International Journal of Engineering & Technology, 7 (4.29) (2018) 72-75

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Association Analysis of Cyberbullying on Social Media using Apriori Algorithm

Zuraini Zainol[1]*, Sharyar Wani[1], Puteri N.E. Nohuddin[2], Wan M.U. Noormanshah[2], Syahaneim Marzukhi[1]

[1]Department of Computer Science, Faculty of Science and Defence Technology, National Defence University of Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia
[2]Institute of Visual Informatics, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor, Malaysia
**Corresponding author E-mail: zuraini@upnm.edu.my*

Abstract

With the phenomenal increase in use of Social Networking Service (SNS) and mobile technology, the consequences of cyberbullying have become an epidemic. More than 80% youth use cell phones making them extremely vulnerable to the abuse and one in three young people have been found victims of this problem. There are many different methods of detection cyberbullying behaviour patterns however rarely any focuses on analysis based on association especially in Malay language. Learning and detecting using association is a natural communication phenomenon that can help to identify abusive content from the hidden corpora, which often goes unnoticed. Association helps to identify trends, rules and patterns of the bullies and detects abusive content considering whole sets rather than focusing on single instances. The current work focuses on detection of cyberbullying instances by association analysis using the Apriori Algorithm. It mainly focuses on detecting bullying and aggressive behaviour on Twitter. Over 80 different patterns with high confidence levels were detected that can be successfully implemented for the detection process. The high confidence levels are indicative of the efficiency of association analysis for cyberbully detection in SNS.

Keywords: *cyberbullying detection; association rule mining; association analysis; twitter; malay; cybersafety*

1. Introduction

Social media and "always on always-connected devices" have completely changed the phase of social interactions. However, the negative side of its applications, is that it aggravates aggressive and bullying behaviours among users [1]. Cyberbullying refers to bullying, harassment and intimidation through online platforms such as Facebook, Twitter, Instagram, etc. Due to the ease of access and popularity of these platforms, the rate of abuse has significantly increased [2, 3]. In fact, the extent of its consequences have reached epidemic levels e.g.; 50% of young social media users reported cases of being bullied [3]. Researchers suggest that adolescents face these abuse attacks often. Most of the times, the perpetrator and the victim know each other. Cyber bullying typically lasts for longer periods and can happen at any point of time. Moreover, the audience reach is higher. Thus, making it severely painful for the victims [4].

Cyberbullying is a by-product of adolescent aggression and electronic communication. Cyberbullying takes varied forms such as cyber-harassment, cyberbullying harassment by proxy, cyberstalking, denigration, exclusion/gossip groups, falsify identity, flaming, impersonation, online grooming, outing, phishing, sexting, trickery [5]. From the electronic communication and technology perspective, cyberbullying uses different ways and means such as blogs, burn books, cameras, cell phones, chat rooms, email, gaming devices, happy-slapping, instant messaging platforms, personal pictures of others, etc. [5]. In summary, cyberbullying may be defined as, "an aggressive intentional act carried out by a group or individual, using electronic forms of contact" repeatedly and over time against a victim who cannot defend him or herself" [6]. Written-verbal behaviour, visual behaviour, exclusion and impersonation have been classified as four mains methods of cyberbullying. The detail of the aforementioned classification is presented in Table 1 [7, 8]. The aim of this paper is to analyse cyberbully words and discover cyberbully patterns and trends on Twitter using Malay language.

The remainder of the paper is organized as follows: Section 2 explores related work and discusses different methods that have been used for cyberbullying detection. Section 3 discusses the experiments details, where Apriori and Association Rule Mining are used to analyse the research problem. Section 4 presents the experimental results. Finally, the conclusion and future work of the study are made in Section 5.

Table 1: Categories of Cyberbullying and Cyberbullying Activities

Category of Cyberbullying	Cyberbullying Activities
Written-Verbal Behavioural	phone calls, text messages, e-mails, instant messaging, chats, blogs, social networking communities, websites
Visual Behaviour	posting, sending or sharing compromising pictures and videos through mobile phone or internet
Exclusion	purposefully excluding someone from an online group
Impersonation	stealing and revealing personal information, using another person's name and account

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2. Related Work

Various techniques have been developed to detect abusive content on various social media platforms such as YouTube, Instagram and Yahoo Answers, etc. [9-12]. Textual and structural features have been analysed to predict user behaviour for YouTube comments, word embedding and supervised learning classification are often used for distinguish abusive content on Yahoo Finance [9, 10]. Similarly manual annotation, manual characterization and bag-of-words classification is often used for detecting cyberbullying on YouTube [13]. Hee et. al. [14] utilizes linguistic characteristics analysis for detection of abuse on Ask.fm. Supervised approaches such as regression model, naïve bayes, support vector machines, decision trees, graph based approaches and probabilistic sentiment analysis have also been used for detection of cyberbullying on various cyber platforms [3, 11, 13-17].

