Mobile Forensics for Cyberbullying Detection using Term Frequency - Inverse Document Frequency (TF-IDF)

Imam Riadi[1], Sunardi[2], Panggah Widiandana[3]

[1] Department of Information System, Universitas Ahmad Dahlan, Yogyakarta
[2] Department of Electrical Engineering, Universitas Ahmad Dahlan, Yogyakarta
[3] Master Program of Informatics, Universitas Ahmad Dahlan, Yogyakarta

ARTICLE INFO

Article history:

Received 04 November 2019,
Revised 01 December 2019,
Accepted 12 January 2020.

Keywords:

Cyberbullying,
TF,
IDF,
Forensics,
Mobile.

ABSTRACT

The case of cyberbullying in Indonesia was ranked third in the world in 2015 and as much as 91% was experienced by children. RSA Anti-Fraud Command Center (AFCC) report reports that in 2015 45% of transactions were carried out through mobile channels, while 61% of fraud occurred through mobile devices. WhatsApp in July 2019, 1.6 billion users access the WhatsApp messenger every month. The data opens a reference for investigators to better anticipate cybercrime actions that can occur in the WhatsApp application because more users are using the application. In this study using the TF-IDF method in detecting cyberbullying, that occurs to be able to add a reference for investigators. The conclusions that have been obtained from the simulation of conversations between four people in a WhatsApp group get the results of the cyberbullying rate that the user "C" has a cyberbullying rate of 50% from the data proving that the TF-IDF method can help investigators detect someone who will commit cyberbullying actions but in its development, a better way is needed when preprocessing so that the abbreviation or changing words can still be detected perfectly.

Corresponding Author:

Imam Riadi,
Department of Information System,
Universitas Ahmad Dahlan, Yogyakarta, Indonesia.
Email: imam.riadi@is.uad.ac.id

1. INTRODUCTION

Crimes that occur in the digital era develop along with technological developments [1]. Cyberbullying in Indonesia in 2015 ranked third in the world and 91% of them were children [2]. A report released by the RSA Anti-Fraud Command Center (AFCC) states that from 2013 to 2015, there was an increase in cybercrime activity reaching 173% worldwide with total losses reaching US$ 325 billion. The report also reported that in 2015 45% of transactions were carried out through mobile channels, while 61% of fraud occurred through mobile devices [3]. Social media such as Facebook, WhatsApp, Twitter, etc., is one of the ways for someone to do cybercrime. WhatsApp is an instant messaging (IM) application for smartphones that can run on various operating systems such as Apple iOS, BlackBerry, Android, Symbian Nokia Series 40 and Windows Phone. WhatsApp helps someone to be able to chat online, share files, exchange photos and other features that attract users [4]. According to statistical data, the Statista website shows the number of WhatsApp users in July 2019, 1.6 billion users access WhatsApp messenger every month [5]. The data opens a reference for investigators to anticipate better cybercrime actions that can occur in the WhatsApp application because more users are using the application.

The law on cybercrime is regulated in the law on ITE in Indonesia. ITE crimes can be criminalized by civil or civil law according to the level of crime committed, the process of cybercrime arrest by the authorities based on evidence of crime stored on a smartphone or on other hardware that can be used as evidence in a court of law such as user name, IP address and stamp time [6]. In the field of technology, forensic analysis of digital or electronic evidence is referred to as computer forensics or digital forensics [7]. Mobile forensics is needed to conduct forensic analysis relating to evidence in the form of cellular devices [8]. In conducting mobile forensics, it requires a reference on how to analyze someone who identifies cyberbullying on mobile, making it easier for investigators to find a cyberbullying action.

The problem that is often found in cyberbullying is that it is difficult to identify that the victim and the perpetrator committed cyberbullying because the checks were carried out by eye and there is no strong reference to prove the cyberbullying case has been carried out by the perpetrator against the victim. The research is expected to add a reference for investigators to obtain or identify cyberbullying actions that have been circulating. The method used to identify or detect cyberbullying actions using the TF-IDF method is one method to search for a word in a text by preparing it before searching for a word that is in the text. The use of the TF-IDF method will search for the same words in the keywords in the database so that the same words in the text or conversations in one person will be weighted and see the sentences that lead to cyberbullying actions.

