Cyberbullying Detection on Social Media in Indonesia with Text Mining

Nedya Farisia[1], Yova Ruldeviyani [1], dan Eko Kuswardono Budiardjo [1]

[1]Magister Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Indonesia, Jl. Salemba no 4, Jakarta, 10430

E-mail: nedya.farisia@alumni.ui.ac.id, yova@cs.ui.ac.id, dan eko@cs.ui.ac.id

Abstract

Social media is growing rapidly at the moment and provide convenience to communicate. But such convenience widely misused to treat other people with not decent before the entire internet community commonly called cyberbullying. If cyberbullying fail to prevent, it will be difficult to track down and deal with it. One of the main weapons to prevent acts of cyberbullying is to perform detection on social media. Detection of cyberbullying can be done by determining whether a post offend the sensitive topic of a personal nature such as racist or not. By determining the related words such sensitive topics and filter sentiment, cyberbullying tweet detection is done by using the method of classification Hyperpipes, Tree-based J48, and SVM. The results show that the algorithm hyperpipes and *decision tree* produces the best evaluation results with the accuracy of 85.32% and 86.24%.

Keywords: Cyberbullying, social media, text mining, data mining

Introduction

Cyberbullying is a form of human rights violation to hurt or humiliate someone through communication technology such as the internet, cell phones, or other technologies. (AHR, 2013). Cyberbullying is mostly on Facebook and Twitter, and a small part through email and SMS, almost half of the victims are children and not a few of them commit suicide. The purpose of this study is to find the best approach in detecting cyberbullying tweets in Indonesia. This approach is used to make it easier for the authorities to delete cyberbullying posts and blacklist violators. In Indonesia, it is not known how to detect posts/tweets that are classified as cyberbullying with text mining. Therefore, the research question can be formulated as follows:

1. “How to detect cyberbullying on social media in Indonesia by utilizing text mining?”
2. “Which classification algorithm can detect cyberbullying the most accurately on social media in Indonesia?”

The research is limited to conversations that occur on social media in Indonesia only. In addition, this research can only be conducted on social media that allows its content to be accessed by the public (Twitter). Due to data limitations, the types of cybercrime that can be

analyzed are flaming and denigration. Crawled data is limited to November 2015 and May 2016.

The benefit of this research to the community is to increase the security of IT users in using social media. The benefit of this research to the government is that it makes it easier to prevent cyberbullying. And the benefit to science is as an alternative reference for how to detect cyberbullying in Indonesia.

Theoretical Review

To determine whether a comment/post is cyberbullying/cannot be determined by topic (Dinakar, Jones, Catherine, Henry, & Picard, 2012). Because the target victim of cyberbullying is a specific individual, topics related to bullying are sensitive matters that are personal to the victim. Psychological research results (Mishna, Cook, Gadalla, Daciuk, & Solomon, 2010) said that most children and adolescents have high sensitivity related to the topic of sexuality, race, physicality, and intelligence. Repeated postings on this topic can cause the victim to believe what the bully is saying, which can damage the victim's health (Dinakar, Jones, Catherine, Henry, & Picard, 2012).

Because this study uses Indonesian, most of the preprocessing methods used are from Margono's research (2014). Margono has found a term that has a strong correlation with cyberbullying in Indonesia. This study adds the preprocessing method used by Margono (2014) to that used by Nahar (2013), namely a simple sentiment filter. This study also uses topics to detect such as Nahar's (2013) research. It's just that the topic is not formed by clustering but has been determined from the start based on the findings of Margono (2014). This study uses the classification algorithm recommended by previous studies, namely SVM and J48. It's just that both of them come from English data. Hence the hyperpipes algorithm derived from bruteforce was added. This study uses the same software as the previous research by Nahar (2014) and Dinakar (2012) to extract knowledge, namely WEKA.

Theoretical Framework

Figure 1 shows how to use text mining to detect cyberbullying on social media. The following is an explanation of each method chosen and the reasons:

1. Establish corpus

To detect conversations that offend someone's personality, the first step is determined by a collection of Indonesian bullying terms taken from Margono's research (2014) plus phrases/terms from domain experts. The addition of phrases/terms is an idea taken from the research of Dinakar (2012). Phrases/terms that are added are a subtle way for someone to bully because they don't immediately say harsh words. Phrases/terms are limited by sensitive topics that have been previously discovered by Margono, namely intelligence, difabel (handicaps), behavior, and animals.