Work in [3] presents algorithm to label aggressive and bullying behaviour in Twitter. The data is mined using WEKA and categorization is based on studying the characteristics of the genre, alongside annotated by crowdsourcing techniques. Based on the categorization, the machine-learning model is built with a Random Forest Classifier. The categorization results using the model are reported at 91%. Detection of gender bullying on twitter using a distant supervision method has been reported in [18].

Improved bag-of-words model using TF-IDF, neglecting grammar and word order but keeping multiplicity have been applied to detect online harassment. The accuracy and precision were reported around 39% and 62% respectively [19]. An automated cyberbullying detection method has been proposed based on classification of comments based on their topics such as sexuality and race, etc. The method detects at an accuracy of about 67% [13]. Graph features such as node number and edge number, etc. have been used by [20] for bullying detection by representing social networks as graphs. It is important to consider that cyberbullying is a social phenomenon. Therefore, a successful detection method needs to consider the social aspect for successful detection. Tree based detection is one such method based on analysing the parent-child relationship between comments. Comments are broken down as parent and child, scored and finally classifiers are often used for detection of any possible abuse. These tasks are done using Pattern, VADER, Weka and SMOTE, etc. The results have been reported at an average of 69%, bearing different values in different use cases [21].

Singh et. al. [22] advocates using fusion approaches to combine heterogeneous sources for cyberbullying detection. The approach is often based on multiple text such as density of bad words, part of speech tags, etc. and social network based features such as number of nodes, number of edges, degree centrality, etc. The researchers deem it is important to calculate the interdependency values from heterogeneous aspects for an efficient detection process. The experimental dataset focuses entirely on twitter. The results indicate that the proposed approach is highly efficient compared to some of the previously used methods such as early fusion, etc.

Bigelow, et al. [23] uses Latent Semantic Indexing over the social media site Formspring.me for detection of cyberbullying. The LSI methodology is often based on Singular Value Decomposition and does not use the formerly developed cyberbullying dictionaries. LSI efficiently helps to detect cyberbullying content in small posts with misspellings, abbreviations and unusual punctuation, classical features of a SNS communication among youth. The system is successful in detecting the abusive content seven to nine times the baseline. Mining using association rules and FP-Growth has been presented in [24]. They used association rules and Apriori Algorithm to find trends and patterns of cyberbullying on Twitter in Indonesian language. The results presented using both association rule and Apriori Algorithm seems promising.

3. Overview of Framework for Discovering Frequent Keyword Patterns and Relationships (DFKPR)

In previous study [25], we proposed a framework for discovering interesting frequent patterns and relationships using Apriori Algorithm. Also, the algorithm has been discussed in this work. The framework consists of three main stages: (i) data collection and pre-processing, (ii) frequent keyword pattern identification, and (iii) interesting relationships discovery. This framework applied frequent keyword pattern mining and association rules mining for extracting interesting relationships between keyword and the Hadith Chapters from the Book of Friday prayer. By using the technique of Association Rule Mining (ARM), frequent patterns and interesting relationships that can benefits Muslim scholars in making full use of Hadith in assisting them in their daily Islamic life and practice. ARM has been widely applied in many application domains such as healthcare [26], predicting flood areas [27], text mining [25], education, etc. In this study, we implement the framework of [25] over the tweets dataset on cyberbullying (see Figure 1).