2. RESEARCH METHOD

The method in this study began with digital evidence in the form of messages from the perpetrators and victims of cyberbullying research methods can be seen in Figure 1. The figure shows how to identify cyberbullying of evidence that will be raised by cloning the evidence before analyzing the evidence that is in the evidence. The results of cloning will get evidence in the form of data, and after getting an existing chat message, then do text mining to look for important words from the perpetrators and victims. The similarity is used when getting important words and similarities are searched by using the TF-IDF method between keywords and important words that are in the evidence for cyberbullying detection that has been done on the WhatsApp evidence group. Then cyberbullying identification can be known.

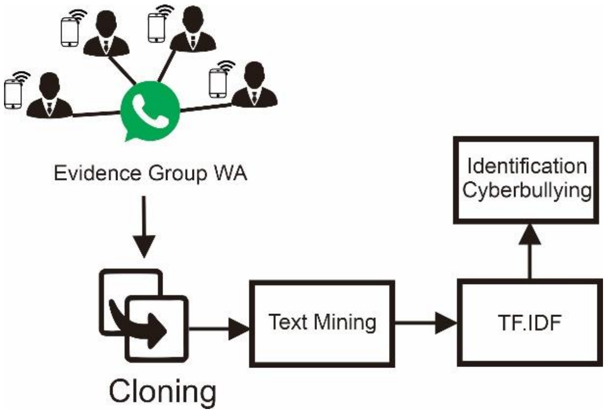

Figure 1. Cyberbullying identification flow

3.1. Preprocessing

Preprocessing is the initial stage of text mining. This stage includes all the routines and the process for preparing data that will be used in the knowledge discovery operation of a text mining system [9]. Preprocessing consists of several stages, namely case folding, tokenizing, stopword, and stemming. Case folding is a step that changes all the letters in a document into lowercase letters. Only letters "a" through "z" can be accepted. Characters other than letters are omitted and are considered delimiter [10]. Tokenizing is the process of decomposing the description that was originally in the form of sentences into words and eliminating delimiter-delimiter such as periods (.), Commas (,), quotation marks ("), parentheses (()), spaces and numeric characters that are in that word [11]. Stopword is a vocabulary that is not a feature (unique word) of a document. For example, "*di*", "*oleh*", "*pada*", "*sebuah*", "*karena*" and so forth. Stopwords are defined as irrelevant concerning the main subject of the database, although they may often be contained in documents. Stopwords include determinants, conjunctions, prepositions, and the like [12]. After going through the stopword removal process, the next action is the stemming process. Stemming is the process of mapping and decomposing various forms of a word into its basic word form (stem) [13].

3.2. Term Frequency

Term Frequency is a way to find the weight of a document. Where will be seeking the number of occurrences of the term in the document? The greater the appearance of a term, it will affect the amount of weight and the suitability value. Following is the equation of Term Frequency can be seen in (1).

$$W(d,t) = TF(d,t) \tag{1}$$

Information:
TF(d,t): the frequency of the term t in each document, which will then be used for the calculation of TF.IDF weighting.

3.3. Inverse Document Frequency

Inverse Document Frequency is a method for calculating the distribution of terms in documents [14]. The following is the equation of Inverse Document Frequency which can be seen in (2).

$$idf_t = \log 10\left(\frac{N}{df_t}\right)+1 \tag{2}$$

Information:
N : the total number of all documents in a conversation that occurred on the WhatsApp application.
idf_t : the number of documents containing the target word. The less the number of documents containing the target word, the greater the weight of the IDF.

