Aspect Based Multilabel Text Classification for Identifying Dangerous Speech Twitter Text

by Totok Wahyu Abadi

Submission date: 14-Jun-2023 11:21AM (UTC+0700)
Submission ID: 2115711529
File name: 2022049952.pdf (310.96K)
Word count: 4301
Character count: 22678

Aspect Based Multilabel Text Classification for Identifying Dangerous Speech Twitter Text

Yulian Findawati
Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
Department of Informatics, Universitas Muhammadiyah Sidoarjo, Indonesia
yulianfindawati@umsida.ac.id

Kresna Adhi Pramana
Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
kresnaadhipramana5758@gmail.com

Agus Budi Raharjo
Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
agus.budi@its.ac.id

Totok Wahyu Abadi
Department of Computer Communication and Social Science, Universitas Muhammadiyah Sidoarjo, Indonesia
totokwahyu@umsida.ac.id

Diana Purwitasari
Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia
diana@if.its.ac.id

***Abstract*—As part of hate speech, dangerous speech is any expression that can increase the risk of committing violence against other people. So far, hate speech research only explains whether some sentences is categorized as hate speech. It does not explain aspects of the sentences that make them called dangerous speech. Aspects of dangerous speech are social context, historical context, dehumanization, the accusation in the mirror, women and children attack, loyalty to the group, and group threat. This study uses the multi-label text classification method to determine dangerous speeches on Twitter texts based on seven aspects. Then, we assign a weighted score from those aspects to differentiate dangerous and hate speech. Based on the test results show the best performance is the Naive Bayes method with label-based subset accuracy (±36%), instance-based (*average*) accuracy (±86%) and classification accuracy (±77%). However, even though Naive Bayes has the best performance in terms of instance based (*average*) accuracy, the average difference between all methods with Naive Bayes is only ± 0.014, this indicates that other methods also produce quite good performance.**

Keywords—dangerous speech, multi-label text classification, weighted sum model, twitter texts

I. INTRODUCTION

Dangerous speech could be in any expression like texts or images, which most likely will increase the risk of committing violence to other people [1]. Subjects discussed in dangerous speech are generally related to race, ethnicity, religion, class, or sexual orientation. Recent studies stated that the contents of dangerous speech could have one or more combined aspects of social context, historical context, dehumanization, the accusation in the mirror, women and children attack, loyalty to the group, and group threat [2].

An example of Indonesian Twitter text translated into English containing dangerous speech is as follows: "*The thieves are Chinese. BLBI are pigs, billions of corruptions. Including Ahok, 2 billions corruption, Take down AHOK*". To put into context, BLBI is a shocking corruption scandal case in Indonesia and Ahok is the first public official with Chinese ancestry, which was somewhat out of the ordinary at that time. Some aspects of dangerous speech can be found in tweets like corruption as "social context" or "historical context" for the BLBI case and the Chinese Race, which has been considered an enemy for years. Another aspect is derived from the term pig for treating humans or people related to the BLBI case as an animal leads to "dehumanization" or describing other people in ways that deny or diminish their humanity. The repetitive words of corruption also indicate "the accusation in the mirror" aspect, which indicates reversing reality or asserting severe and often mortal threats from a target group. Aside from those four aspects, the last one of "group threat" aspect comes from the term of take down by removing Ahok's position in his current public official. The texts implied that Ahok was characterized as insufficiently loyal or traitorous and might damage the community's purity, integrity, or cleanliness.

