Detection of Criticism and Hate speech Text Formulation on Online Social Network Twitter for Semantic Recommendation System Framework

Migunani[1,*] Adi Setiawan[2] Irwan Sembiring[2]

[1]*Information System, Science and Computer Technology University, 50195, Indonesia*
[2]*Information Technology, Cristian Satya Wacana University, 50715, Indonesia*
[*]*Corresponding author. Email: migunani@stekom.ac.id*

ABSTRACT
User opinions on high-volume social media and various themes provide relevant information for sentiment analysis. This information can be collected and analyzed using a natural language processing with a monitoring system to support classification of criticism and hate speech. Regarding monitoring results, a knowledge-based recommendation system with sentiment analysis is supported to send messages to user in order to use positive sentences are not offensive, polite, wise and motivational for users with hateful attitudes. It is important to formulate sentences that can differentiate between criticism and hate speech. By compiling a formula sentence as a classification reference for the text obtained in a twitter tweet whether as criticism or including hate speech. Detection of sentences containing criticism and hate speech using Bag of Word and Convolutional Neural Network to detect hate speech dan criticism sentence via Twitter. The detection results are used for the semantic recommendation system framework that includes sentiment analysis and classification of hate speech.

Keywords: *Text Formulation, Criticism, Hate Speech, Bage of Word, Convolutional Neural Network, Recommendation System.*

1. INTRODUCTION

Currently, active users on social media are growing rapidly. In a study of social media communication indicating that by the end of 2020, online social network users will be 2.95 billion users [1]. The high number of Online Social Network (OSN) users is influenced by the increase of mobile devices users (smartphones and tablets). Social media are now used to express opinions and feelings. Sentences and notes contain sentiments and emotions expressed in social media messages and give insights into the behavior of users of social media. For example, phrases with negative meaningful words could indicate criticism, even hatred utterance, as far as slander is concerned. On the other hand, if someone is in a positive feeling, they will be more confident with emotional stability [2]. The sentiment of social media user has different intensities from low, medium, high level or it changes from high to low level or vice versa. Many online forums like Facebook, YouTube, and Twitter therefore regard hate speech as harmful and have policies to delete hate speech content [3,4,5]. Due to societal concerns and how widely hate speech is becoming on the Internet [6], the study of the automated detection of hate speech is strongly motivated. The spread of hateful content can be reduced by automating its detection.

1.1. Hate Speech

However, social media and other online communication methods have begun to play a greater role in hate crimes. In the UK, hate speech against migrants and Muslim communities has increased significantly in response to recent events, including leaving the EU, Manchester and the London attacks [7]. In New Zealand, suspects have a long history of hate posts in various hate-related terror attacks, suggesting that social media contributes to their radicalization. Social media can play a more direct role in some cases: video footage from the terror attack suspect in Christchurch, New Zealand in 2019, was broadcast live on Facebook [8]. Although the results of some recent approaches for detecting hate speech in textual content have been promising [9,10,11]. The solutions proposed use profound learning techniques to classify text as hate

speech. One limitation of these approaches is that their decisions can be opaque and difficult for people to understand why the decision was made.

Several works on hate speech detection have been used in English in machine learning and in twitter in English [12][13][14]. In the meantime [12] aims to compare features appropriate for detecting hate speech in English. The main features are word n-gram and character n-gram. Word n-gram feature used for unigram and bigram combination. Gender and location are also used as additional features, and BLR has been selected as the classification algorithm. The results showed that n-gram character outperforms the n-gram feature for hate speech detection, in English, with a 10% accuracy difference. In addition, the gender and location had no effect on the performance of the algorithm classification.

In [15] research on the detection of hate speech against religion the characteristic in Indonesian were word unigram and bigram, the number of hateful words and phrases and the number of words with a negative feeling but not the template of phrases. The comparable algorithms were naive bays and vector machine support. The dictionary is also built to count the number of words or phrases related to hate speech. Because of the unbalanced number of religious tweets and not in the class of non-hate-speech-against-religion, the resulting dictionary was not very well suited for the words relating to religion, but more for the dictionary of hate.