3.4. TF-IDF Weighting

The TF-IDF formula is multiplying the weight of TF with the IDF of each word. The TF-IDF formula can be seen in (3).

$$\mathrm{w} = \mathrm{tf_t} \times \mathrm{idf_t} \quad \text{Or} \quad w = tf_t \times \log 10\left(\frac{N}{df_t}\right)+1 \tag{3}$$

Information:
w : weight or the result of multiplication between term frequency and inverse document frequency.
tf_t : term frequency is the number of terms of each conversation in the WhatsApp group
idf_t : term-document frequency is a lot of documents that contain terms on Query

3.5. Normalization of Max-Min

The Min-max method is the simplest in the process of a linear transformation of the original data. After the Min-max normalization process, a balance comparison value can be obtained between the value before the normalization process and the value after the normalization process [15]. The max-min normalization equation can be seen in (4).

$$d' = \frac{d - \min(p)}{\max(p) - \min(p)} X\ 100\% \tag{4}$$

Information:
d' : the new value obtained for normalization results in the form of a percentage where the largest value is max (p).
d : the value to be normalized is the value to be seen what percentage of the value is if the largest value is max (p)
$\min(p)$: min is the smallest value that appears from this attribute, this value is the smallest as the lower limit of normalization
$\max(p)$: max is the largest value of these attributes, this value is the largest as the upper limit of normalization

3. RESULTS AND DISCUSSION

Data were taken from simulation data in a group consisting of four people having a conversation in the group. The scheme can be seen in Figure 2 that shows the perpetrators of cyberbullying of victims through the WhatsApp group and produce a dialogue or conversation that can be seen in Table 1. Table 1 contains two columns, and the first two rows are users or people communicating a conversation, then the second column is the contents of the conversation expressed in a group. Evidence of data that has been obtained will be carried out the identification process at what level of cyberbullying is done, flowchart identification of cyberbullying can be seen in Figure 3. Figure 3 is the stage for cyberbullying identification to facilitate handling or analyzing bullying. The stages for identification are starting from case folding, tokenizing, stop word, stemming, TF-

IDF, and normalization to get a cyberbullying weighting presentation. Figure 3 is a flowchart to detect bullying in a group conversation that starts from case folding, tokenizing, stopword, stemming, applying the TF-IDF method, and Normalization Min-Max.

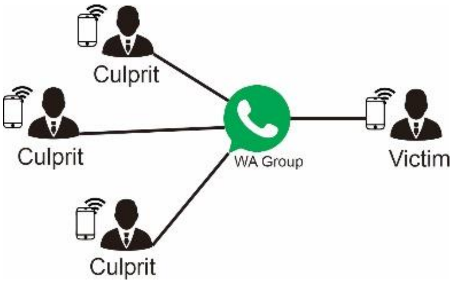

Figure 2. Bullying Scheme

Table 1. Conversations in the WA Group

User	Conversations
a	*: “GOBLOK Lo bud kenapa pak guru lo ingetin untuk ngasih tugas”*
b	*: ”iya ini si budi bukannya enak2 gak ada tugas”*
c	*: “iya caper amat sih lu jadi orang, sok2an jadi pahlawan kesiangan”*
d	*: “soalnya pak guru biasanya ngasih tugas untuk kitakan?”*
e	*: “ya tapi jangan lo ingetin juga kampret”*
f	*: “Sudah jangan terlalu dipermasalahkan kasian si budi”*
g	*: “habisnya dia carimuk banget”*
h	*: “budi tolol...budi tolol”*
i	*: ”PR tu wajib tapi jika tidak itu anugrah wkwk”*
a	*: “ah TOLOL lu nambah-nambahin kerjaan gue aja”*
f	*: “sudah jangan terlalu dipermasalahkan kasian si budi”*
c	*: “Sialan lo budi, harusnya kita bisa enak-enak minggu ini tapi gegara budi semua agenda hancur sudah”*
b	*: “makannya lo jadi orang jangan sok-sokan carimuk didepan guru”*
g	*: “yapsi betul sekali dia itu emang carmuk”*
d	*: “sorry temen2 tadi aku pikir pak guru emang tidk akan memberi tugas kepada kita”*
b	*: “dasar gila lo bud, tugas kita sudah banyak malah lo tambah-tambahin”*
c	*: “temenan sama orang idiot itu emang menyusahkan kita”*
e	*: “gobloknya gak ada habisnya lah pokoknya”*
a	*: “iya betul tu si budi emang TOLOL.”*
h	*: ”IDIOT”*
i	*: ”betul-betul-betul :D”*