Dangerous speech consists of context and message. Thus, by understanding context and messages in a text, we could identify dangerous speech and not only hate speech. Context consists of 2 aspect, namely social context and historical context while message consist of 5 aspect, there are dehumanization, the accusation in the mirror, women and children attack, loyalty to the group, and group threat [2]. *Dehumanization* is describing other as inferior to humans, for example by liken them to disgusting or deadly animals, insects, bacteria, or demons [2] *Accusation in a mirror* is asserting that the audience faces serious and often mortal threats from the target group in other words. *Assertion of attack on women/girls* are suggesting that women or girls of the audience's group have been threatened, harassed, or defiled by members of a target group. *Threat to Group Integrity or Purity* is giving the impression that one or more members of a target group might damage the purity or integrity or cleanliness of the audience group [2]. The aspect from context aspect dangerous speech are social context and Historical context. For example, in *historical context*, is there a history of violence between the groups or describing another group as planning. For example, in *social context* are longstanding competition over resources, previous episodes of violence, difficult living conditions, an ongoing war [2].

Works related to dangerous speech are often extended as hate speech analysis. However, texts of hate speech do not always represent dangerous speech because aspects in the contents should be identified first, like studies focusing solely on classifying three labels of hate speech, offensive, and neither from English tweets [3]. Some studies have classified controversial topics including feminism, immigrants, and Islamic-leftism into six class labels (hate speech, abusive, offensive, disrespectful, fearful, and normal) [4]. In contrast, others were a binary classification of hate speech on Twitter texts targeting English and Spanish immigrants, especially women [5]. There are variation class labels of hate speech such as offensive, abusive, hateful, aggressive, cyberbullying, spam, and normal on English tweets [6]. Other than aspects in the contents, certain works considered the speaker tone in texts used speech levels [7], such that the used dataset was built for abusive words and hate speech, including targets, categories, and speech levels.

Machine learning classification methods like Support Vector Machine (SVM), Naive Bayes (NB), Random Forest Decision Tree (RFDT), Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) are frequently explored in multi-label text classification for abusive language and hate speech detection [8]. The works also include fine-grained tasks such as detecting target, category, and level of hate speech in Indonesian tweets with an average accuracy value of 76%. Besides machine learning, deep learning methods like RFDT, BiLSTM, and BiLSTM with pre-trained BERT models have resulted in better accuracy of around 76%. Slightly better models with SVM+CC, SVM+LP, CNN, and CNN+DistiBERT combining machine learning and deep learning have presented an increased accuracy value of ± 75%. Statistical-based Gradient Boosting was utilized to classify data which is similar to hate speech called as multi-label toxicity with higher accuracies of ± 98% [9]. Another multi-label problems on tweets is classifying emotions with bagging classifier [10]. And how to detect traffic events based on twitter text using CNN and LSTM multilabel classification[14]

Motivated by previous studies, current work investigates dangerous speech identification on Indonesian Twitter texts. During political campaigns, especially with the common usage of mobile phones, people tend to become haters in social media like Twitter, which makes us select tweets related to those activities. Most aspects of dangerous speech are high likely found in Twitter texts during those political events since followers of one candidate will ignite public opinion based on unfavorable situations from other candidates. Comparable analysis on hate speech identification rather than dangerous speech during political events confirmed our work's value [11]. After identifying aspects of dangerous speech on Twitter texts, the proposed method assigns a score indicating dangerous speech or hate speech. Similar to that purpose, previous works on checking the quality of data based on multi aspects were conducted to assign priority scores in a product with the Weighted Sum Model (WSM) [12]. Thus, our proposed steps to identify dangerous speech are recognizing aspect labels and assigning a weighted score to differentiate dangerous and hate speech. The identification of dangerous speech applies multi-label text classification to recognize aspects of dangerous speech and a weighted Sum Model (WSM) to decide the status of dangerous speech.

II. METHODOLOGY

Our proposed method to identify dangerous speech with aspects is displayed in Fig. 1.