2. Preprocessing

The preprocessing step in this study also comes from the research of Margono (2014) only that a simple sentiment filter is added to increase accuracy such as Nahar's (2013) research.

3. Extract knowledge:

There are three algorithms in this study that will be tested, namely hyperpipes, J48 (decision tree), and SVM. The three algorithms were chosen for the reasons previously mentioned, namely based on the ability to minimize errors, minimize unnecessary calculations, withstand noise, be able to analyze high-dimensional data, recommended by previous studies (Nahar (2014) and Dinakar (2012)), and have high accuracy. Which is good when done brute force (hyperpipes). The purpose of brute force here is that a small amount of data is tested on all WEKA algorithms and an algorithm that has high accuracy is selected.

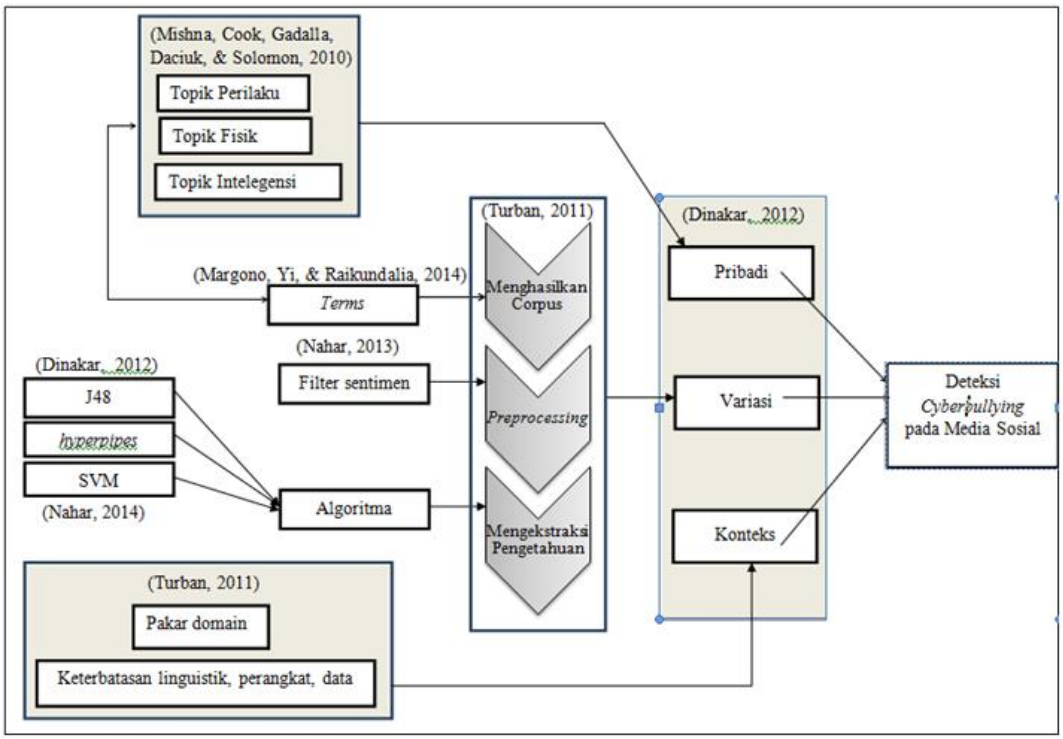


Figure 1 Theoretical Framework

Methods

The stages of research carried out in this study are described in **Table 1**.