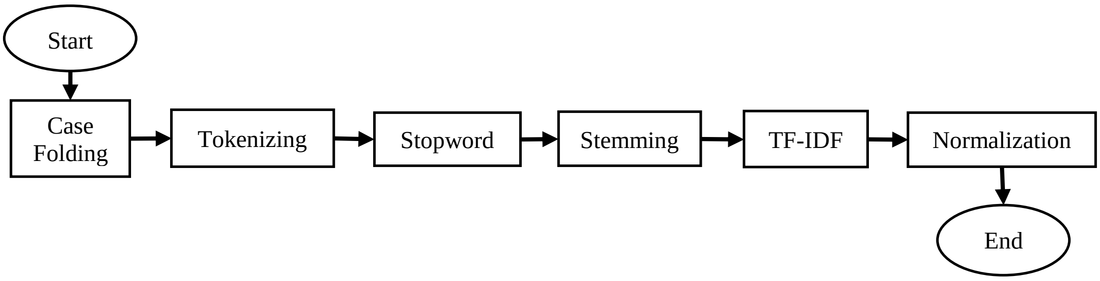

Figure 3. Bullying Identification Flowchart

3.1. Case Folding

The results of changes through the case folding stage can be seen in Table 2. Table 2 contains two columns and the first two rows are users or people communicating a conversation, then the second column is the contents of the conversation expressed in a group. The difference between Table 1 and Table 2 has changed from uppercase to lowercase letters, as in the word "TOLOL" to "tolol." Case folding results are used to convert the entire conversation to a standard form so that it makes it easier to prepare the text.

Table 2. Case Folding

User	Conversations
a	*: "goblok lo bud kenapa pak guru lo ingetin untuk ngasih tugas"*
b	*: "iya ini si budi bukannya enak2 gak ada tugas"*
c	*: "iya caper amat sih lu jadi orang, sok2an jadi pahlawan kesiangan"*
d	*: "soalnya pak guru biasanya ngasih tugas untuk kitakan?"*
e	*: "ya tapi jangan lo ingetin juga kampret"*
f	*: "sudah jangan terlalu dipermasalahkan kasian si budi"*
g	*: "habisnya dia carimuk banget"*
h	*: "budi tolol...budi tolol"*
i	*: "pr tu wajib tapi jika tidak itu anugrah wkwk"*
a	*: "ah tolol lu nambah-nambahin kerjaan gue aja"*
f	*: "sudah jangan terlalu dipermasalahkan kasian si budi"*
c	*: "sialan lo budi, harusnya kita bisa enak-enak minggu ini tapi gegara budi semua agenda hancur sudah"*
b	*: "makannya lo jadi orang jangan sok-sokan carimuk didepan guru"*
g	*: "yapsi betul sekali dia itu emang carmuk"*
d	*: "sorry temen2 tadi aku pikir pak guru emang tidk akan memberi tugas kepada kita"*
b	*: "dasar gila lo bud, tugas kita sudah banyak malah lo tambah-tambahin"*
c	*: "temenan sama orang idiot itu emang menyusahkan kita"*
e	*: "gobloknya gak ada habisnya lah pokoknya"*
a	*: "iya betul tu si budi emang tolol"*
h	*: "idiot"*
i	*: "betul-betul-betul :d"*

3.2. Tokenizing

The results of changes through the tokenizing stage can be seen in Table 3. Table 3 contains two columns, and the first two rows are users or people communicating a conversation, then the second column is the contents of the conversation expressed in a group. The difference between Table 2 and Table 3 that has changed is the comma "," in each sentence the sign is a word separator between words. The results of this tokenization are used to divide the sentence into words that have been spoken by the offender in a conversation in the group so that it is easy to do the elimination of non-essential words contained in the conversation.