Referring to the Big Indonesian Dictionary, the meaning of criticism is sometimes good and bad considerations of a work, opinion and so on. Criticism conveyed is actually in order to improve someone's opinion or behavior. It is different from hate speech, slander and insults which are carried out with offensive narratives. Even disrespectful and unwise and not intended to correct someone's opinion or behavior. On the contrary, it is not based on hatred for the person. Criticism is done by using a choice of words that are not offensive, polite and wise. Since April 21, 2008, expressions of hatred carried out on social media have been regulated in article 45 paragraph (2) junto article 28 paragraph (2) of law number 11 concerning on Electronic Transaction Information constitution 2008 in Indonesia [16].

Sometimes, someone can accidentally say or write a sentence in a written form that contains hate speech. Distinguishing whether it is criticism or hate speech requires boundaries that the sentence is criticism or has entered the area of hate speech. The need for understanding between criticism and hate speech will save someone from the complaint offense. Hate speech is any communication that disparages a person or group on the basis of several characteristics such as race, ethnicity, gender, sexual orientation, nationality, religion or other characteristics [17].

The base word for disparagingly related to hate speech at [17], is an adjective in Indonesian. According to the Language and Development Agency, Ministry of Education and Culture of the Republic of Indonesia, the word trivial has the same meaning of insult as despicable, weak, lecherous, low, and tainted. So that sentences with the meaning of insulting, weakening, humiliating, criticizing and defaming are included in hate speech. Meanwhile, the word criticism is a noun in Indonesian which means criticism or response, or a parcel which is sometimes accompanied by descriptions and good and bad considerations of a work, opinion, and so on. In the Indonesian thematic thesaurus criticism is an opinion or a warning.

1.2. Semantic Recommender System

The use of profiles to represent users' long-term information needs and interests is a common feature in most semantic recommendation schemes. User profiles have therefore become a key part of efficient filtering of recommended systems, as a poor profile can lead to poor quality and irrelevant user advice. Integrating machine-learning with ontology-based recommendation systems has led to many remarkable revolutions when it comes to improving recommendation system processes [18, 19].

In [20] a recommender based on travel content was proposed that uses ontology information to calculate the degree of similarity between user preferences and interest in providing individualized recommendations. The system uses machine training techniques to generalize user preferences. In addition, the authors proposed an ontology-based recommendation system with rules that were developed through data mining techniques in paper [21]. Teachers can use this system to predict the progress and performance of learners. The results were very satisfactory, and the rules produced were checked with a small error forecast. The combination of ontology-based recommendations with machine training techniques is a promising approach for enhancing the accuracy of recommendations.

Semantic recommendation system for social media users who are detected as committing hate speech. The detection results of social media users use a machine learning approach as a state for the recommendation system to be processed in order to produce recommendations for hate speech actor in the form of suggestions, advice, or treatments so that they no longer make hate speech.

2. METHOD

In this section, we will discuss datasets and methods for classifying criticism and hate speech using a deep learning approach.

2.1. Formulation of Criticism and Hate speech Sentence.

Text on social media that twitter will be analyzed using a deep learning approach. The text obtained will be distinguished whether the text is an utterance of criticism or an utterance of hatred. At this stage we propose a new formulation of sentence structure, especially in Indonesian. The sentence structure will be separated into four parts, namely the word criticism or hate speech, the object that is the purpose of the word, a complementary explanation and the part that is considered why the text was written or spoken. The compilation of hate speech text or sentences is based on hate speech statement [1] by adding new sentence structures not as hate speech but as critical utterances with different word choices.

2.2. The Dataset

The sentence template and data set process consist of three main steps, creating a sentence criticism and speech template, collecting the dataset.

2.2.1. Data Collection

Using Twitter Token Access, Twitter data is obtained. The Twitter data relates to race, ethnicity, gender, sexual orientation, nationality, religion. Why are the data related to this word used because of its close relationships between critique and hate speech?

2.2.2. Data Labelling

Whether it contains critical or hate speech, tweet data on the dataset will be tagged. There are two labels, if the phrase with critique is labeled "critical" and those with hate are labeled "hate."

2.3. Hate Speech Detection

Our research aims to compare features and professional algorithms to see which features and algorithms are the best performing combinations. Three phases of critical and hate speech detection are: 1) prerecession, 2) extraction of features and 3) classification and assessment.

