DECISION SUPPORT MODEL FOR DETERMINING CYBERBULLYING TWEET

DARWIN SAMALO[1] , DITDIT NUGERAHA UTAMA[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

[2]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

E-mail: [1]darwin.samalo@binus.ac.id, [2]ditdit.utama@binus.edu

ABSTRACT

The study was driven by the difficulty in distinguishing cyberbullying and bullying and realizing the impact if they are not differentiated. The impacts include the increasing number of cyberbullying incidents, the difficulty of teenagers in distinguishing between banter and cyberbullying, and much cyberbullying masked as banter and vice versa. Most researchers focus only on cyberbullying models, unaware that cyberbullying and banter have a fine line that must be distinguished. The proposed DSM-based model can distinguish this using eleven parameters: violence, hate, aggression, swearing terms, dominant personality, emojis, relationship, the severity of harm, imbalance of power, repetition, and visibility among peers. These parameters will be grouped into four categories based on the method used: manual labeling, lexical category, Laplace's rule of succession, and multi-stage fuzzy. As the test results show, only 27 of the 438 cyberbullying datasets are actually cyberbullying. These results prove that the proposed model can distinguish between cyberbullying and banter, in which case the dataset of the previous model was initially classified as cyberbullying but instead was classified as bullying after further analysis.

Keywords: *Decision Support Model, Multi-stage Fuzzy Logic, Cyberbullying and Banter, Classification, Mining Twitter Data*

1. INTRODUCTION

Cyberbullying is a widespread problem in the social media world. However, a joke or piece of banter may be misinterpreted by certain users as being violent and offensive (cyberbullying). Steer et al. [1] found that most teenagers frequently struggle to identify the motivations or intents of a tweet, which is frequently misinterpreted and interpreted as cyberbullying. According to Dynel [2], banter is a sort of humorous communication between individuals who are comfortable with one another that involves mocking or jokes. Cyberbullying, on the other hand, is the deliberate act of threatening, humiliating, and harassing someone online or through a digital medium [3].

The number of cyberbullying victims could be inaccurate owing to the addition of banter, as banter is misinterpreted as cyberbullying since they are very similar [1]. Furthermore, the number of victims of cyberbullying could rise if the distinction between the two is unclear because many instances of banter are mistaken for cyberbullying [1]. On the other hand, not only is banter misinterpreted as cyberbullying, but cyberbullying can also be masked as banter. According to the report [4], 65% of people think banter can be used as an excuse for bullying. There is currently no model capable of distinguishing between cyberbullying and banter. For instance, the new model developed by Ziems, Vigfusson, and Morstatter [5] can classify cyberbullying but cannot yet distinguish between banter and cyberbullying. Similarly, the model developed by Tripathy et al. [6], which is capable of classifying cyberbullying, fails to differentiate between cyberbullying and banter. Both studies acknowledge the significance of separating these two concepts.

Since it is difficult to determine whether a tweet is classified as cyberbullying or banter by teenagers, this encourages writers to develop a model that can objectively determine whether a tweet is classified in the category of cyberbullying or banter. Therefore, the author created a Decision Support

Model (DSM) to decide if a tweet is considered banter or cyberbullying. Utama [7] claims that DSM can assist in making more fair, impartial, and scientific decisions.

In order to develop the model, it needs to analyze the tweet's category. Racist tweets will automatically be considered cyberbullying because they are unacceptable [8]. Manual labeling will be performed to classify the category of a tweet. A study by Febriany & Utama [9] developed an algorithm for detecting Indonesian profanity based on a corpus. This study categorizes tweets into animal, psychology and idiocy, disabled person, attitude, and general categories. Thus, the model proposed by Febriany & Utama [8] will be adopted in the study to classify a cyberbullying tweet into five categories based on the corpus established by Rezvan et al. [10], namely sexual, appearance-related, intellectual, racial, political, and generic.

Moreover, the model also performs a lexical analysis. This lexical category enables the model to determine whether this tweet contains cyberbullying-related indicators. Empath [11], an open-source Python library, will be used for lexical analysis. Lastly, the model also implements multi-stage fuzzy logic, which will be the final decision-maker for this model.

The author proposes a DSM to determine whether a tweet is classified as cyberbullying or banter. This model has a few stages: manual labeling, lexical analysis, and implementing multi-stage fuzzy logic. This work can be a benchmark for teenagers who want to report cyberbullying incidents. This model can also be employed by social media users seeking to differentiate between cyberbullying and banter. Additionally, it serves as a reminder to researchers about the importance of distinguishing between cyberbullying and bullying. Furthermore, this research is expected to provide valuable insights to the academic community, serving as a reference for other scholars interested in conducting similar research or building similar models.