Table 3. Tokenizing

User	Conversations
a	*: "goblok", "lo", "bud", "kenapa", "pak", "guru", "lo", "ingetin", "untuk", "ngasih", "tugas"*
b	*: "iya", "ini", "si", "budi", "bukannya", "enak2", "gak", "ada", "tugas"*
c	*: "iya", "caper", "amat", "sih", "lu", "jadi", "orang", "sok2an", "jadi", "pahlawan", "kesiangan"*
d	*: "soalnya", "pak", "guru", "biasanya", "ngasih", "tugas", "untuk", "kitakan?"*
e	*: "ya", "tapi", "jangan", "lo", "ingetin", "juga", kampret"*

f	*: “sudah”, “jangan”, “terlalu”, “dipermasalahkan”, “kasian”, “si”, “budi”*
g	*: “habisnya”, “dia”, “carimuk”, “banget”*
h	*: “budi”, “tolol”, “budi”, “tolol”*
i	*: ”pr”, “tu”, “wajib”, “tapi”, “jika”, “tidak”, “itu”, “anugrah”, wkwk”*
a	*: “ah”,” tolol”, “lu”, “nambah”, “nambahin”, “kerjaan”, “gue”, “aja”*
f	*: “sudah”, “jangan”, “terlalu”, “dipermasalahkan”, “kasian”, “si”, “budi”*
c	*: “sialan”, “lo”, “budi”, “harusnya”, “kita”, “bisa”, “enak”, “enak”, “minggu”, “ini”, “tapi”, “gegara”, “budi”, “semua”, “agenda”, “hancur”, “sudah”*
b	*: “makannya”, “lo”, “jadi”, “orang”, “jangan”, “sok”, “sokan”, “carimuk”, “didepan”, “guru”*
g	*: “yapsi”, “betul”, “sekali”, “dia”, “itu”, “emang”, “carmuk”*
d	*: “sorry”,″ “temen2”, “tadi”, “aku”, “pikir”, “pak”, “guru”, “emang”, “tidk”, “akan”, “memberi”, “tugas”, “kepada”, “kita”*
b	*: “dasar”, “gila”, “lo”, “bud”, “tugas”, “kita”, “sudah”, “banyak”, “malah”, “lo”, “tambah”, “tambahin”*
c	*: “temenan”, “sama”, “orang”, “idiot”, “itu”, “emang”, “menyusahkan”, “kita”*
e	*: “gobloknya”, “gak”, “ada”, “habisnya”, “lah”, “pokoknya”*
a	*: “iya”, “betul”, “tu”, “si”, “budi”, “emang”, “tolol”*
h	*: ”idiot”*
i	*: ”betul,” “betul,” “betul,” “:d.”*

3.3. Stopword

The results of changes through the stopword stage can be seen in Table 4. For example, “*di*,” “*oleh*,” “*pada*,” “*sebuah*,” “*karena*,” and so forth. Table 4 contains two columns, and the first two rows are users or people communicating a conversation, then the second column is the contents of the conversation expressed in a group. We can see the difference between Table 3 and Table 4 which has changed, namely the italic and underlined words such as the words “*untuk*,” “*ada*,” “*jadi*,” “*kita*,” “*bisa*,” “*ini*,” “*tapi*,” “*semua*,” “*sudah*,” “*jangan*,” “*akan*,” “*kepada*,” “*sudah*,” “*banyak*,” “*sama*,”*?*”, and “*itu*” will be removed from the conversation sentence.