2.3.1. Preprocessing

In the preprocessing stage are five steps, i.e. 1) dividing tokens in white space, 2) removing all splitting from words. 3) remove all words which do not consist exclusively of alphabetical characters. 4) Remove all words with the elimination of unnecessary and meaningless words that are known stop words. 5) Remove all words of less or equal length to a single character.

2.3.2. Features Extraction

A model bag-of-words is used to extract text features. Text input can be used with neural network algorithms. The document is converted into a vector representation in this case. The number of items in the vector of the document is equivalent to the number of words in the vocabulary. The bigger the vocabulary, the longer the vector representation, the less vocabulary is preferred for this section. Words are marked in a document and results are displayed in the corresponding location. The purpose of this section is to convert reviews into vectors prepared for training a first neural network model. This section is divided into turning reviews into token lines and encoding reviews with a representation of the Bag of Words model.

2.3.3. Classification and Evaluation

In this step we will develop multi-layer perceptron models that can be either positive or negative for encoded documents. The models will be simple feedforward network models in the deeper learning library of kera with fully connected layers called dense. This section consists of 3 sections: 1) first sentiment analysis model, 2) compare text scoring modes, 3) predict new assessments.

Coevolutionary neural networks is an architecture of multi-stage neural networks developed for classification [23]. Each phase consists of a layer type:

a. Convolutional layer as the main component of CNN.

b. Pooling Layers as an integral component of CNN.

c. Embedding Layer is a special text classification problems CNN component.

d. Fully Connected Layer is a classic hidden layer Feed-Forward Neural Network (FNN).

CNN is used to distinguish sentences or documents for hate speech classification based on the word that has been trained. Why use CNN because it can select features that stand out in a different way by their position in the input sequence.

3. RESULT AND DISCUSSION

In this section we explained about result and analysis.

3.1. Results

The word hate in the vocabulary of hate speech in Indonesian has a basic meaning, namely very dislike. Simply put, for example the sentence "I don't like you" is not categorized as hatred yet. But if you add the warmth of "I really don't like you", is this an explicit hate speech? The word hate in Indonesian is distinguished into three types of words, whether it is a verb, adjective and word or noun. Table 1 Show types of Indonesian words contain hatred.

Warning statements in Indonesian can be a number of words that mean exaggerating or there are additional words that mean to warn. The word as the smallest language unit stands alone. Phrases as a combination of two or more words without a predicate. At least one clause as a grammatical unit of a group of words consists of a subject and a predicate and is capable of turning into a sentence. Sentence is a relatively independent language unit with a final intonation pattern and is actually or possibly made up of clauses. The difference between criticism or hate speech can be formulated in Tables 2 and 3.

Words, clauses or sentences in a critical specification can be distinguished based on the choice of words, the service for what the word is, the explanation and the existence of good and bad considerations.

We propose a formula for word composition, clauses or sentences that contain critical speech based on Table 2, which is:

critical speech = *word(a-e)* + *intended(a-f)* + *explanation(pos)* + *consideration (good, bad.* Figur 1 show critical speech sentence formula.

Table 1. Types of Indonesian words contain hatred

Kinds of words	Words contain hate
Verb	hurt, shame, hate, disgust, insult, hurt, offend, feel disgusted, demeaning, scolding, hating, annoying, upsetting, disappointing, irritating, pathetic, choking, grumbling, angry, cursing, moaning, complaining, mumbling, cursing, wailing, whining.
Adjective	arrogant, nauseated, bored, fed up, talkative, nagging, acting up, bitchy, wide mouth, itchy mouth, cynical, cranky, annoyed, angry, fussy, grumpy, disgusted.
Nominal	antipathy, grudge, hatred, confusion, anxiety, jealousy, restlessness, anger, envy, worry, panic, resentment, hostility, enmity, grumpy, irritable, grumbler, violent, high blood, fierce, ferocious, displeasure, bitterness, disappointment, irritation, dissatisfaction, bitterness, anger, rage, mumbling resentment, complaint, scolding, moaning of grief, concern.