A. Data Preparation

Since our work enhances previous works on hate speech, the dataset is collected from those works [7] [8] without data labeled as non-hate speech texts, which leaves 864 Twitter texts labeled as hate speech. For training data, 648 tweets have been annotated by a socio-linguistic expert with those seven aspects [2] regarded as contexts (C1: social context and C2: historical context) and messages (M1: dehumanization, M2: accusation in a mirror, M3: assertion of attack against women,

TABLE I. SAMPLE OF ANNOTATION LABELS FOR DANGEROUS SPEECH BASED ON CONTEXT AND MESSAGE ASPECTS

Data-Id	Hate Speech Tweet	Label $l_{Cx}(D_i)$ and $l_{My}(D_i)$						
		C1	C2	M1	M2	M3	M4	M5
D1	The thieves are Chinese. BLBI are pigs, billion of corruption. Including Ahok, 2 billions corruption, Take down AHOK	1	1	1	1	0	0	1
D2	RT @Lupuz0503: Prestasi Ahok, Selain mjadi mafia koruptor n penista agama, dirinya bhasil tenggelamkan JKT melaui congoran Cebong RT @Lupuz0503: Ahok's achievements, apart from being a corrupt mafia and religious blasphemer, he was able to drown JKT through Cebong's congoran	1	1	1	0	0	0	0
D3	Ulama Kompak Nyatakan #HaramPemimpinKafir Pilih Ahok = Murtad! Cohesive Ulama says #HaramPemimpinKafir choose Ahok = Murtad!	1	1	0	0	0	1	1
	Number of tweets based on individual aspect label	419	111	319	670	6	5	112
	Weight for a dangerous speech aspect based on expert judgement, ω_{Cx} and ω_{My}	0.67	0.33	0.13	0.46	0.06	0.13	0.22
		$\sum \omega_{Cx} = 1$			$\sum \omega_{My} = 1$			

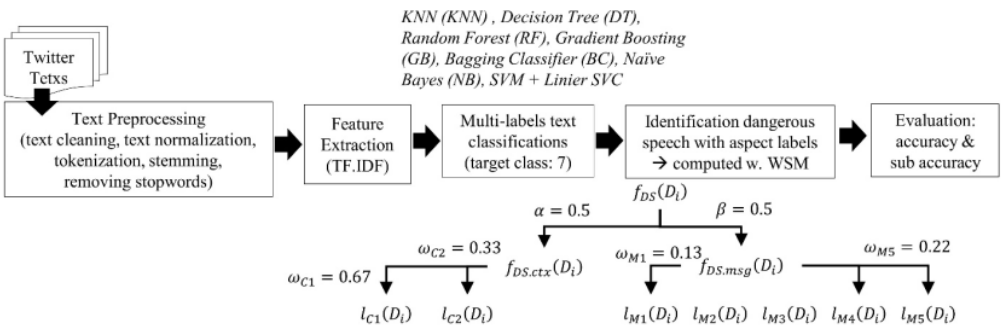

Fig. 1. The proposed procedures to identify dangerous speech with considerations on seven aspects

TABLE II. DETAILED DATA AND LABELED ASPECTS IN DATA PREPARATION

Number of Aspects	Number of data		Description
5	10	0.6%	Lacks of M4 data
4	41	4.7%	Most data have C1, C2, M1, M2
3	27	18.6%	Most data have C1 and M1
2	318	36.8%	Dominant aspects are C1-M2 or M1-M2
1	338	39.1%	Dominant aspects are M1 or M2

M4: questioning in-group loyalty, and M5: threat to group integrity or purity). Testing data (216 tweets) were annotated by a non socio-linguistic expert. Some tweets based on the annotated labels are listed in Table 1, with detail combinations are in Table 2.

It could be seen that most data have the M2 aspect of accusation in a mirror (670 out of 864). Aspects of contexts (C1, C2) and aspects M1-M2 for "dehumanization-accusation" are common in hate speech, whereas aspects M3 and M4 are not so much.