2. LITERATURE REVIEW

2.1 Related Works

The study by Steer et al. [1] defined banter as enjoyable social contact that can be aggressive but harmless and merely playful. Through banter, people can express their personalities and desires and strengthen their bonds with others by being humorous and lighthearted [12]. On the other hand, cyberbullying is a social interaction that might involve aggressive behavior but is intended to harm others. Several indicators can distinguish between banter and cyberbullying. One of them includes the level of relationship between the people involved. By not having a close relationship, the perpetrator can avoid retaliation and the desire to run away from responsibility for their actions. Practitioners also suggest considering the presence of emojis in this online social interaction. Recipients can potentially misinterpret the intent of the text if it does not contain emojis. This misinterpretation is more common online, as the victim cannot see the perpetrator's facial expressions, tone of voice, or body language. Emojis can be used to eliminate this ambiguity [13].

The difference between cyberbullying and bullying is also described in research by Betts & Spenser [14]. Cyberbullying is a cowardly, anonymous act seeking to disrupt social networks. This act of anonymity is also a strategy for perpetrators to protect themselves from their actions. On the other hand, cyberbullying behavior is described in this study as a criminal act, making people object to their words, sharing other people's personal information, disrupting social networks, and making threats. All of these things are categorized as dangerous acts. While banter is described as a humorous interaction between friends, it is consistent and harmless. So, it emphasizes the previous argument that the relationship between the perpetrators is crucial in determining whether the behavior is cyberbullying or harmless banter.

Additionally, Buglass et al. [15] investigated the distinction between bullying and bullying. In offline and online contexts, banter is a form of social communication that can improve relationships between friends. This statement concurs with Assem’s [16] finding that banter usually occurs between family and friends. However, they also note that the line between cyberbullying and banter is thin, so there are instances where banter is misinterpreted as cyberbullying. This study identifies three significant distinctions between banter and cyberbullying. The first is the victim's perceived intent. Most of the victim's perceptions were positive, whereas the perpetrator was only joking.

In contrast to cyberbullying, where the perpetrator feels an indication of malicious intent, the perpetrator of verbal bullying frequently uses hurtful, harassing, and annoying language. Second, the relationship between the two involved subjects is the same as in previous research. Typically, the relationship includes friends, peers, and family, and those who are related like to engage in banter. Essentially, banter occurs between close friends. In contrast, cyberbullying typically involves more complex relationships, such as exploitation. Lastly, the final distinction is the communication direction. For example, where banter has a reciprocal direction, individuals engage in banter against one another. As for cyberbullying, the direction of communication obtained is typically unidirectional, as it is evident that cyberbullying involves both perpetrators and victims.

While the model for classifying cyberbullying has been developed by Van Hee et al. [17], this work categorizes cyberbullying into several classes: curse, defamation, defense, encouragement, insult, sexuality, and threat. The model utilizes machine learning with the Support Vector Machine (SVM) method. The accuracy of this model is 64.32% for the English Dataset and 58.72% for the Dutch Dataset. To obtain this accuracy, the model performs feature and hyperparameter optimization. They also mentioned issues related to the lack of availability of cyberbullying datasets.

Rosa et al. [18] have also developed cyberbullying detection models utilizing machine learning techniques such as SVM, Logical Regression, and Random Forests. Initially, they did a systematic review to see the big picture of the classification scheme for cyberbullying at the time. After reviewing the models developed in earlier research, they conducted their tests using two datasets (Formspring and the Latest Bullying Dataset V3.0). The results are pretty disappointing, with an f-1 score of 45% and a recall of 46% for the Formspring dataset. They determined that the work and previous research could not accurately integrate the core aspects of the cyberbullying definition.

Additionally, most researchers concentrate solely on improving the quality of the model by adding manipulation and extensive pre-processing of data while ignoring the quality of the underlying dataset. In this investigation, it was determined that several datasets were of poor quality due to the data annotations (there are no clear criteria for cyberbullying). Furthermore, they state that models with f1-scores below 0.80 that do not adhere to the fundamental concept we have stated regarding presenting outcomes solely to the cyberbullying class may be inappropriate for real-world use. This statement motivated Ziems, Vigfusson, and Morstatter [5] to design a reliable cyberbullying detection model. However, according to the machine learning community, the definition of cyberbullying is still unclear; they are merely repeating what social scientists have stated. Therefore, the resulting model will inevitably have a distinct formulation. This study defines cyberbullying by five factors: aggressive language, repetition, harmful intent, peer visibility, and power imbalance. The model has been successfully developed but has not been able to classify banter.

Furthermore, researchers acknowledged that many models are prone to misclassifying incidents of cyberbullying and bullying. Tripathy et al. [6] proposed the ALBERT-based fine-tuning model for classifying cyberbullying. However, they clarified that this model is prone to mistakes and can misinterpret banter between friends as cyberbullying. These statements led to some adolescents claiming that the artificial intelligence (AI)-based model is inaccurate [19]. Due to a model flaw, this may lead to unnecessary conflicts.