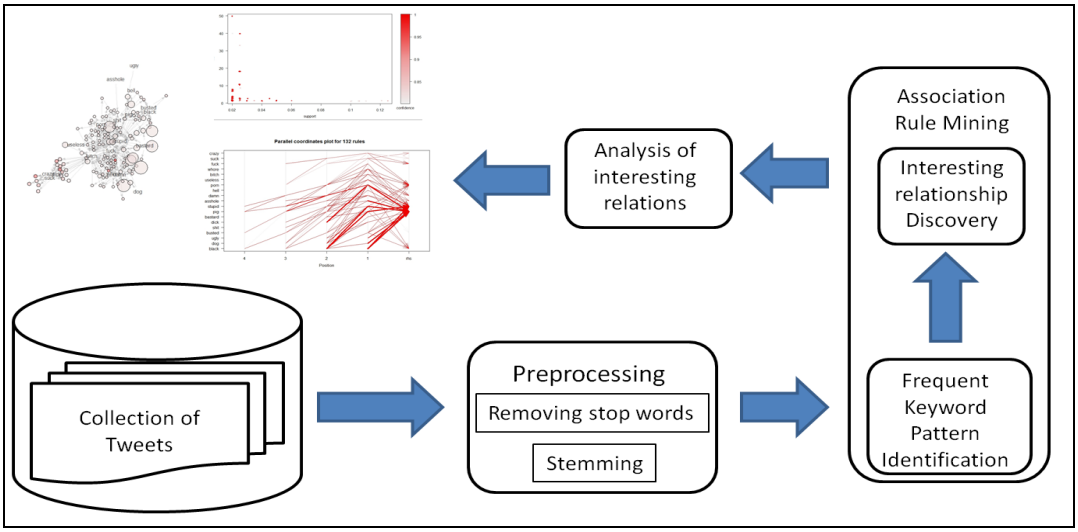


Fig. 1: The Framework for Discovering Frequent Keyword Patterns and Relationship (DFKPR) adopted from [25]

2.1. Data Collection and Pre-processing

This section provides the results from the experiment carried out using DFKPR framework. Data collection is a crucial phase in this research. Researchers spent a lot of time on collecting text data from Twitter and preprocessing task. In this study, we collected text data from Twitter using RapidMiner 7.5.3. RapidMiner is a powerful visual workflow designer for building predictive analytic workflows. RapidMiner also offers hundreds features of data preparation (e.g., sampling, transformations, weighting and selection, attribute generation, etc.), machine learning algorithms to support data science projects [28]. In this study, we applied the Twitter Connector for accessing the Twitter data (phrases, tweets or user profile information) directly from RapidMiner Studio. The 'Search Twitter' operator (available by the RapidMiner "Social Media" extension) enables users to search the top 1000 most recent posting tweets, provided that a Twitter connection has been configured first in this experiment.

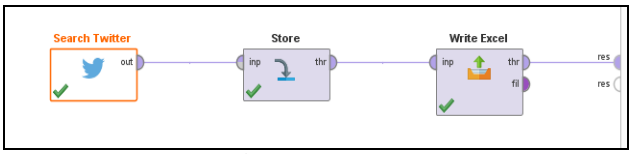


Fig. 2: A screenshot for extracting text data in RapidMiner

Figure 2 shows the implementation of three operators in the process of collecting Twitter data. In the "Search Twitter" operator, users can set up the requirement of tweets by filling up the given parameters (see Figure 3). For example, the number of tweets to be returned is set up to 1000 and only recent tweets need to be returned to RapidMiner Studio is specified in the "Result type" parameter.

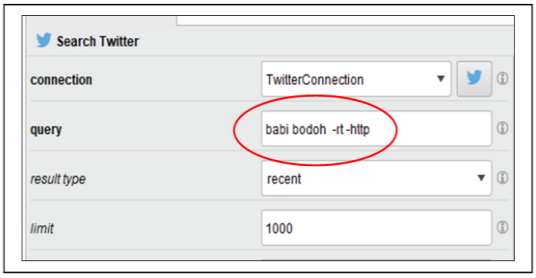


Fig. 3: A screenshot for setting the tweets details that contain specific keywords

We also filtered our search by removing retweets (-rt) and links (-http) (see Figure 3). The result containing all posting tweets text containing the specified keywords or terms with their tweets metadata will be displayed in the RapidMiner format (see Figure 4).

ExampleSet (475 examples, 1 special attribute, 11 regular attributes)

