several experiments such as that of Foxcroft [28]. We tried the vector size experiment with a value of 5 (below the lower bound) and 20 (above the lower bound) to see the results of using different vector size values. We will discuss experiments with window values in section 4.2. Table III shows the best accuracy of each vector size value using window size 10.

TABLE III.
RESULT OF VECTOR SIZE VALUE

Vector Size	Accuracy	Precision	Recall	F1 Score
5	91.2%	91.3%	91.2%	91.2%
10	92.4%	92.4%	92.4%	92.4%
20	93.4%	93.5%	93.4%	93.4%

The results of this research indicate that vector size higher than or equal to the limit gets better results than those under it [29]. The experimental results in Table III show that the vector size above the lower bound value has the best accuracy. Likewise, the vector size value which becomes the lower bound has a better value when compared to vector size 5 which is a value below the lower bound.

The results of the analysis in this experiment show that the vector size 5 value cannot learn all the topics in the training data so there are more prediction errors than the vector sizes 10 and 20. For example, in the training data, 122 tweets are containing the word "Jokowi" and discuss more than 5 topics. In the testing data, 56 tweets were containing the word "Jokowi" and consisted of more than 5 topics as well, so when the experiment used a vector size 5, Doc2Vec could not capture all the contexts that discussed topics related to Jokowi and caused prediction errors. Meanwhile, when the vector size is increased to 10 and 20, the prediction error rate for tweets containing the word "Jokowi" decreases. Prediction errors that exist in experiments with a vector size of 10 related to the topic of Jokowi can also be predicted well when the vector size value becomes 20.

4.3. Window Parameter in Doc2Vec

The experimental results in Table III show that a dimension of 20 can produce the best accuracy. So in this section, we try to use a vector size parameter value of 20 and several window values as mentioned in section 3.5 on Experimental Setting. Experiments with different window values refer to the experiments conducted by Foxcroft [28]. The dataset used in his research uses a smaller number of words than our study, so we try to experiment with window values ranging from 1 to 10. While in [28] we use windows with a range of 1 to 6.

TABLE IV.
RESULT OF WINDOW VALUE

Window	Accuracy	Precision	Recall	F1 Score
10	93.4%	93.5%	93.4%	93.4%

Because the analysis results show that not all tweets have the same number of words, then we conclude that a window with a value of 10 can produce the best accuracy in the case of hoax detection with Twitter data. For example, in training data, tweets are consisting of 20 words, while testing data on the same topic only consists of 5 words. If the window is used a little, it will cause a difference in context when testing data checks the vector from the training data.

4.4. C Parameter in Support Vector Machine

After experimenting with parameters in Doc2Vec, the next step is to see the distribution of the C value on SVM on accuracy. In this step, we try to use the value of C in the range of 1 to 20. Below are the results of the experiment with the

value C.

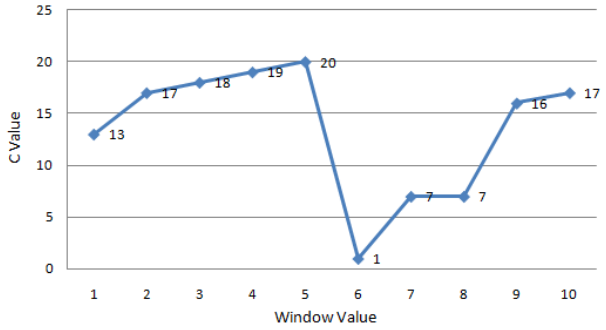

Fig. 3. Distribution of C Value

C is the penalty parameter that controls the width of the soft margin and is related to determining the training error [30]. It means that if the C value is high, the model can not generalize well when predicting new data. Fig.3 above shows the distribution of C values to get the maximum accuracy of each experiment in Table IV about experiments with window value.

V. CONCLUSIONS

In this research, a hoax detection system was successfully built by using context information that was generated from Doc2Vec and classified by SVM. The accuracy of hoax detection using this system can reach up to 93.4% and the precision up to 93.5%. This result has a significant improvement against previous methods [16][17][18]. Although doc2vec does not explicitly recognize a hoax based on characteristics of hoax, it can detect many hoax as classified by Wardani [19].

ACKNOWLEDGEMENT

We would like to thank all reviewers who have provided helpful suggestions and input to make this paper have better quality.