2.2 Decision Support Model

The Decision Support Model (DSM) is a model that can help decision-makers make proper, logical, and rational decisions, regardless of the results [20]. There have been numerous studies producing DSM, such as on hotel selection [21], [22], employee recruiting [23], restaurant selection [24], [25], [26], and countless others. According to a book by Utama [7], DSM can assist decision-makers in making more objective, scientific, and fair decisions. A good decision can be accomplished by relying on validated and logical assumptions. Since the model is an imitation of reality, it can be incorrect; nevertheless, using DSM, the model can correctly execute the decision-making process to deliver objective and scientifically accurate judgments.

According to Chappin et al. [27], there are two types of decisions: operational and strategic. Strategic decisions are more long-term in nature, which can affect future performance. In contrast, operational decisions are the opposite of strategic ones, which are more for the short term and are often carried out regularly. Both types of decisions can still be made in conjunction with a DSM. Waste management is an example of a DSM challenge for strategic decisions, as described in Utama's [28] paper titled "Social Media-Based Smart DSM for Strategic Decision Making: Waste Management Case." On the other hand, in Wang et al.'s [29] research, one of the DSM difficulties for operational decisions is picking hotels based on tourist preferences.

Additionally, the book written by Utama [7] defines the phases of doing DSM. As seen in Figure 1, the research procedure resembles a wheel or loop because the constructed model will always be expandable in terms of both techniques and constraints. However, the constructed model needs to be revised, as it will not be identical to the actual outcome; hence, adjustments will be necessary. Here are the steps necessary to develop a DSM: case analysis, decision analysis, parameterization, data collection, DSM construction, decision proposal, model verification, and validation.

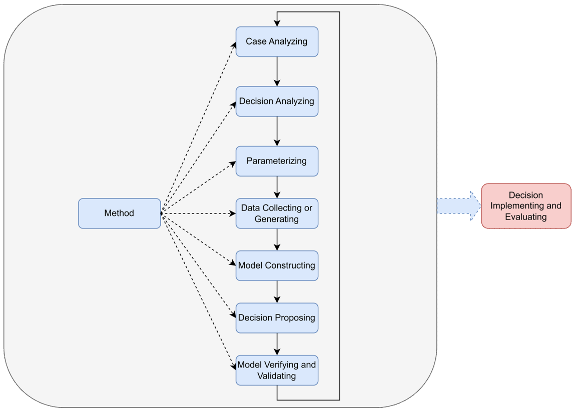
Case Analyzing
Decision Analyzing
Parameterizing
Method
Data Collecting or Generating
Model Constructing
Decision Proposing
Model Verifying and Validating
Decision Implementing and Evaluating

Figure 1: Decision Support Model Process [7]

2.3 Lexcal Category

Lexical categories can be used to categorize a sentence into several related classes. Python already has a package that can be utilized for lexical analysis, Empath [11]. The source dependency relationship (words) is based on ConceptNet. ConceptNet provides a hierarchy of information and a source of category names and seed words. For example, war (seeds) is a conflict (context/class). After that, the model will be expanded by using deep learning's skip-gram architecture (taking a new word as input and trying to predict the context of the word). After enough training, the result will be mapped to vector space. Then, cosine similarity is used to find the nearby terms in the space (vector space model). Cosine similarity measures the similarity between two vectors in an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing roughly in the same direction.

2.4 Manual Labeling

Manual labeling can be performed to obtain a label from an unlabeled new dataset. The manual labeling by Febriany & Utama [9] can be adapted to this work. Before performing this manual labeling, it is better to do the pre-processing stage for data. First, there must be a corpus that serves as a labeling base. Then, the data will be processed manually by checking whether each word in the sentence is in the specified corpus. If the word has the label x in the corpus, the count of label x is increased by one.. After all the words are checked, the label with the maximum value of the count number from each label will be the label of the sentence. For this study, the algorithm was slightly changed. The algorithm does not check all corpus, but only racist. Therefore, if there are racist words, the data will be labeled as cyberbullying. This change is done to avoid the multi-maximum value in the label, which can confuse the algorithm.

2.5 Fuzzy Logic

Fuzzy logic is an object reasoning and computing system in which the objects used for reasoning and computation have fuzzy or fuzzy boundaries [30]. Fuzzy logic is appropriate for ambiguous situations and approximate reasoning [31]. This fuzzy logic enables the model to understand the parameters better than if only true or false values were used. Fuzzy logic combines natural language with logic and converts it into a precise value [7].

3. MATERIALS AND METHODS

In this section, we elaborate on the dataset and proposed model. Three main sub-sections are described in this section.

3.1 Dataset

The dataset used in this model will be mined manually using the Twitter API. Mining could be eased by utilizing the Tweepy Library. The academic researcher's account will be used to maximize the use of this Twitter API. The corpus that used to be the query searching for tweets is the corpus provided by Rezvan et al. [10]. A total of 4,469 data were successfully retrieved based on the corpus. Data mining is required to avoid poor dataset quality and to take data in accordance with the model's requirements. The retrieved data is only data with the type of reply, indicating that this is a two-way communication that can analyze user relationships. The constraint date for the data is set at 2018, so the model only retrieves data from 2018 and later. The data presented in this study are openly available at https://figshare.com/articles/dataset/final_data-tweet_csv/21861756.