...	Id	Creat...	F...	Fro...	T...	...	...	...	Text
1	894788...	Aug 8, ...	Afi	577...	i...	2...	tl	...	@imanalaska BODOH BABI NAMANYA KAU NI HAHAHAHAHA...
2	894774...	Aug 8, ...	K...	415...	h...	4...	in	...	@hudaaalatif Babi apa manja pegang pegang tangan. Ingat k...
3	894747...	Aug 8, ...	T...	243...	?	-1	in	...	ye aku bodoh banggang babi...aku terima hinaan tu...
4	894733...	Aug 8, ...	s...	770...	s...	7...	in	...	bodoh tau tak bodoh. Kau dah buat booking AWAL, bayar la A...
5	894727...	Aug 8, ...	F...	295...	?	-1	in	...	Bodoh gila dpt duit terus lepas tangan,sumpah mcm babi. htt...
6	894711...	Aug 8, ...	za...	547...	?	-1	in	...	Babi mana la yang selit chewing gum kat tempat duduk ni. Bo...
7	894700...	Aug 8, ...	s...	871...	ak...	8...	in	...	@akmalrosely Kau la babi bodoh tak sampai sehari sial
8	894699...	Aug 8, ...	cu...	594...	?	-1	in	...	Bodoh ah eg babi betul
9	894679...	Aug 8, ...	q...	733...	?	-1	in	...	Ni yg bodoh nii bile da sober jd lg bodoh babi trgntung
10	894630...	Aug 8, ...	p...	219...	?	-1	in	...	yup. babi, sial, kimak, gampang, bodoh, buto, bangang, cibai, ...
11	894617...	Aug 8, ...	b...	236...	?	-1	in	...	Babi lah kalau ingat balik haih rasa bodoh sangat

Fig. 4: A screenshot for setting the tweets details that contain specific keywords

In this study, we collected twitter data that contains most common 'malays' cyberbullying keywords such as 'bodoh', 'sial', 'gila', 'babi'. 'haram', 'anjing', etc. Based on our observation on the Twitter, these keywords appeared frequently in the Twitter amongst Malaysian Twitter users. We selected the Malay language because we would like to mine Malay cyberbullying words experienced by Twitter users. The Twitter data consists of 8275 tweets, which are then converted and stored in the form of Excel format. These tweets are collected using RapidMiner between 15th July, 2017 and 15th August, 2017. This Twitter data is now ready to be used as an input data and further analysis using R Mining.

The next step is to carry out the preprocessing task for cleaning the Twitter data. Preprocessing is very important task in text mining, as it prepares data for further analysis. The pre-processing helps to remove errors and inconsistencies from the dataset [29]. Furthermore, it will improve the efficiency the text mining process, the quality of text data [30,31] and provide an accurate result [32]. In this experiment, the pre-processing of Twitter data comprises of multiple tasks such as importing data, cleaning and structuring the Twitter data for later analysis. The cleaning task consists of two steps: (i) convert all tweets text to lower case for standardization, (ii) remove punctuation marks, whitespace, symbols and numbers, etc. After that, we remove stop words and stemming in the Twitter data. In this study, both steps are manually performed as Twitter dataset is originally collected in Bahasa Malaysia.

2.2. Frequent Keyword Pattern Identification

The next stage of the proposed framework is Frequent Keyword Pattern Identification. The cleaned dataset is then stored as an input file for this module. This module identifies a list of frequent keywords that exist in the Twitter dataset. Frequent Keyword Pattern Identification module is developed based on a well-known Association Rule Mining (ARM) technique which is Apriori Algorithm. The identified frequent keyword patterns are referred to as patterns that are frequently found in the itemsets where the frequency in the target dataset is not less than a (user-specified) threshold value. The Apriori algorithm also generates frequent keyword patterns and builds association rules from these patterns. In this experiment, we assumed that the keywords in each Twitter data as items in the transactions, whereas each tweet data is considered as a transaction.

The module produces the number of frequent keyword pattern using three different metric values such as 'support', 'confidence' and 'lift'. The parameter 'support' refers to the frequency counts of each frequent keywords occurrences in each Twitter dataset, whereas 'confidence' and 'lift' are applied to proof the validity of association rules that have been generated based on the discovered frequent keyword patterns. Thus, in this study, the support is measured as the number of data records where item X $\cup$ item Y appears and divided with the total transactions (tweets). The 'confidence' specifies the number of data records where item X $\cup$ item Y appears and then divided with the number of data records where item X appear. Finally, 'lift' measures the importance or interestingness of the discovered frequent keyword patterns and rules.

2.3. Interesting Relationships Discovery

In this sub-section, the generation of association rules of each frequent keyword patterns will indicate the relationships between discovered patterns to the Twitter datasets are interesting. To assist the association analysis in this study, we use two types visualization techniques for examining and illustrating the association rules. The visualization techniques in this study are graph-based visualization and plot interactive graph visualization. Nonetheless, the interpretation of interesting association analysis and rules

should be assisted with the subject matter expertise. Each of this visualization will be discussed in the following section.

4. Experimental Results and Analysis

In this section, the experimental results for the proposed frequent keyword patterns discovery in cyberbully tweets are discussed. 8275 tweets were used as an input dataset for frequent pattern mining. In this experiment, 16 keywords represent as items while 8275 tweets are transactions. The frequencies of each item are presented in Figure 5.