Table 2. Formulation of critical speech

Critical speech			
Word	Addressed	Explanation	Consideration (opsional)
a) Warning	a) Race	significantly positive	Good and bad
b) Criticism	b) Ethnicity		
c) Advice	c) Gender		
d) Responses	d) Sexual orientation		
e) Peeling	e) Nationality		
	f) Religion		

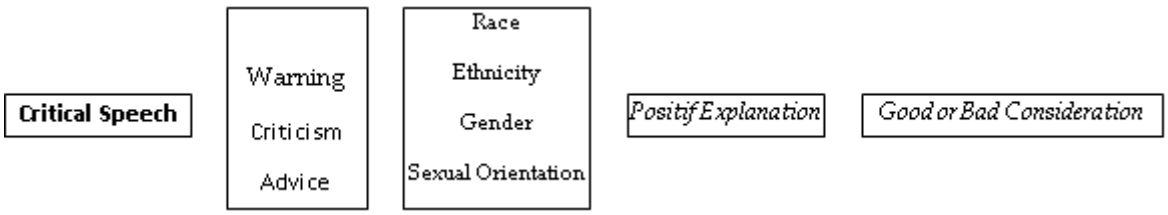

Figure 1 Critical Speech sentence formula

- critical speech: meaningful sentences are complete speech criticism
- word (a-e): words are contained in the sentence, namely: a reprimand, criticism, advice, and comments.
- Intended (a-f): the words contained in the sentence are intended for: race, ethnicity, gender, sexual orientation, nationality, and religion.
- Explanation (pos): there is additional explanation in sentences that are positive.
- Consideration (good, bad): There is good and bad judgment in the sentence.

Word order, clauses or sentences in hate speech can be differentiated based on the choice of words, what the word is intended for, its explanation and the absence of good and bad considerations.

We propose a formula for word composition, clauses or sentences containing critical speech based on table 3, namely:

hate speech = *word(a-g) + intended(a-f) + explanation (pos, none) + consideration (none).* Figure 2 show speech sentence formulas

Tab**le 3.** Formulation of hate speech

Hate Speech			
Word	Addressed	Explanation	Consideration (opsional)
a) Underestimating	a) Race	a) Has a negative meaning	Nothing
b) Reproach	b) Ethnicity	b) Nothing	
c) Insult	c) Gender		
d) Weakening	d) Sexual orientation		
e) Defamatory	e) Nationality		
f) Degrading	f) Religion		
g) Contaminating			

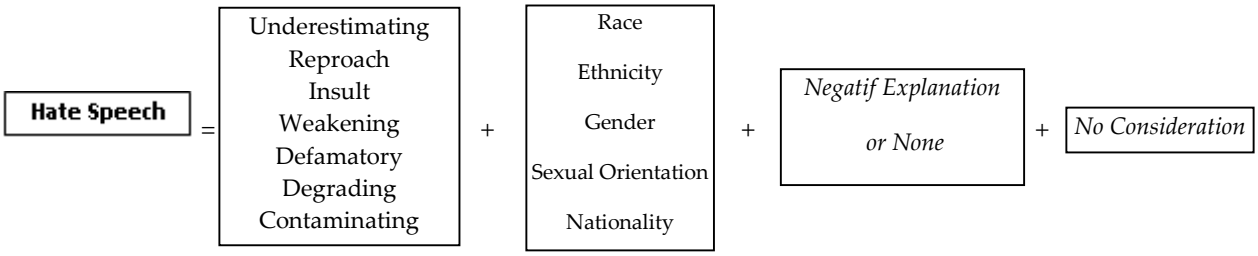

Figure 2 Hate Speech sentence formulas

- Hate speech: The sentence means hate speech in full sentence
- word(a-f): The words contained in the sentence, namely: belittling, reproaching, insulting, weakening, demeaning, and defaming
- intended(a-f): The words contained in the sentence are intended for: race, ethnicity, gender, sexual orientation, nationality, and religion.
- explanation (neg, none): There are additional explanations in sentences that have negative meanings or no explanation at all.
- consideration(none): The absence of good and bad judgment in the sentence.

The classification of critical speech and hate speech can be distinguished based on the words used in the sentence with a combination of the addition of these words aimed at race, ethnicity, gender, sexual orientation, nationality, and religion. If the sentence with the explanation is positive and there are good and bad considerations, it is considered critical. While sentences with negative explanations or no explanation and consideration are classified as hate speech.