Table 4. Stopword

User	Conversations
a	*: “goblok”, “lo”, “bud”, “kenapa”, “pak”, “guru”, “lo”, “ingetin”, “untuk”, “ngasih”, “tugas”*
b	*: “iya”, “ini”, “si”, “budi”, “bukannya”, “enak2”, “gak”, “ada”, “tugas”*
c	*: “iya”, “caper”, “amat”, “sih”, “lu”, “jadi”, “orang”, “sok2an”, “jadi”, “pahlawan”, “kesiangan”*
d	*: “soalnya”, “pak”, “guru”, “biasanya”, “ngasih”, “tugas”, “untuk”, “kitakan?”*
e	*: “ya”, “tapi”, “jangan”, “lo”, “ingetin”, “juga”, kampret”*
f	*: “sudah”, “jangan”, “terlalu”, “dipermasalahkan”, “kasian”, “si”, “budi”*
g	*: “habisnya”, “dia”, “carimuk”, “banget”*
h	*: “budi,” “tolol,” “budi,” “tolol”*
i	*: ”pr”, “tu”, “wajib”, “tapi”, “jika”, “tidak”, “itu”, “anugrah”, wkwk”*
a	*: ”ah”, “tolol”, “lu”, “nambah”, “nambahin”, “kerjaan”, “gue”, “aja”*
f	*: “sudah”, “jangan”, “terlalu”, “dipermasalahkan”, “kasian”, “si”, “budi”*
c	*: “sialan”, “lo”, “budi”, “harusnya”, “kita”, “bisa”, “enak”, “enak”, “minggu”, “ini”, “tapi”, “gegara”, budi, “semua”, “agenda”, “hancur”, “sudah”*
b	*: “makannya”, “lo”, “jadi”, “orang”, “jangan”, “sok”, “sokan”, “carimuk”, “didepan”, “guru”*
g	*: “yapsi”, “betul”, “sekali”, “dia”, “itu”, “emang”, “carmuk”*

d	*: “sorry”, “temen2”, “tadi”, “aku”, “piker”, “pak”, “guru”, “emang”, “tidk”, “akan”, “memberi”, “tugas”, “kepada”, “kita”*
b	*: “dasar”, “gila”, “lo”, “bud”, “tugas”, “kita”, “sudah”, “banyak”, “malah”, “lo”, “tambah”, “tambahin”*
c	*: “temenan”, “sama”, “orang”, “idiot”, “itu”, “emang”, “menyusahkan”, “kita”*
e	*: “gobloknya”, “gak”, “ada”, “habisnya”, “lah”, “pokoknya”*
a	*: “iya”, “betul”, “tu”, “si”, “budi”, “emang”, “tolol”*
h	*: ”idiot”*
i	*: ”betul,” “betul,” “betul,” “:d.”*

3.4. Stemming

The final process of stemming can be seen in Table 5. Table 5 contains two columns, and the first two rows are users or people communicating a conversation, then the second column is the contents of the conversation expressed in a group. We can see the difference between Table 4 and Table 5 that have undergone changes, which are slanted and underlined words such as the words *“ingetin”, “ngasih”, “bukannya”, “kesiangan”, “soalnya”, “biasanya”, “kitakan”, “nambah”, “nambahin”, “kerjaan”, “sialan”, “harusnya”, “sokan”, “didepan”, “emang”, “memberi”, “tambahin”, “temenan”,* and *“menyusahkan”* will be replaced by a standard word like *“ingat”, “kasih”, “bukan”, “siang”, “soal”, “biasa”, “kita”, “tambah”, “tambah”, “kerja”, “sial”, “harus”, “sok”, “depan”, “memang”, “beri”, “tambah”, “teman”,* and "*susah*" from the conversation.