3.2 Proposed Method

The model was developed based on DSM, whose steps are outlined in Figure 2. The study was designed as a DSM. The DSM approach involves creating a system to support decision-making by incorporating mathematical models and algorithms. There are several steps, including analyzing the case and making the decisions that need to be made. The first step involved analyzing the case we wanted to focus on, and the second involved analyzing the potential decisions that the decision support model could recommend. This proposed model performs manual labeling, lexical analysis, and multi-stage fuzzy logic. The model and its parameters will be explained in this section.

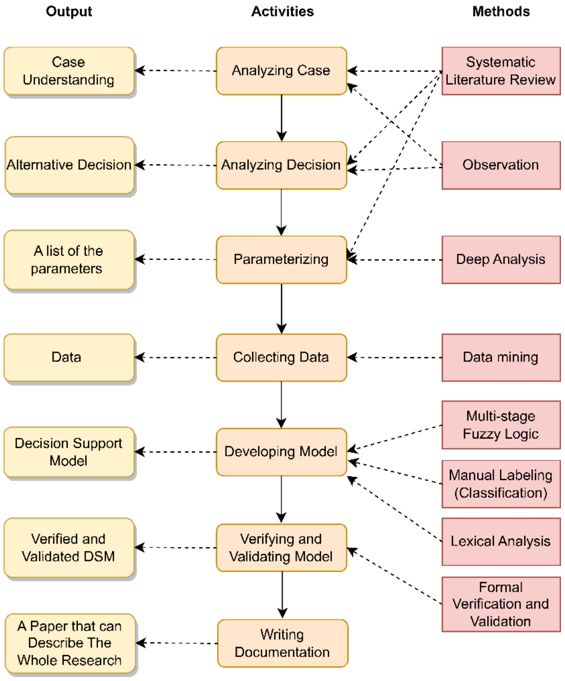

Figure 2: DSM-Based Research Stages

3.3 Model's Parameter

The model used eleven parameters, which can be seen in Table 1. All parameters can be divided into four types based on the method used. The first is using manual labeling for analyzing the severity of harm. Harm is one indication of cyberbullying. However, some levels of harm are very harsh and intolerable, such as racism [8]. Racism will be considered cyberbullying. Second, using lexical categories for analyzing violence, hate, aggression, swearing terms, and dominant personality, all of which are indicators of cyberbullying. The value of this parameter will be obtained by performing a lexical analysis for each category.

Third, using Laplace's rule of succession for analyzing emojis, the person's intentions will be hard to see just from the words of a written tweet. Analyzing the emoji will be very useful in this case. The model will check whether this text contains emojis or not. If emojis exist, the model will be analyzed again using the Laplace estimate equation, according to the data obtained from Novak et al. [32].

The last method uses multi-stage fuzzy. Several parameters will be analyzed in this method; the first is emojis. The outputs of the Laplace rule succession will be analyzed further using fuzzy logic. With fuzzy logic, the model can determine whether the emoji's sentiment is positive, neutral, or negative. Negative ones tend to be cyberbullying. Secondly, the model analyzes the user's relationships. Users who have a good relationship are more likely to banter. Therefore, the model analyzes the parameter based on whether the users follow each other, the total of likes the victim has on the perpetrator's tweet, and the total of replies the victim has on the perpetrator's tweet.

The imbalance of power is also one of the indications of cyberbullying. An imbalance of power will occur if the perpetrator is considered to have greater power than the victim. Therefore, the third parameter analyzes the total count difference between perpetrator and victim followers and whether the user's account is verified. Moreover, aggressive, repeated tweets are considered cyberbullying. Thus, this fourth parameter analyzes the total number of tweets made by the perpetrator to the victim, which contain violence, hate, aggression, swearing terms, and dominant personality. Lastly, analyzing visibility among peers is also essential in this model. Suppose there are other people besides the perpetrator and the victim who like or reply to a tweet. In that case, there is a higher potential for the tweets to become cyberbullying. The visibility among peers indicates other people who help the perpetrator dominate the victim.

Table 1: The Model's Parameters.

Parameter	Reference
Violence	Jones, Waite, & Thomas Clements [33]
Hate	Lehman [34]
Aggression	Steer et al. [1], Betts [35], Lehman [34]
Swearing terms	Kim et al. [36]
Dominant personality	Walker [37]
Emojis	Steer et al. [1]
Relationship	Patterson & Allan [8], Betts & Spenser [14], Buglass et al. [15], Betts [35], Ziems, Vigfusson, & Morstatter [5]
Severity of harm	Patterson & Allan [8]
Imbalance power	Thomas, Connor, & Scott [38], Betts [35], Jones, Waite, & Thomas Clements [33], Kim et al. [36]
Repetition	Ziems, Vigfusson, & Morstatter [5]
Visibility among peers	Ziems, Vigfusson, & Morstatter [5]

3.4 Proposed Model

This model comprises several stages, as illustrated in Figure 3. There are five main stages in this model. Firstly, the data will be mined manually from the Twitter API using an academic researcher's account. A more detailed explanation can be seen in subsection 3.2. After the data is collected, the model performs data preprocessing. Then, several stages are carried out, such as data cleaning, case folding, stop word removal, and lemmatization. The model deletes all duplicate data and processes the tweet using the tweet-preprocessor library (remove URL, mention, hashtag, and reserve words like 'RT' or 'FAV'). Then, the model performs case folding, stop word removal, and lemmatization to maintain and improve the data quality.