Combining between machine learning approaches and semantical or ontological recommender systems for detection and treatment of social media users, especially Twitter, are realized in the framework of "treatment for hate speech". We propose a "treatment for hate speech" framework as a way to reduce or suppress the appearance of hate speech on social media, especially Twitter. The following is the “treatment for hate speech” framework. Figure 3 shown Semantic Recommendation System Framework for Hate Speech.

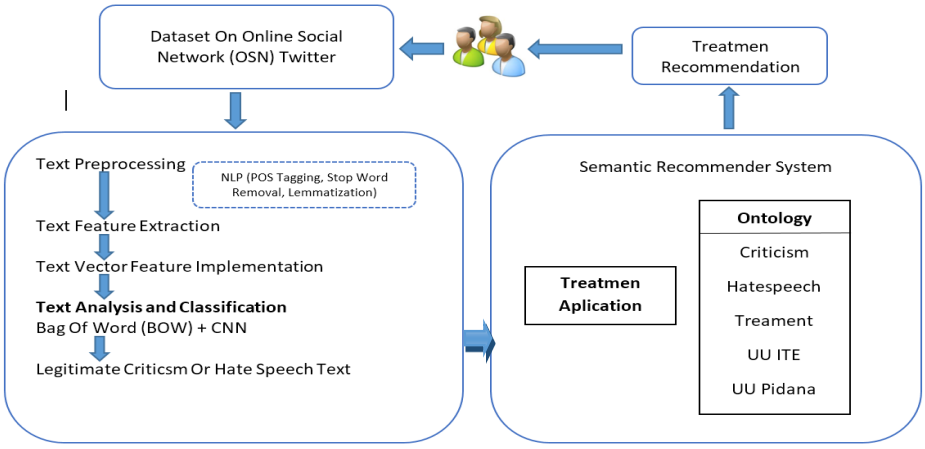

Figure 3 Semantic Recommendation System Framework for Hate Speech

The stages in this work platform are, 1) The dataset obtained through Twitter social media will be prepared for processing using a machine learning approach, 2) The dataset will go through the text preprocessing stage using: White space split tokens, Remove from words all punctuation. Remove all words which do not consist solely of alphabetical characters. Remove all known stop words words words. Remove words that are unnecessary and meaningless. Remove all words of less than or equal length to a single character. 3) Text analysis and classification using Bag of Word and Convolutional Neural Network. 4) Build the ontology for semantics recommender system, 5) build treatment applications. 6. Recommended treatments for hate speech actor.

3.2. Discussion

Distinguishing between criticism and hate speech based on sentence formulation is important to be considered in determining whether the sentence is criticism or has entered the area of hate speech which can have an impact on the realm of law. Critical sentences become the basis for improvement for a condition to be better and lead to something ideal or in accordance with what should be. In contrast to hate speech sentences which can become hostile, hateful, and can damage relationships among humans, race, ethnicity and religion. Classifying criticism and hate speech can help the process of treating or treating hate speech actor not to do so or reducing it gradually through social media-based applications.

A recommendation system with symbols combined with ontology for the treatment or treatment of hate speech is required after the classification stage which results in sentences containing hate speech. The meaning and meaning of hate speech according to [17] so that it can be classified as a formula needed to compile sentences containing hate speech or only in the form of critical sentences. The text formulation will be the syntactic template used to classify sentences. Normalizes data by removing stop words, with the exception of "no", special markers such as retweet and screen names, as well as punctuation marks [12]. Collect unigram, bigram, trigram, and four grams for each tweet and user description. To assess the informativeness of the features, we summed the model coefficients for each feature over 10 times cross validation. This allows for stronger estimates.

We use a different approach, namely by detecting words in twitter tweets that limit criticism or hatred by comparing the sentence formulas. Detection is done by looking for the presence of critical words and hate speech words on Twitter. Figure 4 show checking for critical or hate speech.