Some examples of Indonesian tweets which have been manually translated and annotated are shown in Table 1. In D2, the corruptors are perpetrators of corruption who are detrimental to the community, so they contain aspects of socio-economic context and blasphemy in socio-religious context (aspects of C1 and C2 = 1). With the animal term of "Cebong" to refer Ahok's supporters, D2 has M1= 1 and the word "Ahok" which is considered "a blasphemer of religion" contains the context of accusation in a mirror because D2 is considered as sedition (M2=1). Meanwhile D3 has "kafir" as a socio-religious context and "Haram as Kafir Leaders" from a hashtag of #HaramPemimpinKafir contain the historical context of regional head elections. In short, texts in D3 contain threats not to elect Ahok as a candidate for regional head. Let

TABLE III. ACCURACY VALUES WITH VARIOUS CLASSIFIERS AND EVALUATION METRICS

Multilabel Classifier Method	Subset Accuracy	avg. over (d)-(j)	Accuracy (acc.) C1	C2	M1	M2	M3	M4	M5	Classification (dangerous / hate)
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
ML-KNN	0.254	0.826	0.662	0.870	0.634	0.773	0.995	0.995	0.851	0.519
Decision Tree (DT)	0.351	0.816	0.708	0.865	0.671	0.601	0.995	0.995	0.879	0.727
Random Forest (RF)	0.328	0.851	0.675	0.879	0.712	**0.810**	0.995	0.995	0.888	0.718
Gradient Boosting (GB)	0.351	0.856	0.722	0.879	0.722	0.791	0.995	0.995	0.888	0.750
Bagging Classifier (BC)	0.338	0.849	0.712	**0.893**	0.722	0.736	0.995	0.995	0.888	0.732
Naïve Bayes (NB)	**0.356**	**0.858**	**0.759**	0.875	0.703	0.801	0.995	0.995	0.879	**0.773**
SVM + Linier SVC	0.300	0.847	0.675	0.888	0.685	0.791	0.995	0.995	**0.898**	0.708
Average accuracy for each label			*0.702*	*0.878*	*0.693*	*0.758*	*0.995*	*0.995*	***0.882***	

TABLE IV. AN EXAMPLE OF MISCLASSIFIED ASPECT LABELS

Tweet Text	Stemmming Words	[C1, C2, M1, M2, M3, M4, M5] *ground-truth*	Classifier Method	Predicted Aspect Label
Kenapa #2019*GantiPresiden*? Karena Peng*anggur*an Men*ingkat*? Why #2019*ChangePresident*? Because *employment* Increases?	gantipresiden anggur tingkat	[1,0,0,1,0,0,1]	KNN DT, RF, GB, BC NB, SVM	[0,0,1,0,0,0,0] [0,0,0,1,0,0,1] [0,0,0,1,0,0,0]
Gimana dgn Sumber waras.. Trans Jakarta ko jokowi diem ajah.... Harusnya si kutil babi ahok di ganyang" What about Sumber Waras.. Transjakarta, why is Jokowi silent.... Pig Ahok should have been crushed"	gaimana sumber waras trans jakarta kok jokowi diam harus kutil babi ok ganyang	[1,1,1,0,0,0,0]	BC, KNN DT NB,SVM, RF, GB	[0,0,0,1,0,0,0] [0,0,0,0,0,0,0] [0,0,1,1,0,0,0]
COPOT Darmin Nasution Menko Perekonomian BODOH yang turunkan daya beli sejak menjabat Removes Darmin Nasution Coordinating Minister for the Economy Idiot who has reduced purchasing power since taking office	copot darmin nasution menko ekonomi bodoh turun daya beli sejak jabat	[1,0,0,1,0,0,1]	BC,DT,NB,ML-KNN,RF,GB SVM	[0,0,0,1,0,0,0] [0,0,0,0,0,0,0]
#IklanAhokJahat bagaimana klo kita ganyang beneran hok?itu sih maunya elo ya hok adu domba pribumi dengan tionghoa!! #IklanAhokJahat what if we crush it for real? Is that what you want, hok, to fight the natives with the Chinese!!	iklanahokjahat bagaimana ganyang benar hok hok adu domba pribumi tionghoa jangan	[1,1,0,1,0,0,0]	NB SVM, BG,GB RF DT ML-KNN	[1,0,0,1,0,0,0] [0,1,0,1,0,0,0] [0,0,0,1,0,0,0] [1,1,1,1,0,0,0] [0,0,1,0,0,0,0]

D_i represents a Twitter text and using the seven aspect labels of contexts (C_x with $x = \{1,2\}$) and messages (M_y with $y = \{1 \dots 5\}$), then D3 has $l_{C1}(D_3) = 1$ and so on.