Manual labeling will be carried out after the data is cleaned. Manual labeling is done to check whether the tweet is racial (high severity of harm). If the tweet is classified as racial, then it is not needed for further analysis and will be considered cyberbullying. The manual labeling method will be adapted from the model developed by Febriany & Utama [7]. The racial corpus will be collected from the corpus made by Rezvan et al. [10].

The next stage is lexical analysis to determine the category of the sentence that will become the parameter value (violence, hate, aggression, swearing terms, dominant personality). The data not classified as cyberbullying at the previous stage will be analyzed with lexical analysis. The tweet is considered cyberbullying if any parameter has a value greater than zero. On the other hand, if all parameters have a zero value, the tweets will be categorized as banter, and there is no need for further analysis. Fuzzy logic is the last stage of the model. All data that cannot be classified yet will be analyzed at this stage using fuzzy logic. In this stage, five parameters will be used: emojis, relationships, power imbalance, repetition, and visibility among peers. This model utilizes multi-stage fuzzy logic, the outcomes of which can identify whether this tweet is classified as cyberbullying or banter.

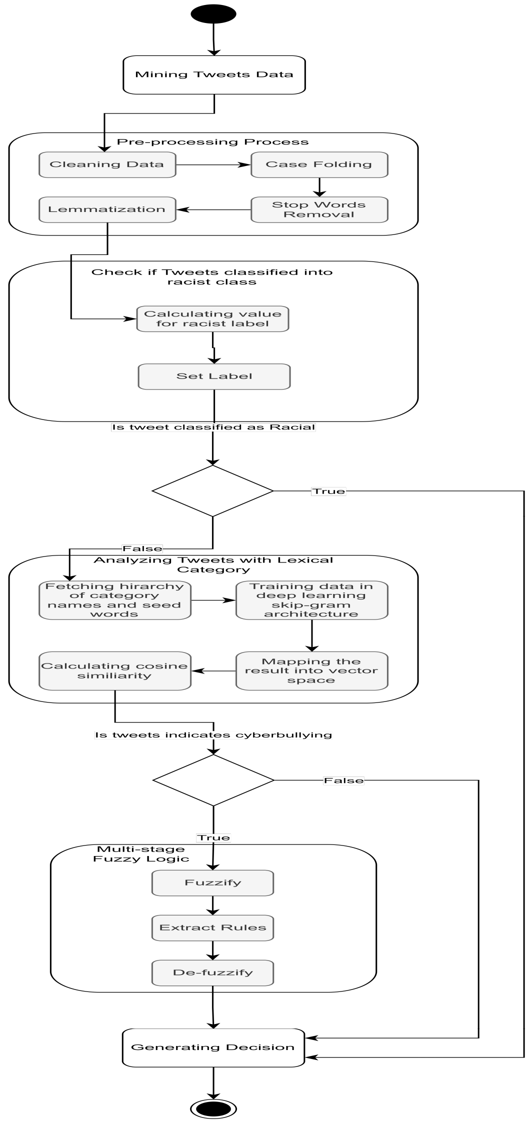
Mining Tweets Data
Pre-processing Process
Cleaning Data
Case Folding
Stop Words Removal
Lemmatization
Check if Tweets classified into racist class
Calculating value for racist label
Set Label
Is tweet classified as Racial
True
False
Analyzing Tweets with Lexical Category
Fetching hirarchy of category names and seed words
Training data in deep learning skip-gram architecture
Mapping the result into vector space
Calculating cosine similiarity
Is tweets indicates cyberbullying
False
True
Multi-stage Fuzzy Logic
Fuzzify
Extract Rules
De-fuzzify
Generating Decision

Figure 3: Activity Diagram of The Model

4. RESULT AND DISCUSSION

4.1 Results and Evaluation

The first step after data mining is the preprocessing process. This stage consists of data cleaning, case folding, stop word removal, and lemmatization. This preprocessing is done so that the text of the tweet can be effectively analyzed. After that, go to the next stage, checking for racism. Of the 4,469 data, 836 data are classified as the cyberbullying class because they have racist words. The results of this step can be seen in Table 2.

Table 2: Result After Classification Using Manual Labeling.

