Group chat analysis of hoax detection during the covid-19 pandemic using the k nearest neighbors algorithm and massive text processing

K Umam*

Department of Computer Science, Faculty of Mathematics and Natural Science, Semarang State University, Indonesia.

*Corresponding author: khoerulumam@students.unnes.ac.id

Abstract. Group chat is the most widely used choice of various short information. Besides being easy to send messages, sharing short messages in group chat is considered effective compared to sending massively to several users. The ease of sending short messages in group chat is often used as the spread of fake news and untrue news or hoaxes, especially during the Covid-19 pandemic, the information shared can be easily shared by anyone without seeing a valid source. The dissemination of information related to Covid-19 without a clear source is a dangerous act, because it can lead users into false information and endanger themselves. Fake message detectors have not been widely implemented in instant message applications, for this reason, there is a need for a detector and a machine to analyze activities in group chat and see whether the message is included in content containing fake news or not. If a group chat has a lot of fake news, you can be sure that the group chat is not good to follow. The use of the K-Nearest Neighbors algorithm is considered quite effective in classifying an object, the results can be determined whether it is included in fake news, miss-information news, or true news. The process of processing messages is carried out by the massive text processing method because the characteristics of the text are different for each user so that text processing can be maximized for later classification. As a result, group chat can be analyzed based on active time, user messages, user activity, and messages sent between users.

1. Introduction

The phenomenon of spreading false information greatly affects people's social life patterns because the spread of false information can change people's mindsets, opinions, and how people make decisions. Information on the internet started to influence the wider community several years ago [1].

A hoax can simply be interpreted as false or untrue information. The term hoax becomes bigger when the use of technology, especially the internet, is getting wider too [15]. The ease with which users share information is one of the reasons why hoax information spreads faster. The main purpose of spreading hoax information is to lead public opinion and manipulate the situation through threats and various fraudulent artificial information [4].

The spread of false information is mostly through social media. Easy access to social media and the absence of supervision, allows users to disseminate any information without having to include the source. One of the media that often spreads false information is Group Chat. In Group Chat, users can freely send text, news, links, videos, images, and documents without having to include a source. In
addition, there are message forward and broadcast message features that allow users to share messages faster and send them to many other users at one time [5].

K Nearest Neighbors (KNN) is an algorithm that is widely used in data classification, data mining, and machine learning because of its simple implementation and can be measured easily. In simple terms, determining the results of the KNN process is by looking at how close the values of the parameters are inputted with the dataset that has been provided is based on a certain k value [14][6].

The use of KNN in classifying data has been widely used. One of the uses of the KNN is in detecting spam messages in e-mail with various methods resulting in different levels of accuracy, such as the extended KNN [3], Spearman Correlation as a method for calculating distance data [11] and combines with other algorithms such as the Recurrent Neural Network (RNN) [13].

Apart from detecting spam e-mails, the KNN algorithm is used to detect comments on social media using the Support Vector Machine [2]. Whereas in research to detect hoaxes, KNN was used together with other methods such as TF/TDM which resulted in an accuracy value between 74.1% to 83.6% [14].

In detecting hoaxes, in addition to KNN, many other algorithms have been developed and can be used as potential in optimizing final results, such as hoax detection with Naïve Bayes optimization [9], using N-Gram analysis [1], and Stochastic Gradient Descent [8].

2. Methods
2.1. Research procedure

The research procedure carried out in this study is shown in Figure 1. There are two main data that will be processed, namely the dataset as a comparison of hoax information taken during the pandemic and last updated in October 2020, and the second message data from the group for the period February to October 2020.

![Figure 1. Research Procedure](image)

2.2. Data collection

Data collection is done by exporting the message data that has been provided by the group chat application service provider. The data is stored in text format with a description of the time, the message sender's contact, message, and other information such as media delivery, adding members to the group, and information about group members leaving. Meanwhile, the comparative dataset for hoax information is taken from the public dataset that has been collected.

2.3. Pre-Processing data

2.3.1. Cleaning data. Data cleaning is the process of deleting data that is not needed in the research process. Because in this study it focuses on text processing, data in the form of media is not included in the next process. Some of the explanatory text contained in the data has also been removed so that it can focus on the hoax messages and data used.
2.3.2. *Data normalization*. Data normalization is done to balance data so that it can be processed properly without more data errors. The details of the processed data are shown in Table 1.

| Data details                  | Total |
|-------------------------------|-------|
| Hoax data                     | 500   |
| Message                       | 4519  |
| The media contained in the message | 1435  |

2.3.3. *Omitting non-standard words*. Non-standard words are often used in daily communication activities. The word cannot be processed properly because it has a biased and ambiguous meaning. In the process, apart from eliminating non-standard words, analysis of the use of good words is carried out so that the data can be processed optimally [7].

2.4. *Processing data*

2.4.1. *Massive text processing*. Massive text processing is message processing by separating the parts that are not needed and the parts that are needed. The required part will be stored in a database for further analysis. The flow of massive text processing is shown in Figure 2. Messages will be grouped into groups of text, media, links, and emoji. Each will be stored in the database. The main objective of massive text processing is the speed of data processing and data grouping so that it is easy to analyze.