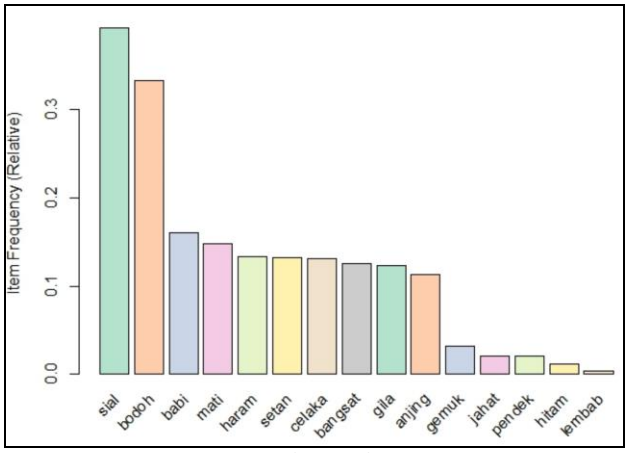


Fig. 5: A screenshot for item frequency histogram

Figure 5 presents the item frequency histogram that illustrates how many times an item has occurred in the dataset as compared to the others. The items 'sial'(3244), 'bodoh'(2755) and 'babi'(1329) had the highest frequencies in the 8275 Twitter data. This indicates that, these items are frequently used keywords by most Malaysian tweeters to bully someone on Twitter.

	lhs		rhs	support	confidence	lift
[1]	{bodoh}	=>	{sial}	0.16374622	0.49183303	1.2545987
[2]	{sial}	=>	{bodoh}	0.16374622	0.41769420	1.2545987
[3]	{haram}	=>	{babi}	0.11444109	0.86012716	5.3555698
[4]	{babi}	=>	{haram}	0.11444109	0.71256584	5.3555698
[5]	{gila}	=>	{bodoh}	0.10151057	0.82758621	2.4857626
[6]	{bodoh}	=>	{gila}	0.10151057	0.30490018	2.4857626
[7]	{bangsat}	=>	{sial}	0.05619335	0.44840887	1.1438297
[8]	{sial}	=>	{bangsat}	0.05619335	0.14334155	1.1438297
[9]	{anjing}	=>	{sial}	0.05353474	0.47228145	1.2047253
[10]	{sial}	=>	{anjing}	0.05353474	0.13655980	1.2047253
[11]	{bangsat}	=>	{anjing}	0.05256798	0.41947927	3.7006300
[12]	{anjing}	=>	{bangsat}	0.05256798	0.46375267	3.7006300
[13]	{babi}	=>	{sial}	0.05135952	0.31978932	0.8157388
[14]	{sial}	=>	{babi}	0.05135952	0.13101110	0.8157388
[15]	{celaka}	=>	{sial}	0.05123867	0.39186691	0.9995989
[16]	{sial}	=>	{celaka}	0.05123867	0.13070284	0.9995989
[17]	{mati}	=>	{bodoh}	0.03782477	0.25467860	0.7649602
[18]	{bodoh}	=>	{mati}	0.03782477	0.11361162	0.7649602
[19]	{babi}	=>	{bodoh}	0.03202417	0.19939804	0.5989179
[20]	{bodoh}	=>	{babi}	0.03202417	0.09618875	0.5989179

Fig. 6: The result after generating Association Rules based on 'support' measurement

In this experiment, we set the minimum support threshold is 0.005 and the minimum confident threshold is 0.06. As shown in Figure 6, 88 association rules are generated. These rules are then ranked based on the value of 'support' parameter. There are many ways to describe the derived rules. For example, rules 1 and 2 $\{bodoh\} \rightarrow \{sial\}$ and $\{sial\} \rightarrow \{bodoh\}$ shows that 16.4% of Malaysian tweeters likely to post the keywords 'bodoh' will also post the keyword 'sial' in their tweets. Rules 3 and 4, $\{haram\} \rightarrow \{babi\}$ and $\{babi\} \rightarrow \{haram\}$ also share similar support values 11.4%. This shows that 11.4% of Malaysian tweeters likely to post the keywords 'haram' will also post the same keyword 'babi' in their tweets.

