Logical framework for hate speech detection on religion issues in Indonesia

W Darmalaksana1,*, F S Irwansyah2, H Sugilar3, D S Maylawati4,6, W D I Azis5 and A Rahman4

1 Department of Ilmu Hadits, UIN Sunan Gunung Djati Bandung, Indonesia
2 Department of Chemistry Education, UIN Sunan Gunung Djati Bandung, Indonesia
3 Department of Mathematics Education, UIN Sunan Gunung Djati Bandung, Indonesia
4 Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia
5 Research and Publication Centre, UIN Sunan Gunung Djati Bandung, Indonesia
6 Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

*yudi_darma@uinsgd.ac.id

Abstract. Hate speech is a hot issue in social media era, including in Indonesia. This paper presents the concept of logical framework to analyse and detect whether the content from text in social media contains hate speech on religion issues or not. This paper use literature review to propose the logical framework for hate speech detection. We propose the text mining as a part of data mining technique that combine with sequential pattern mining to detect hate speech from social media with text data. Based on the proposed logical framework to detect hate speech from social media, it is possible and recommended to be implemented in the future research to prevent hate speech distribution which can cause disunity and commotion in the community.

1. Introduction

Hate speech is a contemporary issue that arises from the wrong use of social media technology. The hate speech has multiple interpretations, there is no standard and general definition that explains the hate speech. From a legal point of view, hate speech can mean speech that only expresses hatred, there is no other meaning other than hate, towards a certain group, such as race, even under certain conditions can trigger violence [1]. Although expressions of hatred today are still exposed to a variety of interpretations, both international law, the law in force in Indonesia, and the law in every religion, any form or action that promotes hatred on the basis of nationalism, race or religion that causes hostility and violence is prohibited. Different countries apply laws for hate speech with different approaches. The United States upholds freedom of speech, but opinions that are allegedly expressions of hatred will become a criminal offense if they have been proven to result in threats of violence [2]. In contrast to Germany which has completely banned the expression of hatred which is a form of threat [3]. In fact, the United Nations Human Rights Commission recommends that each country clearly define the hate speech [2]. Where, utterances are categorized as hate speech if done intentionally and planned, or there is a clear intention to spread the hatred, intended to incite or cause threats and violence.
In Indonesia matters relating to hate speech have been regulated in the Criminal Law Act, the Law on Information and Electronic Transactions, the Law on the Elimination of Racial and Ethnic Discrimination, and the Law on Social Conflicts [4–6]. In fact, the Indonesian National Police regulates the prohibition of hate speech through the Circular of the Indonesian National Police Chief Number SE / 06 / X / 2015 [7], although it still reaps the pros and cons among legal experts. In the circular, there are seven criteria that can be categorized as expressions of hatred including contempt, defamation, defamation, unpleasant acts, provoking, inciting, and spreading false news. Islam as the majority religion in Indonesia has governed how to interact socially with humans properly and wisely. In the Holy Qur'an, Surah Ali Imron verse 159, being gentle and not speaking is a noble behaviour. When we speak harsh words will make others stay away from us. Many verses of the Qur'an and hadith that mention the virtues of speaking good words, among others, are the cause of getting forgiveness and the reason to enter heaven, get a special room in Heaven later, as a practice that can replace alms, save from the torments of hell, and can eliminate the hostility described in the Qur'an Surah Fushilat verses 34-35.

It cannot be denied that the role of social media in neutralizing hate speech is a big part. Internet users in Indonesia have skyrocketed until 2016 reaching 132.7 million people (the results of a survey of the Indonesian Internet Service Providers/ Asosiasi Penyelenggara Jasa Internet Indonesia (APJII)), and around 90% that are accessed every day are social media [8,9]. Various issues such as political issues, religion, race, gender, citizenship, etc. are subjected to expressions of hatred that can provoke people's emotions. There are even persons who deliberately provide services to spread hate speech through social media. At a price of hundreds of millions of rupiah, they spread hate speech through thousands of fake social media accounts to "adu domba" (bring into conflict) between religions for the benefit of individuals, groups, and politics. In a short time, the utterance of hatred is spread and can have the effect of dividing the unity and integrity of the nation and religious community.

The issues of religion, politics, race, gender, and other sensitive matters are the main targets in spreading hate speech. In this script, we create a logical framework for analysing messages sent through social media such as Facebook, Twitter, Instagram, and several other electronic media such as blogs and leading news media related to hate speech. The analysis carried out is the extent to which these messages contain expressions of hatred. This paper provides benefits as a systematic and measurable framework for filtering messages and news from various social media so that internet and social media users are wiser in their activities on social media. So that social media users are not easily influenced by hate speech, do not re-spread hate speech, and can even stop messages or news that contain hate speech. The big benefit is the maintenance of the unity and unity of the Indonesian nation, especially Muslims, which is a religion that teaches peace and gentleness in every language and speech act.