Table 1 Research Methodology

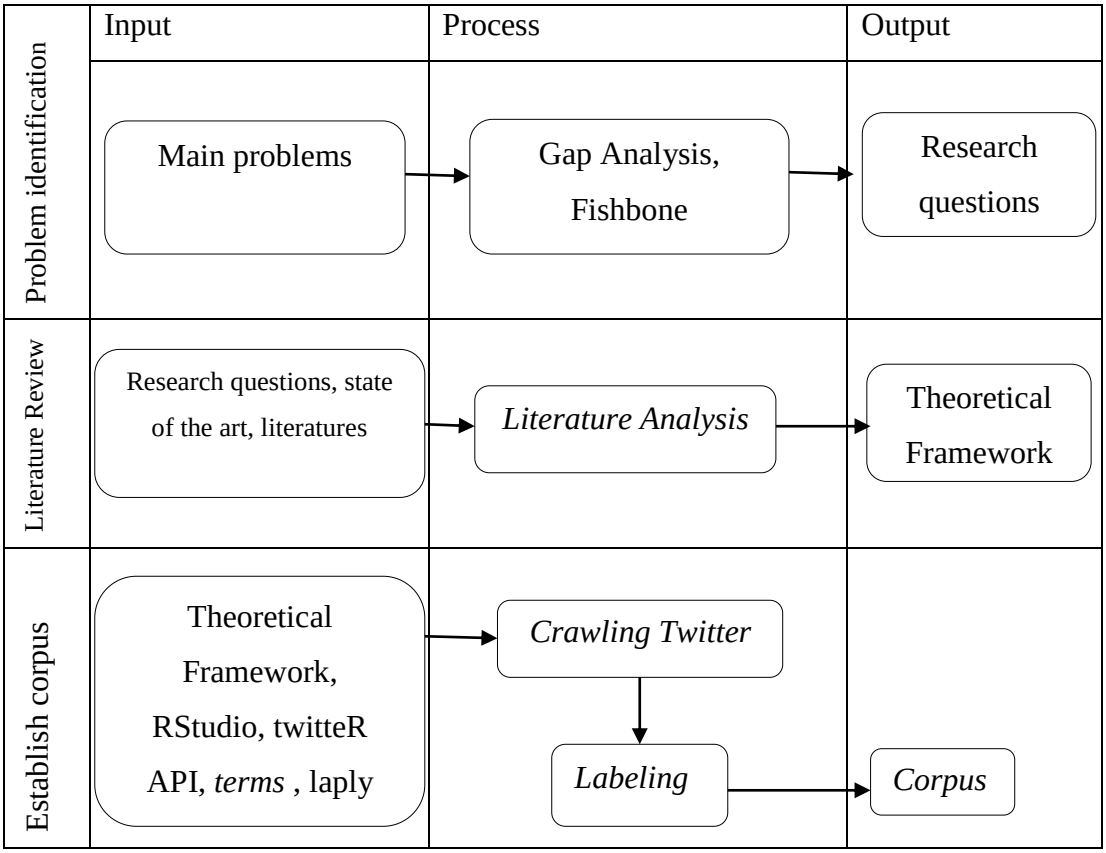
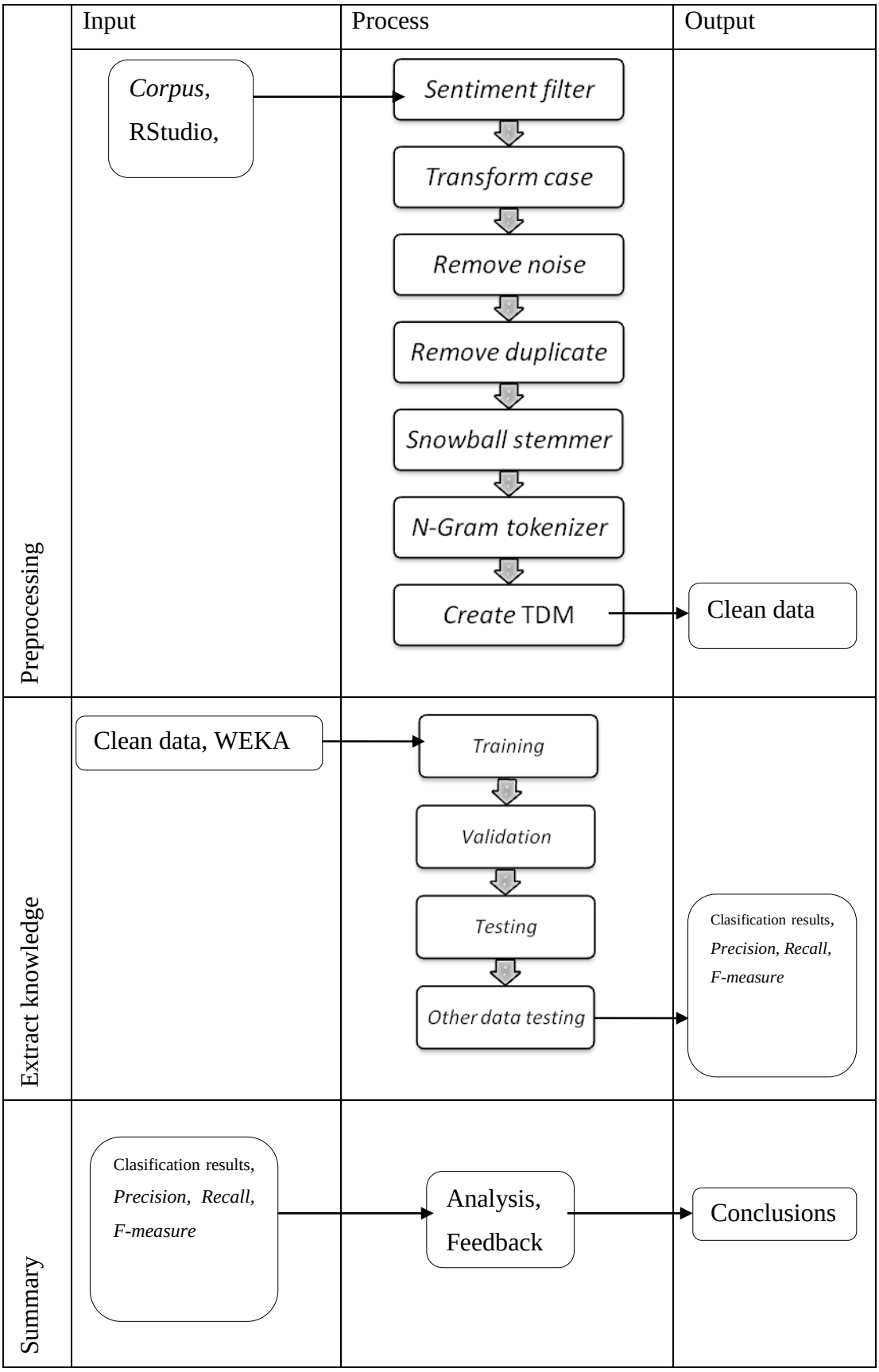