	lhs		rhs	support	confidence	lift
[1]	{pendek}	=>	{gemuk}	0.020060423	0.9764706	30.491676
[2]	{gemuk}	=>	{pendek}	0.020060423	0.6264151	30.491676
[3]	{hitam}	=>	{gemuk}	0.010755287	0.9175258	28.651041
[4]	{gemuk}	=>	{hitam}	0.010755287	0.3358491	28.651041
[5]	{anjing,haram}	=>	{babi}	0.017401813	0.9230769	5.747526
[6]	{anjing,haram,sial}	=>	{babi}	0.007854985	0.9027778	5.621133
[7]	{anjing,babi,sial}	=>	{haram}	0.007854985	0.7303371	5.489137
[8]	{haram}	=>	{babi}	0.114441088	0.8601272	5.355570
[9]	{babi}	=>	{haram}	0.114441088	0.7125658	5.355570
[10]	{anjing,babi}	=>	{haram}	0.017401813	0.6697674	5.033902

Fig. 7: The result after generating Association Rules based on 'lift' measurement

Among the 88 rules are produced in the experiment, rules 1 and 2 have the highest lift at 30.49 (see Figure 7). This shows that whenever Malaysian tweeters posted the keywords 'pendek', they are very likely to post the keyword 'gemuk'. This result can also be visualized in form of an interactive plot in Figure 9.

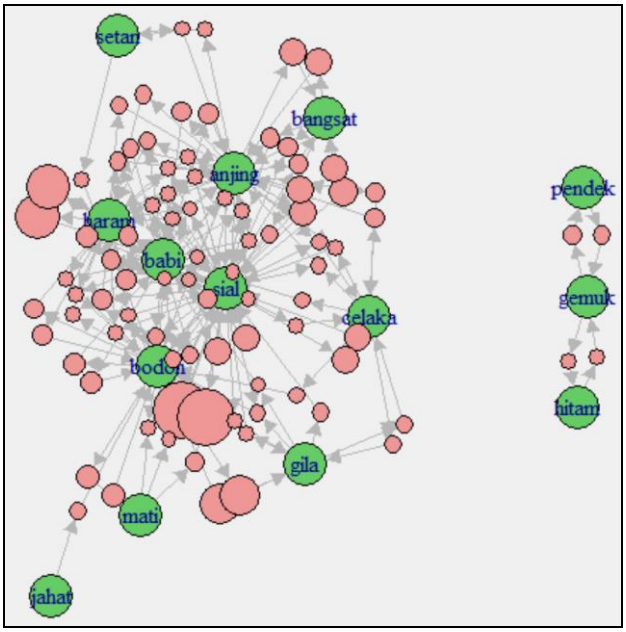


Fig. 8: Visualization of network graph for 10 association rules according to lift

The next step is to visualize the association rules using the network graph. Figure 8 shows the network graph for the top 88 rules according to lift measure. As seen in Figure 8, the size of graph nodes is based on support levels and the colour on lift ratios. The larger nodes or circles imply higher support while red nodes imply higher lift. The incoming lines show the left hand side (LHS) and the right hand side (RHS) are represented by names of keywords (items). The above graph illustrates that most of tweets were associating around the keywords 'sial', 'bodoh' and 'babi'. This shows that these items are frequently used keywords by most Malaysian tweeters to bully someone on Twitter.

Figure 9 illustrates an interactive plot graph that presents 88 rules generated. As seen below, the confidence levels are plotted on the Y axis whereas the support levels on the X axis for each rule. The details of each rule can be examined by pointing the cursor on the interactive plot graph. As shown in Figure 9, the interactive plot graph suggests that rules $\{gemuk\} \rightarrow \{pendek\}$ has the confidence of 0.976, support of 0.0201 and the highest lift value is at 30.5.