Table 5. Stemming

User	Conversations
a	*: “goblok”, “lo”, “bud”, “kenapa”, “pak”, “guru”, “lo”, “ingat”,” kasih”, “tugas”*
b	*: “iya”, “ini”, “si”, “budi”, “bukan”, “enak2”, “gak”, “tugas”*
c	*: “iya”, “caper”, “amat”, “sih”, “lu”, “orang”, “sok2an”, “pahlawan”, “siang”*
d	*: “soal”, “pak”, “guru”, “biasa”, “kasih”, “tugas”, “kita”*
e	*: “ya”, “tapi”, “lo”, “ingat”, “juga”, kampret”*
f	*: “sudah”, “terlalu”, “masalah”, “kasian”, “si”, “budi”*
g	*: “habis”, “dia”, “carimuk”, “banget”*
h	*: “budi,” “tolol,” “budi,” “tolol”*
i	*: ”pr”, “tu”, “wajib”, “jika”, “anugrah”, wkwk”*
a	*: “ah”, “tolol”, “lu”, “tambah”, “tambah”, “kerja”, “gue”, “aja”*
f	*: “sudah”, “jangan”, “terlalu”, “dipermasalahkan”, “kasian”, “si”, “budi”*
c	*: “sial”, “lo”, “budi”, “harus”, “enak”, “enak”, “minggu”, “gegara”, “budi”, “agenda”, “hancur”*
b	*: “makannya”, “lo”, “orang”, “sok”, “sok”, “carimuk”, “depan”, “guru”*
g	*: “yapsi”, “betul”, “sekali”, “dia”, “memang”, “carmuk”*
d	*: “sorry”, “temen2”, “tadi”, “aku”, “pikir”, “pak”, “guru”, “memang”, “tidk”, “beri”, “tugas”*
b	*: “dasar”, “gila”, “lo”, “bud”, “tugas”, “malah”, “lo”, “tambah”, “tambah”*
c	*: “teman”, “orang”, “idiot”, “emang”, “susah”*
e	*: “goblok”, “gak”, “habis”, “lah”, “pokok”*
a	*: “iya”, “betul”, “tu”, “si”, “budi”, “emang”, “tolol”, “goblok”*
h	*: ”idiot”*
i	*: ”betul,” “betul,” “betul”*

3.5. TF-IDF weighting

From the results of the preprocessing, conversation data will be weighted in a conversation using the TF-IDF method, which can be seen in Table 6. TF-IDF weighting or called Term Frequency and Inverse Document

Frequency is a way to detect cyberbullying done in verbal conversations on WhatsApp, weighting or Term Frequency and the Inverse Document Frequency method can be seen in Table 6.

Table 6. Calculation of TF-IDF

Term	**TF-IDF WITH QUERY**									
	Q	**a**	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**
"gila"	1.698	0	1.698	0	0	0	0	0	0	0
"goblok"	1.522	1.522	0	0	0	1.522	0	0	0	0
"idiot"	1.522	0	0	1.522	0	0	0	0	1.522	0
"sial"	1.698	0	0	1.698	0	0	0	0	0	0
"sok"	1.522	0	1.522	1.522	0	0	0	0	0	0
"tolol"	1.522	3.045	0	0	0	0	0	0	3.045	0
	9.489	4.568	3.221	4.744	0	1.522	0	0	4.568	0
Percentage (%)	100	48.14	33.95	50	0	16.04	0	0	48.14	0

The results of the calculation of Table 6 can be concluded that the perpetrator "C" has a level of use of negative words, which are words that are said to bully the victim. Queries do not count because it is only a database of negative words to look for negative words in a conversation. Numbers in fields Q, A, to I such as 1.69, are obtained from the calculation of TF-IDF using (3).

Known:

tf_t : 1 derived from the number of words "*gila*" in the query

N : 10 is the total number of all documents in a conversation that occurred in the WhatsApp application.

df_t : 2 is the number of documents containing the word "*gila*"

Settlement:

$$w = 1 \times (\log 10 \left(\frac{10}{2}\right) + 1)$$

$$w = 1 \times 1{,}69$$

$$w = 1{,}69$$

From the results obtained 0.39794 settlement, then this calculation is done to get the TF-IDF other words in a conversation. Change to percent using normalization. The formula to calculate it becomes a percentage form to see the weight of cyberbullying that has been done using (4).

$$a = \frac{4.568636 - 0}{9.489455 - 0} X\ 100\% = 48.14435\%$$

The above calculation is the normalization of the TF-IDF calculation results following the predetermined queries as negative words, namely 4.56 for "a" actors, 3.22 for "b" actors, 4.74 for "c" actors, 0 for "d," 1.52 for "e" actors, 0 for "f" actors, 0 for "g" actors, 4.56 for "h" actors, and 0 for "i" actors' perpetrators. This number will be deducted by the minimum value of TF-IDF and then divided by the reduction between the maximum value of TF-IDF and the minimum value of TF-IDF. The calculation results are 48.144%, 33.952%, 50%, 0%, 16.048%, 0%, 0%, 48.144%, and 0% from these figures it can be concluded that the perpetrators who have the heaviest weight of bullying are perpetrators "C" because the highest value of the calculated results.