REFERENCES

[1] Ada 800.000 Situs Penyebar Hoax di Indonesia. : https://kominfo.go.id/content/detail/12008/ada-800000-situs-penyebar-hoax-di-indonesia/0/sorotan_media

[2] Kominfo: Hingga Juni Terdapat 850 Hoax Terkait COVID-19. : https://kominfo.go.id/content/detail/27755/kominfo-hingga-juni-terdapat-850-hoaks-terkait-covid-19/0/sorotan_media

[3] 6 Hoaks di Indonesia Jadi Sorotan Media Asing Sepanjang Tahun 2020.https://www.liputan6.com/cek-fakta/read/4366072/6-hoaks-di-indonesia-jadi-sorotan-media-asing-sepanjang-tahun-2020

[4] 2018 Edelman Trust Barometer. : https://www.edelman.com/sites/g/files/aatuss191/files/2018-10/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf

[5] Hasil Survey Wabah Hoax Nasional 2019. https://mastel.id/hasil-survey-wabah-hoax-nasional-2019/

[6] Situngkir, Hokky. Spread of Hoax in Social Media. Bandung Fe Institute in MPRA No. 30674. https://mpra.ub.uni-muenchen.de/30674/

[7] Hoax Distribution Through Digital Platforms in Indonesia 2018. https://dailysocial.id/post/laporan-dailysocial-distribusi-hoax-di-media-sosial-2018

[8] Kompas. Hoaks di Twitter Lebih Gampang Menyebar Daripada Klarifikasi, Mengapa? https://tekno.kompas.com/read/2018/03/14/08040367/hoaks-di-twitter-lebih-gampang-menyebar-dari-klarifikasi-mengapa?page=all. 2018

[8] Alcott, et al. Trends in the Diffusion of Misinformation in Social Media. Stanford University. 2018.

[9] Asya, Akopova. Linguistic Manipulation: Definition and Types". *International* Journal of Cognitive Research in Science, Engineering and Education, Volume 1, Nomor 2. 2013.

[10] Juditha, Christiany. Hoax Communication Interactivity in Social Media and Anticipation. Puslitbang Aplikasi Informatika dan Informasi Komunikasi Publik Kementerian Komunikasi dan Informatika RI. 2018

[11] Parera, J. D. Teori Semantik. Jakarta: Erlangga. 2004.

[12] Kartika, D., et al. Analisis Konteks Gaya Bahasa Berita Hoax Debat Capres di Media Sosial Facebook. Universitas Muhammadiyah Surakarta. 2019.

[13] Saifudin, A. Konteks Dalam Studi Linguistik Pragmatik. LITE: Jurnal Bahasa, Sastra, dan Budaya. Vol 14. No 2. 2018. http://publikasi.dinus.ac.id/index.php/lite/article/view/2323/1462

[14] Stanley, J. Context in Semantics. Forthcoming in Contextualism, Gerhard Preyer, ed. (OUP).https://ling.auf.net/lingbuzz/000024/current.doc?_s=j5feWVndjcBFDC50#:~:text=The%20semantic%20content%20of%20a%20sentence%20relative%20to%20a%20context,syntactic%20structure%20of%20that%20sentence.

[15] Le, Quoc., Mikolov, Thomas. Distributed Representation of Sentences and Documents. International Conference on Machine Learning. 2014.

[16] Rasywir, E., Purwarianti, A. Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. Jurnal Cybermatika Vol 3, No 2. 2015

[17] Prasetijo, A.B., et al. Hoax Detection System on Indonesian News Sites Based on Text Classification Using SVM and SGD. 2017

[18] Afriza, A., Adisantoso. Klasifikasi Berita Bohong Menggunakan Algoritma Rocchio. Institut Pertanian Bogor. 2018

[19] Wardani, MMS. Manipulasi Bahasa Dalam Teori Kabar Bohong (Hoax). Jurnal Kebudayaan Ilmiah SINTESIS Vol 11, No 2. 2017.

[20] Tala, F Z. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Institute for Logic, Language and Computation Universeit Van Amstredam. 2003

[21] Jurafsky, Dan., Martin J.H. Speech and Language Preprocessing. 3rd Edition. 2019. https://web.stanford.edu/~jurafsky/slp3/

[22] Cervantes, Jair., et al. A Comprehensive Survey on Support Vector Machine Classification: Application, Challenge, and Trends. Neurocomputing. 2019.

[23] Farid, Nahla., et al. A Comparative Analysis for Support Vector Machines for Stroke Patients. Recent Advances in Informatics Science. 2013

[24] Han, Jiawei., Kamber, Michelin., Pei, Jian. Data Mining Concepts and Technique. Third Edition. 2012. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[25] Rehurek, Radim. Doc2Vec Paragraph Embedding. https://radimrehurek.com/gensim/models/doc2vec.html#:~:text=min_count%20(int%2C%20optional)%20%E2%80%93,then%20prune%20the%20infrequent%20ones.

[26] Ting, K. M. Confusion Matrix. Encyclopedia of Machine Learning and Data Mining. 2017.

[27] Moradi, M., et al. A Cross-Modality Neural Network Transform for Semi-Automatic Medical Image Annotation. Springer International

Publishing AG. 2016.

[28] Foxcroft, J., et al. Name2Vec: Personal Names Embedding. Springer Nature Switzerland AG. 2019.

[29] Patel K., Bhattacharyya, B. Towards Lower Bounds on Number of Dimensions for Word Embeddings. Proceedings of The 8th International Joint Conference on Natural Language Processing. 2017

[30] Hearst, M.A. Support Vector Machines. IEEE Intelligent Systems. 1998.