Tweet ID	Text	Class
1564524743142801409	ode m*slim dress like dis *rab country f*ck naive people f**lish man	Cyberbullying (Racial)
…	…	…
1564532517369106434	Ni**a bi**xual lol	Cyberbullying (Racial)

Three thousand six hundred thirty-three unclassified data are carried on to the next stage, namely the lexical category. As explained in Section 3, the data will be analyzed and grouped into five categories: violence (V), hate (H), aggression (A), swearing (S), and dominant personality (DP). The value of this data contains the numbers 0-1. Therefore, any data with a value of 0 in all of the category's sentences will be classified as banter; on the other hand, further processing will be carried out if any data with a value greater than zero is found in any category. Table 3 shows the outcomes at this stage.

Table 3: Result After Lexical Category.

Tweet ID	Lexical Category					Class
	V	H	A	S	DP	
1564524972483428354	0.0	0.0	0.0	0.0	0.0	Banter
…	…	…	…	…	…	…
1564536592894672900	0.0	0.0	0.0	0.0	0.0	Banter
…	…	…	…	…	…	…
1564532517369106434	0.0	0.0	0.0	0.07	0.0	-

A total of 2,169 data have been classified into banter class. The remaining 1464 will be further analyzed. Multi-stage fuzzy logic will be performed based on Table 4. Each weight for the sub-parameter has the same weight, which is 0.09. All the fuzzy values from the sub-parameter are multiplied by the weight. 'Verified accounts' and 'following each other' sub-parameters do not perform fuzzy logic because they only have two values, namely true or false. If the 'verified account' is true, there is an indication of imbalance power; therefore, the value of the fuzzy for imbalance power is added to the verified account weight, which is 0.09. Likewise, if the value 'follow each other' is true, then the model can assume they have a bad relationship (stranger); therefore, fuzzy values in the relationship will be added.

Table 4: The Parameter Details of Fuzzy.

Fuzzy Parameters	Sub Parameter	Linguistic Value
Emojis	Negative (ng)	Low, Medium, and High
	Neutral (nt)	Low, Medium, and High
	Positive (p)	Low, Medium, and High
Relationship	Follow each other (feo)	-
	Likes (rl)	Stranger, Friend, and Close friend
	Reply (rr)	Stranger, Friend, and Close friend
Imbalance Power	Total difference followers (tdf)	No, Low, Medium, and High
	Verified account (va)	-
Repetition	Total repetition tweets (tr)	Low, Medium, and High
Visibility among peers	Like (vl)	Low, Medium, and High
	Reply (vr)	Low, Medium, and High

Additional data will be retrieved from the Twitter API for this step. The additional data will be divided into four categories based on parameters. The first parameter is a relationship that will require two additional data to be retrieved. The first data checks whether the author and victim follow each other (true or false). Second, the count of all author tweets that the victim likes. Lastly, the count of all authors' tweets the victim has replied to.

The second parameter is imbalance power, which will take two additional data. The first data is the follower's difference obtained from the count of author followers minus the count of victim followers. The second data is checking verified accounts. If the author account is verified and the victim account is not verified, then the value is true; otherwise, it is false (true or false).

The third parameter is repetition, which only takes one additional data. The data that will be retrieved for this parameter is the total count of tweets the author sent to the victim. Only tweets that indicate cyberbullying will be counted, which were analyzed first using a lexical category. The last parameter is visibility among peers, which requires two additional data that will be retrieved for this parameter: the count of users who like the tweet besides the author and victim and the count of users who replied to the tweet besides the author and victim. Both data can be analyzed to see the visibility of the perpetrator's tweet. The emoji parameter does not need additional data because this parameter directly analyzes the tweet's text.

In the Twitter API, if the user's privacy is private, then the API cannot retrieve any data from that user. So, this model could not analyze if the perpetrator's or victim's accounts are private. The result of this final step can be seen in Table 5.

Table 5: Final Results.

Tweet ID	Emojis			Relationship			Imbalance Power		Repetition	Visibility among peers		Class
	ng	nt	p	feo	rl	rr	tdf	ng	nt	p	feo	
1564521301963849729	0.46	0.27	0.26	F	0	4	33,962	T	1	2	0	Cyberbullying (0.67)
…	…	…	…	…	…	…	…	…	…	…	…	…
1564518667689365504	0	0	0	T	7	8	-7,949	F	1	6	0	Banter (0.26)

In this model, several methods and formulas are adopted: the manual labeling method [7], the lexical category [11], and Laplace's rule of succession [32]. The data in this model is valid because it was directly taken from the Twitter API, and model verification has been done.

Additionally, the model tested the data obtained from Ziems, Vigfusson, and Morstatter [5]. After analyzing this data using the proposed model, out of 438 data in the bully category, only 27 are classified in the cyberbullying category. For the banter class, there are 114 data categorized into this class. The rest, 297 tweets, could not be found because the user set the privacy to "private".

4.2 Discussion

This proposed model can objectively solve subjective problems; it can distinguish between cyberbullying and banter, which so far have been challenging to distinguish. The model would correspond to human logic by applying fuzzy logic. A prior study conducted by Ziems, Vigfusson, and Morstatter [5] aimed to address the issue of banter being misclassified as cyberbullying by proposing a new model that analyzes both text-based and user-based features, as well as measures the relationship between the author and the target. The model also considers several vital parameters, such as aggressive language, repetition, harmful intent, visibility among peers, and power imbalance, to provide a complete understanding of the analyzed issue.