Table 1 Research Method (cont.)



Results

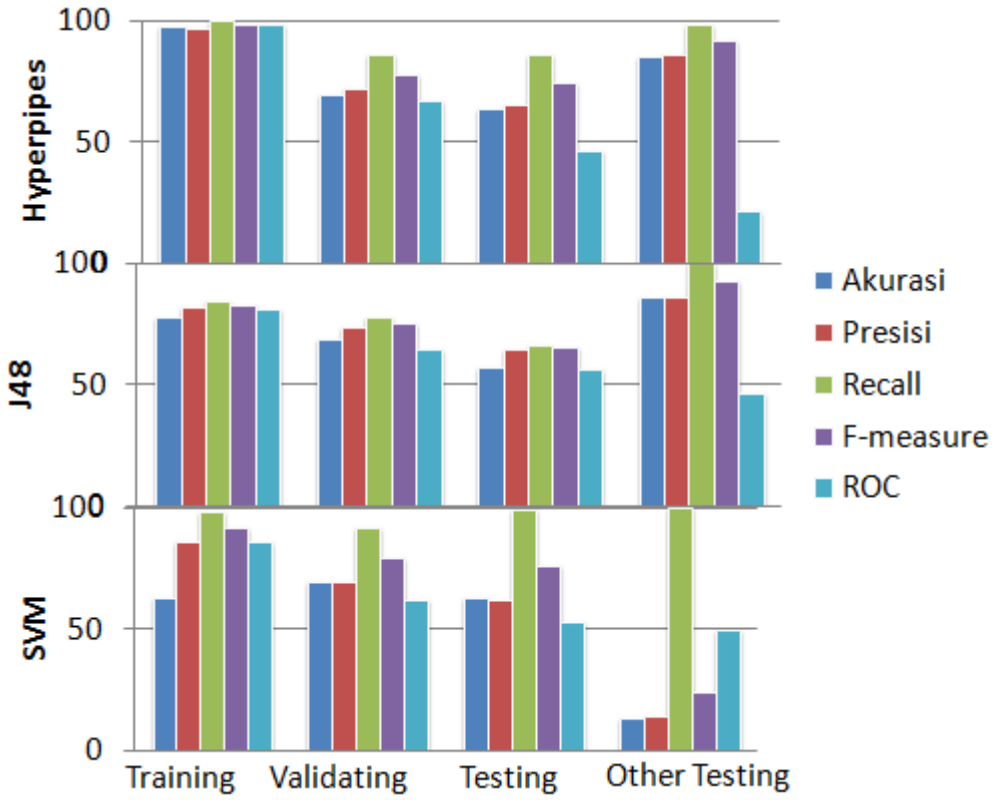


Figure 2 The evaluation results are based on several criteria at the training, validation, testing, and testing stages from other sources (the presidential debate).