B. Text Preprocessing

Standard text preprocessing steps have been applied such as text cleaning, text normalization, converting text into lower cases, stemming with Sastrawi library to convert into basic words for reducing words with almost the same meaning, and removing stopwords because they have least influence in the sentences. The Twitter texts have metadata such as username, URL, RT (re-tweet), '@' character, symbols, numbers, ASCII strings, punctuations, and characters that reduce classification performance [13]. We did not remove #, and we converted words with unclear vocabulary in the text normalization. Before preprocessing, we did tokenization to divide sentences into some parts called token or words with *word_tokenize* function from *nltk.tokenize* Python library.

C. Classification and Evaluation

This work investigated algorithm of k-Nearest Neighbors (kNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Bagging Classifier (BC), Naïve Bayes (NB), Support Vector Machines (SVM), and C-Support Vector Classification (SVC or LinierSVC) to classify multi-labels in dangerous speech based on context and message.

We define dangerous speech identification based on decision-making theory. WSM is one of the most known MCDM (Multi-Criteria Decision Making) methods and the simplest one to evaluate the alternatives based on some criteria or dangerous speech aspects in the Twitter texts. It could be stated by computing aspect labels ($l_{Cx}(D_i)\ or\ l_{My}(D_i)$) and their corresponding weight values ($\omega_{Cx}\ or\ \omega_{My}$) of Twitter texts into a decision of dangerous speech or not ($f_{DS}(D_i) = \{1,0\}$ when $f_{DS}(D_i) > thres$) means transforming multi-dimensional to one-dimensional data (1). The aspects are categorized as contexts ($f_{DS.ctx}(D_i)$) in (2) and messages ($f_{DS.msg}(D_i)$) in (3). Notes that $f_{DS}(D_i) = 1$ means that D_i is a dangerous speech while $f_{DS}(D_i) = 0$ means that D_i is a hate speech. Both context and message functions could be simplified as summations of corresponding aspect weights conditioned by the existence of aspect labels after classification ($l_{Cx}(D_i) = 1\ or\ l_{My}(D_i) = 1$).

The weight values are defined by a socio-linguistic expert in data preparation step as listed in Table 1. Since both context and message are equally important aspects in defining dangerous speech, the weights of α and β are accumulated into 1. This principle applies to each context and message function as well such that $\sum \omega_{Cx} = 1$ and $\sum \omega_{My} = 1$.

$$f_{DS}(D_i) = \alpha f_{DS.ctx}(D_i) + \beta f_{DS.msg}(D_i) \tag{1}$$

$$f_{DS.ctx}(D_i) = \sum_{C_x \text{ with } x=\{1,2\}} \omega_{Cx} \times l_{Cx}(D_i) = \sum_{l_{Cx}(D_i)=1} \omega_{Cx} \tag{2}$$

$$f_{DS.msg}(D_i) = \sum_{My \text{ with } y=\{1 \dots 5\}} \omega_{My} \times l_{My}(D_i) = \sum_{l_{My}(D_i)=1} \omega_{My} \tag{3}$$

We define the threshold value as $thres = 0.55$ and $\alpha = \beta = 0.5$ indicating balance condition for both aspects of context and message. As an example for D3 in Table 1 with its aspect labels of C1, C2, M3, and M5, the computation is as follows.

$$\begin{aligned} f_{DS}(D_3) &= \omega_{C1} + \omega_{C2} + \omega_{M4} + \omega_{M5} \\ &= 0.5 \times (0.67 + 0.33) + 0.5 \\ &\times (0.13 + 0.22) = 0.675 > 0.55 \\ &\rightarrow 1\ (dangerous) \end{aligned}$$

Since the function result of $f_{DS}(D_3) = 1$ is larger than threshold, it indicates that D3 is a dangerous speech.