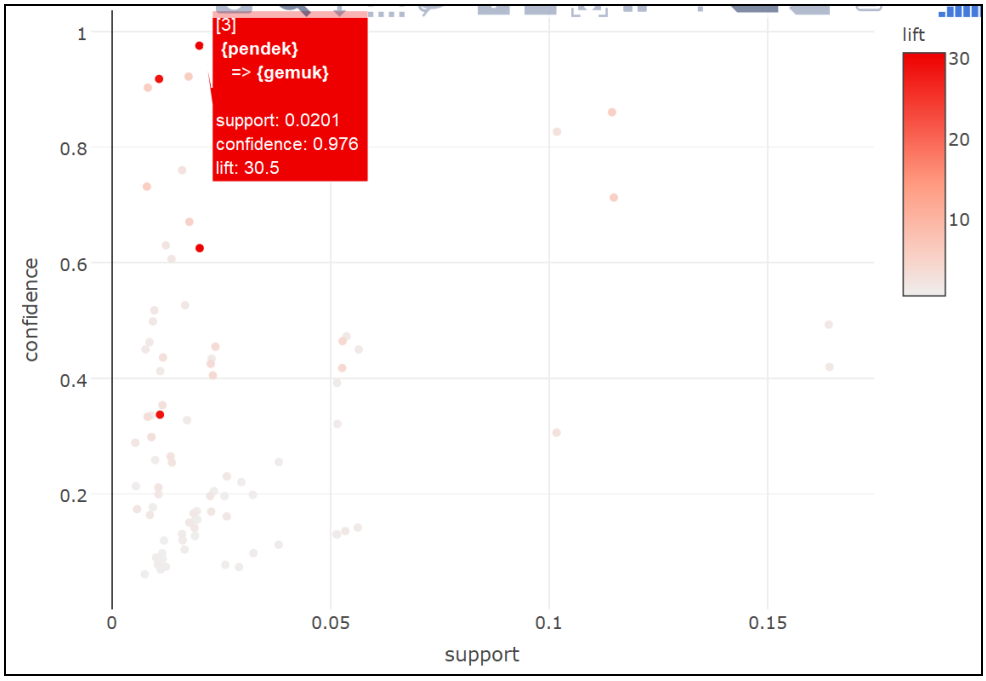


Fig. 9: An interactive plot for 88 association rules

5. Conclusion

The paper presents a cyberbullying detection method by association analysis and Apriori Algoritm using Malay language. The corpus was based on the data from Twitter. The results present a pattern and 88 relationship rules among the cyberbullying terms. The highest confidence level based detection successfully demonstrates the potential of association analysis and Apriori Algorithm for establishment of patterns and detection of cyberbullying in SNS. For future work, this research will extend the tweets dataset and mine deeply the Malay cyberbully words using clustering techniques. This research also will be further extended to detect and predict what kind of Malay cyberbullying patterns and trends on Twitter.

Acknowledgement

The authors would like to thank Faculty of Science and Defence Technology and National Defence University of Malaysia (NDUM) for sponsoring this publication.

References

[1] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, & A. Vakali, "Measuring #GamerGate: A Tale of Hate, Sexism, and Bullying," *Proceedings of the 26th International Conference on World Wide Web Companion*, Perth, Australia, 2017.

[2] Z. Ashktorab, "A Study of Cyberbullying Detection and Mitigation on Instagram," *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, San Francisco, California, USA, 2016

[3] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, & A. Vakali, "Mean Birds: Detecting Aggression and Bullying on Twitter," *Proceedings of the 2017 ACM on Web Science Conference*, Troy, New York, USA, 2017

[4] M. Rybnicek, R. Poisel, & S. Tjoa, "Facebook Watchdog: A Research Agenda for Detecting Online Grooming and Bullying Activities," in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 2854-2859.

[5] S. P. Kiriakidis & A. Kavoura, "Cyberbullying: A review of the literature on harassment through the internet and other electronic means," Family & community health, Vol. 33, No. 2, pp. 82-93, 2010.

[6] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, & N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *Journal of child psychology and psychiatry*, Vol. 49, No. 4, pp. 376-385, 2008.

[7] A. Ioannou et al., "From risk factors to detection and intervention: A metareview and practical proposal for research on cyberbullying," in 2017 IST-Africa Week Conference (IST-Africa), 2017, pp. 1-8

[8] A. Nocentini, J. Calmaestra, A. Schultze-Krumbholz, H. Scheithauer, R. Ortega, & E. Menesini, "Cyberbullying: Labels, behaviours and definition in three European countries," *Journal of Psychologists and Counsellors in Schools*, Vol. 20, No. 2, pp. 129-142, 2010

[9] Y. Chen, Y. Zhou, S. Zhu, & H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," presented at the *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, 2012.

[10] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, & N. Bhamidipati, "Hate Speech Detection with Comment Embeddings," *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015.

[11] H. Hosseinmardi, S. A. Mattson, R. Ibn Rafiq, R. Han, Q. Lv, & S. Mishra, "Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network," in Social Informatics: *Proceeding of the 7th International Conference*, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings, T.-Y. Liu, C. N. Scollon, and W. Zhu, Eds. Cham: Springer International Publishing, 2015, pp. 49-66.

[12] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, & F. Bonchi, "The Social World of Content Abusers in Community Question Answering," *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015.

[13] K. Dinakar, R. Reichart, & H. Lieberman, "Modeling the detection of Textual Cyberbullying," *The Social Mobile Web*, Vol. 11, No. 02, 2011.