![Massive Text Processing](image)

2.4.2. *Data classification using K Nearest Neighbors (k-NN)*. Data classification is done using the KNN algorithm. The KNN algorithm works by calculating the value of the given parameter. Data is calculated based on the level of similarity of the message to the dataset that has been prepared. The message is similar to the dataset provided, meaning that the message has a meaning and meaning that is more or less the same as what the dataset provides.

2.5. *Data analysis*

2.5.1. *Word cloud*. Word cloud is a depiction of how often a word appears in a sentence [7]. In this study, the use of word cloud aims to see in detail what topics are discussed most frequently in the group.

2.5.2. *Group activity analysis*. Based on massive text processing, group data can be analyzed to find out how active the group being tested, when is the peak time used, who sent the most messages, and on what date the group was active in replying to messages.
3. Result and Discussion
The data processing process begins with the process of collecting data from a group chat. Data from
group chat is then performed data cleaning and data normalization so that unnecessary data is not
processed in this study. Chat data will be grouped by date, time, sender contact, message, and additional
information. Additional information is the information contained in the message, including information
when the user was added, information on deleting messages, information on media uploads, and
information about leaving the group.

The chat data is then processed to find out what topics are often discussed and when the group is
active. The activity of the group is shown in Figure 3.

![Figure 3](image)

**Figure 3.** (a) Most active date in the group. (b) Hours when users are actively discussing topics in
the group.
In addition to group activity, message types can also be analyzed so that it can be seen which types are often shared in the group and who sends messages the most. The results of the analysis of the types of messages and the users who send messages can also be seen so that you can find out who is the most active and most often sends messages and what messages are sent.

While the most frequently discussed topics are mapped with word cloud as shown in Figure 4.

![Figure 4. The most frequently discussed topics in the group.](image)

Of the total 4519 chat data, analysis was carried out to detect hoaxes with KNN and the value of $K=1$ as shown in Figure 5.

![Figure 5. The value of each test on the dataset.](image)

Statistical details from the results of the KNN test are shown in Table 2.
Table 2. Test results statistics.

| Statistics | Total |
|------------|-------|
| Mean       | 0.536837 |
| Min        | 0.142910 |
| Max        | 0.910547 |
| Std        | 0.070235 |
| 25%        | 0.497828 |
| 50%        | 0.549499 |
| 75%        | 0.910547 |

From the results given, the higher the value, the greater the possibility that messages sent within the group have the potential to be detected as a hoax message, and conversely the smaller the value, the less likely it is that the message will be detected as a hoax message.

4. Conclusion

In the case of hoax detection from a group chat with massive text processing analysis and KNN with raw data from group chat and public datasets, it can be concluded that the most active group occurred between February 2020 and March 2020 and continued to be active until June 2020. In that month, in Indonesia, the COVID case is being reported a lot. With the most active time, which is around 8.00 AM to 9.00 AM and noon around 2.00 PM. The topics most discussed were UNNES and also about COVID. It can be seen from word cloud that in addition to the two topics, there are several discussions such as the corona case, the corona case update, and the Indonesian corona. As for hoax detection, each message has different data characteristics. With an average value of 0.536837, it can be concluded that the group chat is still limited to normal. This means not talking too much about false news. With the results of a minimum value of 0.142910 and a maximum value of 0.910547, it shows that sometimes there are indeed some members who share fake news whose data is not clear, but many also provide reliable data.

References

[1] Ahmed H, Traore I and Saad S 2017 *Lec. Not. in Comp. Science* Springer 10618 127
[2] Aziz A, Foozy C F M, Shamala P and Suradi Z 2018 *Ind. J. Elec, Eng, Comp, Sci.* 12 607
[3] Chawla E G and Saini R 2016 *Int. J. Tren, Res. Dev.* 3 479
[4] Hernández J C, Herrández C J, Sierra J M and Ribagorda A 2002 *Proc. IEEE Ann. Int. Carn. Conf. on Sec. Technology* 102
[5] Divine H N 2019 *J. of Soc. and Pol. Sciences* 22 98
[6] Imron M A and Prasetyo B 2020 *Journal of Soft Computing Exploration* 1 56
[7] Mustaqim T, Umam K and Muslim M A 2020 *J. of Phy.: Conf. Series* 1567
[8] Prasetijo A B, Isnanto R R, Eridani D, Soetrisno Y A A, Arfan M, and Sofwana A 2017 *Proc. 2017 4th Int, Conf. Infor, Tech, Comp, Elec, Eng.* 45 49
[9] Santos H A, Rachmawanto E H, Hidayati U 2020 *Scientific Journal of Informatics* 7 228
[10] Santos H A, Rachmawanto E H, Nugrahana A, Nugroho A A, Setiadi D R I M and Basuki R S 2020 *Telecom. Comp. Elec. and Control* 18 799
[11] Sharma A and Suryawanshi A 2016 *Int. J. Comp. Appl.* 136 28
[12] Sukarsa I M, Buana P W, Arynasta I P K 2020 *Scientific Journal of Informatics* 7 189
[13] Udaya K T, Sravanthi V, Anita G, Rajesh M and Rajesh M 2020 *J. Adv, Res, Dyn, Cont, Syst.* 12 845
[14] Zhang S, Li X, Zong M, Zhu X and Cheng, D 2017 *ACM Trans, Intel, Sys, Technol.* 8 3
[15] Zuliarso E, Anwar M T, Hadiono K and Chasanah I 2020 *IOP Conf, Ser.: Mater, Sci, Eng.* 835 1