TABLE V. DETAILED DATA AND LABELED ASPECTS IN DATA WITH RESULT DANGEROUS SPEECH

Number of Aspects	Number of data		Description
5	5	01.3%	Most data have C1,C2,M1,M2,M5
4	41	11%	Most data have C1, C2, M1, M2
3	162	43.6%	Dominant aspect are C1-M1-M2 (16.7%) and C1-M2-M5(9.4%)
2	171	45.8%	Dominant aspects are C1-M2

Based on the entire dataset of 864 instances, it shows that 371 data are dangerous speech data and 493 data are hate speech data. Based on the data, it shows that dangerous speech sentences are influenced by C1>M2>M1>C2. Of the 371 tweets classified as dangerous speech, it shows that 100 percent contain C1, indicating that this aspect is very important in the identification of dangerous speech, 24% contain C2, 26.9% contain M1, 97% contain M2, 10% contain M3 and M4, and 16.9% contain M5. Therefore, in the next research, the method must be able to recognize C1 and M2 well, because the average accuracy of C1, M1 and M2 is still quite low.

Unlike the evaluation of single-label classification models, our multi-label classification problem applies instance-based and label-based metrics. The instance-based metric is totaled by averaging all test data, referring to a standard accuracy. The label-based metric or subset accuracy is computed for each label and averaged over all labels. Subset accuracy evaluates the fraction of correctly classified test data, i.e., whether the predicted label set is identical to the ground-truth label set. Intuitively, subset accuracy tends to be overly strict, especially when the size of the label space is enormous [11].

III. RESULTS AND DISCUSSION

Based on previous works on hate speech, we investigated some classifiers with kNN as a baseline as listed in column (a) of Table 3. After creating classifier models with training data (648 instances), the testing results of 216 tweets are shown in Table 3. Columns (d)-(j) show single-label accuracies for each aspect, i.e., the accuracy for C1 labeled in 216 instances using Naïve Bayes is ± 76%. Column (c) displays the averaged values of single-label accuracies in (d)-(j) using a specific classifier, while the last row in Table 3 shows accuracies being averaged across different classifiers in each aspect. Although M3 and M4 have higher accuracy (0.995), the corresponding data in Table 1 for those aspects showed an imbalanced condition (6 and 5 instances or less than 1% from 864 tweets), thus making their accuracy questionable. Aspects of C1 and M1 have low accuracy, indicating difficulties in recognizing social contexts and dehumanization because of their specific terms, i.e., the term of unemployment or "pengganggguran"(*unemployed*) in Indonesian was stemmed

into “anggur”(*grapes*) which makes the classifier model fails to label the text with social context (Table 4). This is because words containing social context are quite broad and more unpredictable due to growing social cases such as “sumber waras”. “Sumber waras” is the name of the corruption case that occurred in Indonesia. In addition, the following words such as “turun daya beli” (*decreased purchasing power*), “ibu pertiwi diporakporandakan”(*country destroyed by*), “adu domba pribumi” (*bring into conflict*), etc. These words didn’t exist during the training data process. In addition, the method is also less able to recognize words that contain aspects of dehumanization, especially words that contain dehumanizing aspects but use words other than animals as insults such as “kacung”(*lackey*), “bangsad”(*brock*), “najis”(*unclean*), “iblis betina”(*devil female*), “jamban”(*latrine*), etc. These words are not recognized because the words do not exist during the training process. Various results of classifier models have showed that Naïve Bayes performs better in C1 (acc.= 0.759), Bagging Classifier in C2 (acc.= 0.893), while Gradient Boosting or Bagging Classifier are for M1.