Data derived from social media, one of which is in the form of text. Text is unstructured or semi-structured data that requires special handling to be carried out the process of exploring the implied meaning or knowledge in it [10,11]. Therefore, text data needs to be changed into structured data representation. Structured representation of a text in general there are two types, namely the form of a single word or better known as bag of words and the form of multiple words. Bag of words is a form of structured representation of text by gathering all words in a text document without looking at the interrelationships of words [11], whereas multiple word representations are text representations that collect words in a text document by paying attention to the interrelations between words so that with multiple word representations semantic meaning in text documents can be maintained better, because it can capture the relationship between words/ phrases, even clauses and sentences [12,13]. Sequential patterns are a form of representation of multiple words, so that compared without sequential patterns in the form of single words the meaning or knowledge of text documents can be better guarded [14,15]. Sequential patterns used in plain text are called Sequences of Word (SOW) with word sequences that pay attention to the order in which words appear, one of which is Frequent Word Sequence (FWS) which is a set of FWS by observing the order in which each word appears [16–19].

In this paper we use a programming algorithm and software engineering approach, starting from the concept of frameworks, modelling, to algorithms that can be used for text mining in producing text analysis that contains hate speech. For example, the BIDE (BI-Directional Extention) algorithm as one
of the sequential pattern mining algorithms. Where, sequential pattern mining is a part of data mining techniques that extract patterns in the form of sequential data, also called set of features which store hidden information \([15,20,21]\). BIDE algorithm is the development of the PrefixSpan algorithm by applying feature selection to form frequent closed sequences (FCS) \([20,22,23]\). FCS is a more efficient FWS feature, because there are no further sub sequences of frequent super sequences. A more efficient text representation feature will facilitate and streamline the hate speech analysis process. Although in terms of processing time and memory usage is not more efficient than PrefixSpan, but the proven features produced are more efficient and do not reduce meaning \([22,24]\). This is in accordance with the need for hate speech analysis in this text, to produce an efficient analysis of the right size.

2. Methodology

This manuscript begins with a study of the literature relating to hate speech, frameworks, software engineering, text mining, sequential pattern mining, algorithms that support text mining and sequential pattern mining. The data used in this manuscript was collected from several social media. Data is filtered only related to Islamic issues. Social media as a source of data for this research include Facebook, Twitter, Instagram, Blogs, and Forums or other online media and websites.

2.1. Text mining

Text mining is a technique to generate new knowledge from text automatically or semi-automatically, where the extracted knowledge is useful and comes from very large amount of text data \([11]\). Data managed in text mining techniques are unstructured text data. The difference between data mining and text mining is the process of extracting features or patterns derived from different forms of data. In data mining feature extraction comes from structured data, while in text mining comes from semi-structured data, it cannot be said to be completely unstructured or said to be structured, although most of it comes from unstructured data \([20,25]\).

The stages of doing text mining include text pre-processing, text transformation (feature generation), feature selection, data mining / pattern discovery, and implementation or evaluation \([11]\). The text pre-processing stage is the process of cleaning up the data, in the process also marked part of speech and parsing where the sentence is a stand-alone graph. Next is the process of text transformation or feature generation which is the process of generating all features that can be in the form of structured text representations in the form of single words or bag of words or multiple words. The next step is feature selection to select features that are still redundant and only useful features. Data mining stage is to obtain knowledge from data text by applying data mining methods, for example clustering, classification, or association along with data mining techniques or algorithms. Finally, the implementation / evaluation stage is implementing and evaluating patterns generated from the data mining process to see its accuracy so that appropriate knowledge can be generated.

The pre-processing stage in text is the stage of preparing unstructured text data into a representation of structured data so that the data is ready for further processing \([26]\). The expected result of pre-processing is a representation that is ready to be used for the mining process by transforming text into structured representation from the extraction process to selection \([27,28]\).