The new proposed model has further improved upon the previous study by incorporating several vital parameters, including the analysis of emojis, hate speech, and other relevant factors. The results of this new model identify 27 tweets as instances of cyberbullying, whereas the previous model identified 438. Without the ability to accurately differentiate between cyberbullying and banter, the number of reported instances of cyberbullying would likely be significantly higher. It should be noted that this model is limited to the Twitter ecosystem and will not be able to perform analysis if a user has private privacy settings.

5. CONCLUSION AND FUTURE WORKS

Distinguishing between cyberbullying and banter is crucial, considering the increasing number of cyberbullying, the difficulty of teenagers distinguishing between banter and cyberbullying, and much cyberbullying masked as banter and vice versa. This proposed DSM-based model can objectively distinguish between cyberbullying and banter using several parameters: violence, hate, aggression, swearing terms, dominant personality, emojis, relationship, the severity of harm, power imbalance, repetition, and visibility among peers. The model uses manual labeling, lexical analysis, and fuzzy logic. As a result, only 27 data were classified as cyberbullying, compared to 438 in the previous model [5].

This study has made a significant contribution to the classification of cyberbullying by proposing a new model that considers not only user or text features but also several vital parameters, such as the relationship between users. This model can differentiate between cyberbullying and banter, a challenging issue in previous research. The results of this study highlight the importance of acknowledging and exploring the subtle distinction between banter and cyberbullying for developing a reliable classification model. Future researchers should continue to build upon the findings of this study, further enhancing our understanding of this issue. Additionally, they could delve deeper into these aspects to better comprehend the complexities involved in identifying and classifying such behaviors in various online environments. It is imperative to recognize the subtle distinction between banter and cyberbullying, which is often murky and frequently overlooked.

REFERENCES:

[1] O. L. Steer, L. R. Betts, T. Baguley, & J. F. Binder, "I feel like everyone does it"-adolescents' perceptions and awareness of the association between humour, banter, and cyberbullying, *Computers in Human Behavior*, vol.108, no.106297, 2020.

[2] M Dynel, No Aggression, Only Teasing: The Pragmatics of Teasing and Banter, *Lodz Papers in Pragmatics*, vol.4(2), pp.241-261. 2008.

[3] L. Mark, & K. T. Ratliffe, Cyber Worlds: New Playgrounds for Bullying, *Computers in the Schools*, vol.28(2), pp.92-116, 2011.

[4] The Cybersmile Foundation, "Banter or Bullying No Offence", Available: https://www.cybersmile.org/wp-content/uploads/Banter-or-Bullying-No-Offence-High-Res.pdf.

[5] C. Ziems, Y. Vigfusson, & F. Morstatter, Aggressive, repetitive, intentional, visible, and imbalanced: Refining representations for cyberbullying classification. *International AAAI Conference on Web and Social Media*, pp. 808-819, 2020.

[6] J. K. Tripathy, S. S. Chakkaravarthy, S. C. Satapathy, M. Sahoo, & V. Vaidehi, ALBERT-based fine-tuning model for cyberbullying analysis, *Multimedia Systems*, vol.28(6), pp.1941-1949, 2022.

[7] D. N. Utama, Logika Fuzzy untuk Model Penunjang Keputusan, Yogyakarta: Garudhawaca. 2021.

[8] L. J. Patterson, & A. Allan, Adolescent perceptions of bystanders' responses to cyberbullying, *New Media & Society*, vol.19(3), pp.366-383, 2017.

[9] A. Febriany, & D. N. Utama, Analysis Model for Identifying Negative Posts Based on Social Media, *International Journal of Emerging Technology and Advanced Engineering*, pp.96-103, 2021.

[10] M. Rezvan, S. Shekarpour, F. Alshargi, K. Thirunarayan, V. L. Shalin, & A. Sheth, Analyzing and learning the language for different types of harassment. *PLOS ONE*, vol.15(3), no.e0227330, 2020.

[11] E. Fast, B. Chen, & M. S. Bernstein, Empath: Understanding Topic Signals in Large-Scale Text, *2016 CHI conference on human factors in computing systems*, pp. 4647-4657, 2016.

[12] S. L. Pang, & J. A. Samp, Goals, Power and Similarity: Responses to Banter in Initial Interactions, *Western Journal of Communication*, pp.1-22, 2022.

[13] A. Kumar, S. R. Sangwan, A. K. Singh, & G. Wadhwa, Hybrid deep learning model for sarcasm detection in Indian indigenous language using word-emoji embeddings, *Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[14] L. R. Betts, & K. A. Spenser, "People think it's a harmless joke": young people's understanding of the impact of technology, digital vulnerability and cyberbullying in the United Kingdom, *Journal of Children and Media*, vol.11(1), pp.20-35, 2017.