After preprocessing the text from Twitter, keywords of criticism or hate speech will be searched according to their respective characteristics. Next, we will look for words related to previously found words of criticism or hate speech. Identify the second word which becomes a statement that the previous word is intended for criticism or hate speech. In addition, the explanation of the sentence will be examined if there is an explanation in the sentence if it is positive or negative. If the explanation is positive then the sentence and there are words that describe a good or bad consideration, the sentence is included in the classification of criticism. Conversely, if the word has a negative meaning and there is absolutely no consideration, it can be classified as hate speech. At level-1 the formula only contains two words, the first word is a word that includes criticism or hate speech, while the second word is the word that is the object of criticism or hate speech. At level-2 the sentence is equipped with an explanation, if it is positive, it will increase the belief that the tweet sentence is criticism, on the contrary, if the explanation is negative, the tweet is hate speech. At level-3, if the sentence is equipped with good or bad considerations, the tweet sentence is a

criticism, on the contrary, if there is no consideration, it is categorized as hate speech.

The novelty in this research is a method or way of classifying sentences containing criticism or hate speech based on a text formulation consisting of four parts and four stages of checking sentences, especially in Indonesian, the results of which can be classified as critical sentences or hate speech sentences. Even though it only consists of two parts, namely at level-1 short sentences can be classified. At level-2 and level-3 sentences become completer and more become easier to classify.

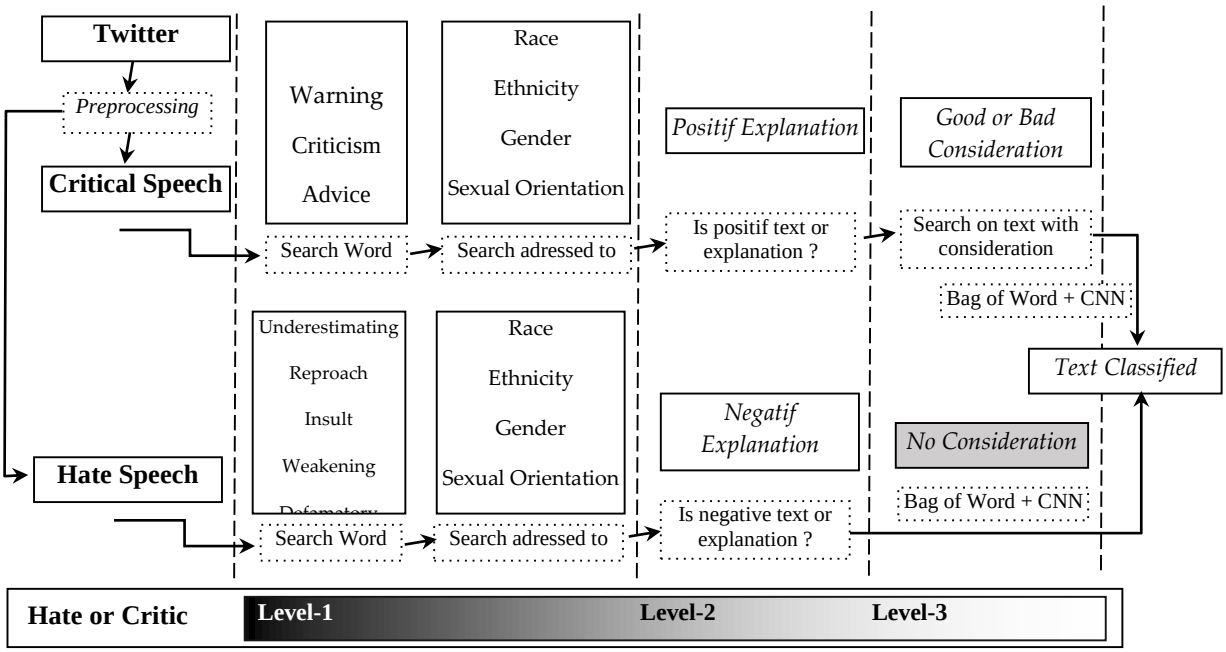

Figure 4 Checking for Critical or Hate Speech

4. CONCLUSION

A new method in detecting a sentence whether it is included in the criticism category or it is included in the category of hate speech. The sentence structure formula can be divided into four parts to determine that the sentence means criticism or hate speech. The four sections are divided into (1) words that represent criticism or hate speech such as warning, critics, advice, response, peeling, underestimating, reproach, insult, weakening, defamatory, degrading and contaminating. (2) Words for critique or hate, such as race, ethnicity, gender, sexual orientation, nationality and religion. (3) words of explanation that can strengthen, if it is positive then criticism is the opposite if it is negative then hate speech. (4) optional consideration parts. Sentences with good or bad considerations are categorized as criticism, whereas without consideration at all can include hate speech.