Subset accuracy in column (b) of Table 3 ensures that all labeled aspects of an instance are similar to the labels defined in the ground-truth. The lowest accuracy value is resulted from the baseline classifier kNN. An example of four misclassified aspects in Table 4 supports that kNN classifier has performed worst. The wrong labels are marked with red-bold fonts by comparing the ground-truth and the predicted results.

Values of the last column (k) in Table 3 compare the final label of dangerous or hate after WSM computation. As mentioned before, with the threshold value of 0.55, Naïve Bayes is more reliable to identify dangerous speech of Twitter texts based on the classification accuracy of 0.773.

Although our experiments have been performed using rather limited data, in general Naïve Bayes is more suggested from the evaluation metrics (Table 3) of label-based subset accuracy (±36%), instance-based (*average*) accuracy (±86%) and classification accuracy (±77%), showed in column (b), (c), and (k) respectively.

IV. Conclusion

This paper delivers a procedure to identify dangerous speech of Twitter texts based on aspects rather than a binary class of hate speech. Because of multi-aspects with different weights to support dangerous speech, the proposed procedure includes a weighted sum model to transform multi-dimensional aspects within texts into a one-dimensional label of dangerous or (plain) hate speech.

Because of imbalanced data within each aspect, future works will accommodate some appropriate measures such as finding more data in certain aspects and proper feature extraction like avoiding stemming or extending the vocabulary. With deep learning usage, our procedure will also incorporate some popular architectures such as BiLSTM, CNN, or BERT and their variants.

References

[1] J.L. Maynard & S. Benesch, “Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention”, *Genocide Studies and Prevention: An International Journal*, Vol. 9, No. 3, pp. 70-95, 2016.

[2] S. Benesch, T. Glavinic, S. Manion, & C. Buerger, Dangerous Speech: A Practical Guide, *the Dangerous Speech Project*, Apr. 2021, Accessed on: Nov. 2, 2021. [Online]. Available: https://dangerousspeech.org/guide/

[3] T. Davidson, D. Warmsley, M.W. Macy, & I. Weber, “Automated hate speech detection and the problem of offensive language”, *In: Proc. of the Eleventh International Conference on Web and Social Media (ICWSM 2017)*, Montréal, Canada, pp. 512–515, 2017.

[4] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, & D. Yeung, “Multilingual and multi-aspect hate speech analysis”, *In: Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 4674–4683, 2019.

[5] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M.R. Pardo, P. Rosso, & M. Sanguinetti, “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”, *In: Proc. of the 13th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2019)*, Minneapolis, USA, pp. 54–63, 2019.

[6] A.M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, & N. Kourtellis, “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior”, *In: Proc. of the Twelfth International Conference on Web and Social Media (ICWSM 2018)*, California, USA, pp. 491–500, 2018.

[7] M.O. Ibrohim & I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter”, *In: Proc. of the Third Workshop on Abusive Language Online (ALW3)*, Florence, Italy, pp. 46–57, 2019.

[8] R. Hendrawan, Adiwijaya, & S. Al Faraby, “Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media”, *In: Proc. of the 2020 International Conference on Data Science and Its Applications (ICoDSA)*, Bandung, Indonesia, pp. 1–7.

[9] I. Gunasekara & I. Nejadgholi, “A Review of Standard Text Classification Practices for Multi-label Toxicity Identification of Online Content”, *In: Proc. of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium, pp. 21–25, 2018.

[10] W. Liu, J. Pang, N. Li, X. Zhou, & F. Yue, “Research on Multi-label Text Classification Method Based on tALBERT-CNN”, *International Journal of Computational Intelligence Systems*, Vol. 14, pp. 201, 2021.

[11] I. Alfina, R. Mulia, M.I. Fanany, & Y.Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study”, *In: Proc. of the 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Bali, Indonesia, pp. 233–238, 2017.

[12] G. Kumar & N. Parimala, “A weighted sum method MCDM approach for recommending product using sentiment analysis”, *International Journal of Business Information Systems*, Vol. 35, No. 2, pp. 185-203, 2020.