4. CONCLUSION

The conclusions that have been obtained from the simulation of conversations between four people in a WhatsApp group get the results of the cyberbullying rate that the user "C" has a cyberbullying rate of 50% from the following data can prove that the Term Frequency and Inverse Document Frequency methods can help investigators detect cyberbullying that occurs in WhatsApp group conversations and know the intensity level of negative words in bullying. Further improvement is needed to be able to detect cyberbullying more perfectly, such as the preprocessing process, it is necessary to normalize to minimize word detection errors so that the detection process is accurate even though there are abbreviated words or changing words like "sok2an" and so on. Based on the results of research that has been done can run well and smoothly and can achieve the expected targets or goals.

REFERENCES

[1] I. Riadi, A. Fadlil, and A. Fauzan, "Evidence Gathering and Identification of LINE Messenger on Android Device," *Int. J. Comput. Sci. Inf. Secur. (IJCSIS)*, vol. 16, no. June, pp. 201–205, 2018. Research Gate

[2] A. Fauzan, I. Riadi, and A. Fadlil, "Analisis Forensik Digital Pada Line Messenger Untuk Penanganan Cybercrime," *Annu. Res. Semin. ISBN 979-587-626-0*, vol. 2, no. 1, pp. 159–163, 2017. Online

[3] I. Riadi, A. Fadlil, and A. Fauzan, "A Study of Mobile Forensic Tools Evaluation on Android-Based LINE Messenger," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 10, pp. 201–206, 2018. DOI: 10.14569/IJACSA.2018.091024

[4] RSA, "Current State of Cybercrime," 2016. Online

[5] R. Umar, I. Riadi, and B. F. Muthohirin, "Live forensics of tools on android devices for email forensics," *TELKOMNIKA*, vol. 17, no. 4, pp. 1803–1809, 2019. DOI: 10.12928/telkomnika.v17i4.11748

[6] I. Riadi, S. Sunardi, and A. A. Kadim, "Monitoring Log Aplikasi Mobile Native Menggunakan Framework Grr Rapid Response," *J. Buana Inform.*, vol. 10, no. 1, p. 1, 2019. DOI: 10.24002/jbi.v10i1.1909

[7] G. M. Zamroni and I. Riadi, "Instant Messaging Forensic Tools Comparison on Android Operating System," *KINETIK*, vol. 4, no. 2, pp. 137–148, 2019. DOI: 10.22219/kinetik.v4i2.735

[8] Christof Baron, "Most popular global mobile messenger apps as of July 2019, based on number of monthly active users (in millions)," 2019. Online

[9] J. Feldman, R., & Sanger, *The Text Mining HandBook*. New York: Cambridge University Press., 2007. DOI: 10.1017/CBO9780511546914

[10] C. Triawati, "Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia," *Univ. Telkom*, 2009. Online

[11] F. J. Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, *Text Mining: Predictive Methods for Analyzing Unstructered Information*. New York: Springer, 2005. Online

[12] W. Dragut, E., Fang, F., Sistla, P., Yu, C., & Meng, *Stop Word and Related Problems in Web Interface*. Chicago: Computer Science Department University of Illinois, 2009. DOI: 10.14778/1687627.1687667

[13] F. Z. Tala, *A Study of Stemming Effects On Information Retrieval in Bahasa Indonesia*. The Netherlands: Universiteitvan Amsterdam. Online

[14] S. Suprianto, A. Fadlil, and S. Sunardi, "Aplikasi Sistem Temu Kembali Angket Mahasiswa Menggunakan Metode Generalized Vector Space Model," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 1, pp. 33–40, 2009. DOI: 10.25126/jtiik.2019611184

[15] S. S. and S. K. J. Santosh Kumar Sahu, "A Detail Analysis on Intrusion Detection Datasets," *Int. Adv. Comput. Conf.*, 2014. DOI: 10.1109/IAdCC.2014.6779523