[15] S. L. Buglass, L. Abell, L. R. Betts, R. Hill, & J. Saunders, Banter Versus Bullying: a University Student Perspective, *International Journal of Bullying Prevention*, vol.3(4), pp.287-299, 2021.

[16] S. Assem, The Development of Sentiment Analysis from a linguistic perspective, *The Egyptian Journal of Language Engineering*, vol.9(2), pp.40-52, 2022.

[17] C. V. Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever, B. Verhoeven, . . . V. Hoste, Automatic detection of cyberbullying in social media text, *PloS one*, vol.13(10), no.e0203794, 2018.

[18] Rosa, H., Pereira, N., Ribeiro, R., Ferreira, P. C., Carvalho, J. P., Oliveira, S., . . . Trancoso, I. (2019). Automatic cyberbullying detection: A systematic review, *Computers in Human Behavior*, vol.93, pp.333-345, 2019.

[19] T. Milosevic, K. Verma, M. Carter, S. Vigil, D. Laffan, B. Davis, & J. O'Higgins Norman, Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From Youth Perspective, *Social Media+ Society*, vol.9(1), no.20563051221147325, 2023.

[20] D. N. Utama, Sistem Penunjang Keputusan: Filosofi Teori dan Implementasi, Yogyakarta: Garudhawaca, 2017.

[21] P. K. Kwok, & H. Y. Lau, Hotel selection using a modified TOPSIS-based decision support algorithm, *Decision Support Systems*, vol.120, pp.95-105, 2019.

[22] H. g. Peng, H. y. Zhang, & J. q. Wang, Cloud decision support model for selecting hotels on TripAdvisor.com with probabilistic linguistic information, *International Journal of Hospitality Management*, vol.68, pp.124-138, 2018.

[23] C. E. Pah, & D. N. Utama, Decision support model for employee recruitment using data mining classification, *International Journal of Emerging Trends in Engineering Research*, vol.8(5), 2020.

[24] D. N. Utama, M. R. Putra, M. Su'udah, Z. Melinda, N. Cholis, & A. Piqri, SA-Optimization based decision support model for determining fast-food restaurant location, *Computer Science On-line Conference*, pp. 333-342, 2017.

[25] D. N. Utama, L. I. Lazuardi, H. A. Qadrya, B. M. Caroline, T. Renanda, & A. P. Sari, Worth eat: An intelligent application for restaurant recommendation based on customer preference (Case study: Five types of restaurants in Tangerang Selatan region, Indonesia), *2017 5th international conference on information and communication technology (ICoIC7)*, pp.1-4, 2017b.

[26] M. Hartanto, & D. N. Utama, Intelligent decision support model for recommending restaurant, *Cogent Engineering*, vol.7(1), no.1763888, 2020.

[27] E. Chappin, G. Dijkema, K. v. Dam, & Z. Lukszo, Modeling strategic and operational decision-making—an agent-based model of electricity producers, *21st annual European Simulation and Modelling Conference (ESM2007)*, pp. 22-24, 2007.

[28] D. N. Utama, Media social based smart DSM for strategic decision making: Waste management case, *International Journal of Recent Technology and Engineering,* vol.8(3), pp.7308-7312, 2019.

[29] X. k. Wang, S. h. Wang, H. y. Zhang, J. q. Wang, & L. Li, The recommendation method for hotel selection under traveller preference characteristics: A cloud-based multi-criteria group decision support model, *Group Decision and Negotiation*, vol.30(6), pp.1433-1469, 2021.

[30] L. A. Zadeh, Fuzzy logic—a personal perspective, *Fuzzy sets and systems*, vol.281, pp.4-20, 2015.

[31] L. A. Zadeh, Is there a need for fuzzy logic? *Information sciences*, vol.178(13), pp.2751-2779, 2008.

[32] P. K. Novak, J. Smailović, B. Sluban, & I. Mozetič,Sentiment of emojis, *PloS one*, vol.10(12), no.e0144296, 2015.

[33] S. N. Jones, R. Waite, & P. Thomas Clements, An evolutionary concept analysis of school violence: from bullying to death, *Journal of forensic nursing*, vol.8(1), pp.4-12, 2012.

[34] B. Lehman, Stopping the hate: Applying insights on bullying victimization to understand and reduce the emergence of hate in schools, *Sociological inquiry*, vol.89(3), pp.532-555, 2019.

[35] L. R. Betts, Definitions of cyberbullying, *Cyberbullying*, pp. 9-31, 2016.

[36] H. Kim, Y. Han, J. Song, & T. M. Song, Application of social big data to identify trends of school bullying forms in South Korea, *International journal of environmental research and public health*, vol.16(14), no.2596, 2019.

[37] S. Walker, Workplace Bullying: An individual differences perspective on" diagnosing" important organizational members, *Academy of Business Research Journal*, vol.1, pp.22-36, 2019.

[38] H. J. Thomas, J. P. Connor, & J. G. Scott, Integrating traditional bullying and cyberbullying: challenges of definition and measurement in adolescents – a review, *Educational psychology review*, vol.27(1), pp.135-152, 2015.