Figure 2 shows the visualization of the results of this study based on the criteria of accuracy, precision, recall, F-measure, and ROC. Of all the experiments conducted in this study, the hyperpipes and J48 models have the best accuracy. However, the hyperpipes model is only good for crawling data with the same keywords/terms as the training data. Meanwhile, if used for conversational data outside of crawling keywords, hyperpipes will be a bad classifier. This indicates the low generalization of the classifier. J48 with its exploration ability has a greater generalization so that it will be better used for conversation cases outside the predetermined terms.

Discussion

Analysis of the research results shows that blatant and clear bullying sentences can be easily detected as cyberbullying. Because the three classification models in this study can detect it correctly. What is meant by correctly detected tweets can be seen on TP and TN. The results of the TP and TN classifications from the three classification algorithm models in the table are

intended to be insulting and cursing. They have a repeating and stable pattern and contain profanity and negative expressions. However, all three models misclassify tweets that do not contain this pattern and which require little understanding of the sentence. For example, tweets that only meant to ask and express sadness were detected incorrectly by the SVM model. This means that taking into account some punctuation and emoji (emotional analysis) can help correct errors in this model.

Table 2.1 Examples of tweets that are detected are true (TP and TN) and false (FP and FN)

Model	Examples of detected tweets	
J48	TP	si miras dkira sll bejad padahal yg sadis jg ada yg dkira miras p sadis ada ustad pendeta ada anak baik p bejad dalilnya khilaf jiwasoak
	TN	SIALAN MATI LAMPU. BANGSAT PLN NI :/
	FP	Makin banyak yg bangsat
	FN	rt nih guru ngaji gak mabok muridnya gak pake pakaian seksi tetep bejad co fg kpbhajr
Hyperpipes	TP	Woy njing!! ibumu pelacur ya pantesan kelakuan lu bejad! Dasar anak haram!! @MekelSungg bruakakakak :D
	TN	plays "Polisi bejad"
	FP	Ga salah emg, Allah udh paling maha baik. Gue terjauhkan dr org2 bejad
	FN	@ari09402154 pengikut keluarga cendana mentalnya lebih bejad dibanding pki
SVM	TP	Pendukung @Prabowo08 spt ARB, SDA adl org2 BEJAD tabiatnya kasar termasuk org ini >> @novrinawawi smoga dilaknat oleh ALLAH
	TN	(((BEJAD MASSAL)))
	FP	@darmanug @_Handschar_ @ramberusuh Mengajarkan Allah yg tauhid, apa hubungannya dgn moral yg bejad???
	FN	ciyo bejad bat ew :(

In addition to emotional analysis and balancing the data discussed in the previous paragraph, the convergence (clustering) and sparcity of the data also affect the performance of the algorithm in this study. The training data in this study has a balanced and continuous

character as shown in While the testing data is a bit grouped and the presidential debate data is sparse so that it affects the performance of the hyperpipes and SVM algorithms which are based on the frequency of occurrence and the distance between the data. The test data gives an advantage to the J48 algorithm because the clustered data is studied by it as a new node discovery.

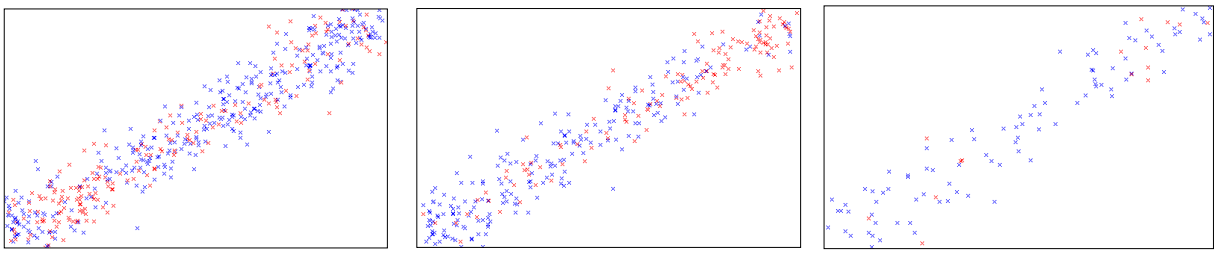


Figure 3 Visualization of cyberbullying data characters (red) on (a) training data (b) testing data (c) presidential debate data