2.2. Sequential pattern

Sequential patterns are patterns that are formed from sequentially repetitive transactions \([20]\). The technique of finding sequential patterns or sequential pattern mining is part of a data mining technique that extracts patterns in the form of sequential data, also called sets of features that store hidden information \([15]\). Sequential patterns emerge from the idea of shopping transactions in supermarkets that are often done by customers, there are the emergence of goods that are often bought together, there are also items that are sequentially purchased after other goods so that a pattern appears \([21]\). However, although the order of appearance is important, the appearance of items in the sequential pattern does not have to be continuous, so there may be other items between one item and another item in a sequential pattern.
3. Results and discussion

A number of experts written in www.cloudtweaks.com in 2015 predict that around 2.5 trillion bytes of data flow on the internet every day, of course, currently exceeding that number. Social media like Facebook, Twitter, WhatsApp, Instagram, and so on become the biggest contributors to the data generated. Of course, the era of social media became the era of big data, where data that flows has large volume characteristics, velocity (megalor data is real time), variety (has a variety of data formats, both structured and unstructured such as text, video, audio), variability (data flow becomes inconsistent, especially in busy times or when trending topics), and complexity (data can be generated from various sources) [29–31]. These facts become the basis for easy and fast news spread, and even become a trending topic. News or messages that contain negative content such as human trafficking, pornography, gossip, hoaxes, and even hate speech can quickly spread. The danger is that hate content can trigger violence and crime, and even divide the unity and integrity of the nation.

In Indonesia, content containing hate speech has been regulated in Article 156 of the Indonesian Criminal Code; Article 157 paragraphs 1 and 2; Article 310 paragraphs 1, 2 and 3; Article 311 paragraph 1; Law No 11 of 2008 concerning Information and Electronic Transactions Article 28 paragraphs 1 and 2; Article 45 paragraph 2; Law No. 40 of 2008 concerning the Elimination of Racial and Ethnic Discrimination Article 16 [4–6]. Even the Police Circular of the Republic of Indonesia Number SE / 06 / X / 2015 contains the characteristics of messages that are categorized as expressions of hatred, among others [7]:

- **Insult.** What is meant by insults is the act of attacking the honour and reputation of others, whether individual or group.
- **Defamation.** Defamation is an act that has an impact on the defamation or honour of someone who is stated both verbally and in writing.
- **Unpleasant deeds.** A treatment that offends someone is called an unpleasant act, forcing others to do, not do, or let something by using the threat of violence, threat of pollution both verbally and in writing.
- **Provoke.** Provoking means to take action to provoke someone's anger either by inciting or deliberately provoking anger.
- **Incite.** Inciting means encouraging, inviting, arousing a person's enthusiasm to do something. In the case of hate speech, incitement can cause negative provocation so that it triggers acts of violence.
- **Spread false news.** Spreading false news or known as hoaxes is also included in the category of hate speech, because the news that is spread is false news and can cause slander and hatred.

It is undeniable that the more sophisticated social media technology, the more sophisticated the ways to disseminate hateful content. Until it appeared, people like Saracen who sold thousands of social media accounts and messages containing hate speech at a price of hundreds of millions. Of course, this is very troubling, because people like Saracens are able to "fight sheep" between religious communities, between political groups by spreading hate speech which certainly contains slander and hoaxes. With social media, hate speech in a matter of minutes or seconds can be viral spread throughout Indonesia, even the world.

Framework design is one of the Software Engineering studies that focuses on building, designing, planning, and implementing a systematic work order and reasoning. Where, logical framework can be used to design, plan, implement, monitor, and evaluate an activity to the project. A logical framework is a set of activities that relate inputs, processes, and outputs to achieve certain goals or goals. Figure 1 shows a proposed logical framework that can be implemented to detect hate speech from social media using text mining technology and sequential pattern mining.
4. Conclusion

Hate speech is a type of crime that can damage the order and harmony of society. Especially if the hate speech contains sensitive issues such as religious issues. The rapid development of technology that cannot be avoided, especially social media, becomes a boomerang if it is not used wisely, one of which is by spreading hate speech. For this reason, with technology, hate speech can be avoided or minimized. Text mining technology can be used as a first step to analyse the content that contains hate speech. This paper compiles a logical framework that can be implemented and built into future scripts to reduce the spread of hate speech. For further works, logical framework to analyse hate speech from social media can be developed.

Acknowledgments

We would like to thank to the Research and Publishing Centre, the Research and Community Service Institute, UIN Sunan Gunung Djati Bandung for supporting and funding the publication of this research.