ACKNOWLEDGMENTS

Thanks to the university of science and computer technology, especially the foundation that has funded the study and preparation of this research aimed at completing the dissertation. Thanks also to those who have helped in the preparation of this article for the dissertation plan.

REFERENCES

[1] I.-R. Glavan, A. Mirica, and B. Firtescu, “The use of social media for communication.” Official Statistics at European Level. Romanian Statistical Review, vol. 4, pp. 37–48, Dec. 2016.

[2] I. B. Weiner and R. L. Greene, “Handbook of personality assessment,” in John Wiley and Sons, N.J, EUA, 2008.

[3] Community Standards;. Available from: https://www.facebook.com/community standards/objectionable_content.

[4] Hate speech policy-YouTube Help;. Available from: https://support.google.com/ youtube/ answer/2801939.

[5] Hateful conduct policy;. Available from: https://help.twitter.com/en/rules-and-policies/hateful-conductpolicy.

[6] Mondal M, Silva LA, Benevenuto F. A Measurement Study of Hate Speech in Social Media. In: ACMHyperText; 2017.

[7] Guardian. Anti-muslim hate crime surges after manchester and london bridge attacks, Last accessed: July 2017, https://www.theguardian.com.

[8] The New York Times. New Zealand Shooting Live Updates: 49 Are Dead After 2 Mosques Are Hit. 2019.

[9] Fortuna P, Nunes S. A Survey on Automatic Detection of Hate Speech in Text. ACM Comput Surv. 2018; 51(4):85:1–85:30. https://doi.org/10.1145/3232676

[10] Davidson T, Warmsley D, Macy MW, Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. ICWSM. 2017;.

[11] Zimmerman S, Kruschwitz U, Fox C. Improving Hate Speech Detection with Deep Learning Ensembles. In: LREC; 2018.

[12] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," Proc. NAACL Student Res. Work, pp. 88—93,2016.

[13] I. Kwok and Y. Wang, "Locate the Hate: Detecting Tweets against Blacks," Twenty-Seventh AAAI Conf. Artif. Intell., pp. 1621-1622, 2013.

[14] P. Bumap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," Policy and Internet, vol. 7, no. 2, pp. 223-242, 2015.

[15] S. H. Pratiwi, "Detection of Hate Speech against Religion on Tweet in the Indonesian Language Using Naive Bayes Algorithm and Support Vector Machine," B.Sc. Tesis, Universitas Indonesia, Indonesia, 2016.

[16] Laws of the republic of Indonesia number 19 in 2016 on information and electronic transactions (UU ITE).

[17] W. Warner and J. Hirschberg, 'Detecting hate speech on the world wide web," Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media, no. Lsm, pp. 19-26,2012.

[18] Dou, D. et al. 2015. Semantic data mining: A survey of ontology-based approaches. Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015) (Feb. 2015), 244–251

[19] Ristoski, P. and Paulheim, H. 2016. Semantic Web in data mining and knowledge discovery: A comprehensive survey. Web Semantics: Science, Services and Agents on the World Wide Web. 36, (Jan. 2016), 1–22. DOI:https://doi.org/10.1016/j.websem. 2016.01.001.

[20] Bahramian, Z. and Abbaspour, R.A. 2015. An Ontology-Based Tourism Recommender System Based on Spreading Activation Model. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 15, (Dec. 2015), 83–90. DOI:https://doi.org/10.5194/isprsarchives-XL-1-W5-83-2015

[21] Boufardea, E. and Garofalakis, J. 2012. A Predictive System for Distance Learning Based on Ontologies and Data Mining. (Jul. 2012), 151–158.

[22] Pang, Bo, Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using SubjectivitySummarization Based on Minimum Cuts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics-ACL doi:10.3115/ 1218955.1218990, In EMNLP, pages 79–86.

[23] S. V Georgakopoulos and V. P. Plagianakos, "Convolutional Neural Networks for Toxic Comment Classification," Comput. Sci. Biomed., 2018

[24] Zhang, Xiang, Junbo Zhao, and Yann LeCun. "Character-level convolutional networks for text classification." Advances in Neural