Conclusion

Cyberbullying can be detected based on certain topics by using text mining. In this study, supervised learning was carried out using the hyperpipes classification algorithm, C4.5 (J48), and SVM. From the analysis obtained on several tests in chapter 4, the following conclusions can be drawn:

1. Sentences that clearly intend to oppress in a blatant way can be easily detected as cyberbullying in this study.
2. Balance, convergence, and sparseness of test data can adversely affect the performance of hyperpipes and SVM algorithms, but not for J48.
3. Classification on Twitter using the decision tree algorithm (J48) is able to detect cyberbullying with the best level of accuracy in this study, which is 86.24%.
4. The hyperpipes classification algorithm has a good accuracy value only on testing data that comes from the same source as the training data. The addition of new words that match the new test data can increase the accuracy of 36%.
5. SVM has not been able to study the classification of cyberbullying well in this study, obtained from the area under the ROC curve which is 0.5.
6. In supervised learning architecture, as in this study, a slightly smaller number of datasets will be more profitable, so the label 'not cyberbullying' should be reduced because the results affect the recall of the label 'cyberbullying'.

Suggestions

1. Some punctuation marks and emojis should also be taken into account as factors that affect cyberbullying detection.
2. The test data should be balanced beforehand by oversampling and downsizing techniques.

References

AHR. (2013). *Cyberbullying, Human rights and bystanders*. Retrieved 2015, from Australian Human Rights Commission: https://bullying.humanrights.gov.au/cyberbullying-human-rights-and-bystanders

APJII; PUSKAKOM UI. (2015). *Profil Pengguna Internet Indonesia* . APJII.

Bannerman, P. L. (2008). Risk and risk management in software projects: A reassessment. *The Journal of Systems and Software* , 2118-2133.

Dinakar, K., Jones, B., Catherine, H., Henry, L., & Picard, R. (2012). Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems , 2* (3).

Djojosoedarso, S. (2004). *Prinsip-Prinsip Manajemen Risiko.* Jakarta: Salemba Empat.

facts. (n.d.). Retrieved Oktober 2015, from http://facts.net/cyber-bullying/#f8

General Manager IT Services Delivery, O. &. (2015, Februari 18). Pandangan Terkait Manajemen Proyek TI di PT. XYZ. (R. Ali, Interviewer)

Hakizabera, A. U., & Ohsato, A. (2010). Early Risk Assessment in Software Development Life Cycle Using Software Metrics. *Proceedings of the International Conference on Information Managemant and Evaluation* , 122-128.

Higuera, R. P., & Haimes, Y. Y. (1996). *Software Risk Management.* Pittsburgh: Software Engineering Institute.

Hijazi, H., Alqrainy, S., Muaidi, H., & Khdour, T. (2014). Risk Factors In Software Development Phases. *European Scientific Journal* , 213-232.

ISO 31000. (2009). *Risk Management.* Geneva: International Organization for Standardization.

Marchewka, J. T. (2009). *Information Technology Project Management.* Illinois: John Wiley & Sons.

Margono, H., Yi, X., & Raikundalia, G. K. (2014). Mining Indonesian Cyber Bullying Patterns in Social Networks. *Proceedings of the Thirty-Seventh Australasian Computer*

Science Conference (ACSC 2014) (pp. 115-124). Auckland: Australian Computer Society, Inc.

Mishna, F., Cook, C., Gadalla, T., Daciuk, J., & Solomon, S. (2010). Cyberbullying behaviours amoung middle and high school students. *Amer. J. Orthopsychiatry* .

Pressman, R. S., & Maxim, B. R. (2013). *Software Engineering: A Practitioner's Approach.* New York: McGraw Hill.

Project Management Institute. (2013). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)* (5th Edition ed.). United States of America: Project Management Institute, Inc.

Schwalbe, K. (2012). *Information Technology Project Management.* Boston: Cengage Learning.

Software Engineering Institute. (2007). *Introducing OCTAVE Allegro: Improving the Information Security Risk Assessment Process.* Hanscom AFB: Carneige Mellon University.

Sommerville, I. (2009). *Software Engineering.* Boston: Addison-Wesley.

WeAreSocial. (2015). *Global Digital Statistics.*

Woody, C. (2006). *Applying OCTAVE: Practitioners Report.* Hanscom AFB: Carneige Mellon University.