[13] I.A.F. Hidayatullah & M.R. Ma'arif, “Pre-processing Tasks in Indonesian Twitter Messages”, *In: Proc. of the 2016 International Conference on Computing and Applied Informatics*, Medan, Indonesia, 2016.

[14] Atikah L, Purwitasari D & Suciati N, “Deteksi Kejadian Lalu Lintas pada Teks Twitter dengan pendekatan Klasifikasi Multilabel berbasis Deep Learning”, In: *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK),* Vol. 9, No. 1, Februari 2022, hlm. 87

Aspect Based Multilabel Text Classification for Identifying Dangerous Speech Twitter Text

ORIGINALITY REPORT

13%	9%	7%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	github.com Internet Source	2%
2	www.semanticscholar.org Internet Source	1%
3	Lecture Notes in Computer Science, 2015. Publication	1%
4	Submitted to University of Greenwich Student Paper	1%
5	Dini Yuniasri, Siti Rochimah, Agus Budi Raharjo. "A Correlation Analysis Between ISO 25010 based Modularity and CK Metrics in Object-Oriented Software", 2020 International Conference on Advanced Science and Engineering (ICOASE), 2020 Publication	1%
6	Submitted to University of Auckland Student Paper	1%
7	web.archive.org Internet Source	1%

8	Faizal Adhitama Prabowo, Muhammad Okky Ibrohim, Indra Budi. "Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter", 2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), 2019 Publication	1 %
9	repository.uin-suska.ac.id Internet Source	1 %
10	Nawal Aljedani, Reem Alotaibi, Mounira Taileb. "Multi-Label Arabic Text Classification: An Overview", International Journal of Advanced Computer Science and Applications, 2020 Publication	<1 %
11	"Table of Contents", 2022 10th International Conference on Information and Communication Technology (ICoICT), 2022 Publication	<1 %
12	Submitted to Laredo Community College Student Paper	<1 %
13	repository.ub.ac.id Internet Source	<1 %
14	www.whoopy.it Internet Source	<1 %

15	A F Hidayatullah, M R Ma'arif. "Pre-processing Tasks in Indonesian Twitter Messages", Journal of Physics: Conference Series, 2017 Publication	<1 %
16	Diana Purwitasari, Dini Adni Navastara, Yulian Findawati, Kresna Adhi Pramana, Agus Budi Raharjo. "Feature Extraction in Hierarchical Multi-Label Classification for Dangerous Speech Identification on Twitter Texts", 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), 2023 Publication	<1 %
17	academic-accelerator.com Internet Source	<1 %
18	Submitted to Liverpool John Moores University Student Paper	<1 %
19	Y Findawati, I N Hikmah, S Sumarno, Y Rachmawati. "HPAI consumer shopping analysis using Apriori algorithm", IOP Conference Series: Materials Science and Engineering, 2021 Publication	<1 %
20	Fadilla Sukma Alfiani, Imamah, Umi Laili Yuhana. "Categorization of Learning Materials Using Multilabel Classification", 2021	<1 %

International Conference on Electrical and Information Technology (IEIT), 2021
Publication

21	Ruangsung Wanasukapunt, Suphakant Phimoltares. "Classification of Abusive Thai Language Content in Social Media Using Deep Learning", 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2021 Publication	<1 %
22	urss.knuba.edu.ua Internet Source	<1 %
23	Karimah Mutisari Hana, Adiwijaya, Said Al Faraby, Arif Bramantoro. "Multi-label Classification of Indonesian Hate Speech on Twitter Using Support Vector Machines", 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020 Publication	<1 %
24	mafiadoc.com Internet Source	<1 %
25	mdpi-res.com Internet Source	<1 %
26	Damayanti Elisabeth, Indra Budi, Muhammad Okky Ibrohim. "Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study",	<1 %

2020 8th International Conference on Information and Communication Technology (ICoICT), 2020
Publication

27 Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga. "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions", Computer Science Review, 2020 <1%
Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		