References

[1] Garner B A 2014 Black’s Law Dictionary 10th Edition (NY: Henry Campbell Black)
[2] Anonymous 2013 "United Nations Human Rights Office of the High Commissioner," ohchr [Online] Retrieved from: http://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf (Accessed on: September 20, 2017)
[3] Brugger W 2003 The treatment of hate speech in German constitutional law (part I) German Law Journal 4(1) 1-22
[4] Anonymous n.d. "Undang-Undang Republik Indonesia, Nomor 40 Tahun 2008 tentang Penghapusan Diskriminasi Ras dan Etnis,"
[5] Anonymous n.d. "Undang-Undang Republik Indonesia, Nomor 7 Tahun 2012 tentang Penanganan Konflik Sosial,"
[6] Anonymous n.d. "Undang-Undang Republik Indonesia, Nomor 11 Tahun 2008 tentang Informasi dan Transaksi Elektronik,"
[7] Anonymous n.d. "Surat Edaran tentang Penanganan Ujaran Kebencian (Hate Speech),"
[8] Anonymous 2018 "Ini Media Sosial Paling Populer di Indonesia," Katadata [Online] Retrieved from: https://databoks.katadata.co.id/datapublish/2018/02/01/media-sosial-apa-yang-paling-
sering-digunakan-masyarakat-indonesia (Accessed on: November 30, 2018)

[9] Pratomo Y 2019 "APJII: Jumlah Pengguna Internet di Indonesia Tembus 171 Juta Jiwa."
Kompas.com [Online] Retrieved from: https://tekno.kompas.com/read/2019/05/16/03260037/apjii-jumlah-pengguna-internet-di-indonesia-tembus-171-juta-jiwa (Accessed on: July 14, 2019)

[10] Torre C J, Matrin-Bautista M J S and Blanco I D 2008 Text Knowledge Mining: And Approach to Text Mining ESTYLF08 17 19

[11] Zohar Y E 2002 Introduction to Text Mining (Autom Learn Group, Univ Illinois)

[12] Doucet A 2006 Advanced document description, a sequential approach ACM SIGIR Forum 40(1) 71-72

[13] Doucet A and Ahonen-Myka H 2004 Non-contiguous word sequences for information retrieval Proceedings of the Workshop on Multitword Expressions: Integrating Processing pp 88-95

[14] Maylawati D S A 2015 Pembangunan library pre-processing untuk text mining dengan representasi himpunan frequent word itemset (hfwi) studi kasus: bahasa gaul Indonesia (Bandung: Institut Teknologi Bandung)

[15] Hoonlror A 2011 Sequential patterns and temporal patterns for text mining (NY: Rensselaer Polytechnic Institute)

[16] Agrawal R, Srikant R and Lent B S 1999 Mapping words, phrases using sequential-pattern to find user specific trends in a text database (Washington, DC: U.S. Patent and Trademark Office)

[17] Ahonen-Myka H 2002 Discovery of frequent word sequences in text Pattern Detection and Discovery pp 180-189 (Berlin, Heidelberg: Springer)

[18] Nalistia R A 2008 Feature-Based Clustering untuk Pengelompokan Dokumen dengan Representasi Text Terstruktur (Bandung: Institut Teknologi Bandung)

[19] Maylawati D S A and Saptawati G P 2016 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang International Conference on Computing and Applied Informatics pp 1-6

[20] Han J, Pei J and Kamber M 2011 Data mining: concepts and techniques (Elsevier)

[21] Srikant R and Agrwal R 1996 Mining sequential patterns: Generalizations and performance improvements International Conference on Extending Database Technology (Berlin, Heidelberg: Springer) pp 1-17

[22] Wang J and Han J 2004 BIDE: Efficient mining of frequent closed sequences Proceedings. 20th international conference on data engineering (IEEE) pp 79-90

[23] Wang J, Han J and Li C 2007 Frequent closed sequence mining without candidate maintenance IEEE Transactions on Knowledge and Data Engineering 19(8) 1042-1056

[24] Maylawati D S A, Irfan M and Zulfikar W B 2016 Comparison between BIDE, PrefixSpan, and TRuleGrowth for Mining of Indonesian Text IOP Conference Series 801

[25] Wei T, Lu Y, Chang H, Zhou Q and Bao X 2015 A semantic approach for text clustering using WordNet and lexical chains Expert Systems with Applications 42(4) 2264-2275

[26] Mahgoub H, Rössner D, Ismail N and Torkey F 2008 A text mining technique using association rules extraction International journal of computational intelligence 4(1) 21-28

[27] Srividiya V and Anitha R 2010 Evaluating preprocessing techniques in text categorization International journal of computer science and application 47(11) 49-51

[28] Kannan S and Gurusamy V 2014 Preprocessing techniques for text mining International Journal of Computer Science & Communication Networks 5(1) 7-16

[29] Arockia P S, Varnekkha S S and Veneshia K A 2017 The 17 V’s of Big Data Int Res J Eng Technol 4 3–6

[30] Tan W, Blake M B, Saleh I and Dusdjar S 2013 Social-network-sourced big data analytics IEEE Internet Computing 17(5) 62-69

[31] Wu X, Zhu X, Wu G Q and Ding W 2013 Data mining with big data IEEE transactions on knowledge and data engineering 26(1) 97-107