[14] C. Van Hee et al., "Automatic detection and prevention of cyberbullying," *Proceeding of the International Conference on Human and Social Analytics* (HUSO 2015), 2015, pp. 13-18: IARIA

[15] M. Dadvar, D. Trieschnigg, & F. de Jong, "Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies," in Advances in Artificial Intelligence: 27th Canadian Conference on Artificial Intelligence, Canadian AI 2014, Montréal, QC, Canada, 2014. Proceedings, M. Sokolova and P. van Beek, Eds. Cham: Springer International Publishing, 2014, pp. 275-281.

[16] V. Nahar, S. Unankard, X. Li, & C. Pang, "Sentiment Analysis for Effective Detection of Cyber Bullying," *Proceeding of the Web Technologies and Applications: 14th Asia-Pacific Web Conference, APWeb 2012*, Kunming, China, 2012, pp. 767-774.

[17] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, & Y. Chang, "Abusive Language Detection in Online User Content," P*roceedings of the 25th International Conference on World Wide Web*, Montrealal, Qubec, Canada, 2016.
[18] H. Sanchez & S. Kumar, "Twitter bullying detection," ser. NSDI, Vol. 12, pp. 15-15, 2011.
[19] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, & L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1-7, 2009.
[20] Q. Huang, V. K. Singh, & P. K. Atrey, "Cyber Bullying Detection Using Social and Textual Analysis," presented at the *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, Orlando, Florida, USA, 2014.
[21] Z. Li, J. Kawamoto, Y. Feng, & K. Sakurai, "Cyberbullying detection using parent-child relationship between comments," *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, Singapore, Singapore, (2016).
[22] V. K. Singh, Q. Huang, & P. K. Atrey, "Cyberbullying detection using probabilistic socio-textual information fusion," in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 884-887
[23] J. L. Bigelow, A. Edwards, & L. Edwards, "Detecting Cyberbullying using Latent Semantic Indexing,", *Proceedings of the International Workshop on Computational Methods for CyberSafety*, Indianapolis, IN, USA, (2016).
[24] H. Margono, X. Yi, & G. K. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," *Proceedings of the Thirty-Seventh Australasian Computer Science Conference* – Vol. 147, Auckland, New Zealand, 2014.
[25] Z. Zainol, P. N. E. Nohuddin, M. T. H. Jaymes, & S. Marzukhi, "Discovering "interesting" keyword patterns in Hadith chapter documents," *Proceeding of the International Conference on Information and Communication Technology*, (2016), pp. 104-108, https://doi.org/10.1109/ICICTM.2016.7890785
[26] R. A. A. Rashid, P. N. E. Nohuddin, & Z. Zainol, "Association Rule Mining Using Time Series Data for Malaysia Climate Variability Prediction," *Proceeding of the 5th International Visual Informatics Conference*, (2017), pp. 120-130, https://doi.org/10.1007/978-3-319-70010-6_12
[27] N. A. Harun, M. Makhtar, A. A. Aziz, Z. A. Zakaria, F. S. Abdullah, & J. A. Jusoh, "The Application of Apriori Algorithm in Predicting Flood Areas," *International Journal on Advanced Science, Engineering and Information Technology*, Vol. 7, No. 3, 2017.
[28] RapidMiner. Available online: https://rapidminer.com/products/studio/, last visit:29.05.1017
[29] Z. Zainol, A. M. Azahari, S. Wani, S. Marzukhi, P. N. E. Nohuddin, & O. Zakaria, "Visualizing Military Explicit Knowledge using Document Clustering Techniques " *International Journal of Academic Research in Business and Social Sciences*, Vol. 8, No. 6, (2018), pp. 1127-1143, available online: http://hrmars.com/hrmars_papers/Visualizing_Military_Explicit_Knowledge_using_Document_Clustering_Techniques.pdf, last visit:25.07.2018
[30] Z. Zainol, S. Marzukhi, P. N. E. Nohuddin, W. M. U. Noormaanshah, & O. Zakaria, "Document Clustering in Military Explicit Knowledge: A Study on Peacekeeping Documents," in *Proceeding of the 5th International Visual Informatics Conference*, (2017), pp. 175-184, https://doi.org/10.1007/978-3-319-70010-6_17
[31] Z. Zainol, P. N. E. Nohuddin, T. A. T. Mohd, & O. Zakaria, "Text Analytics of Unstructured Textual Data: A Study on Military Peacekeeping Document using R Text Mining Package " in *Proceeding of the 6th International Conference on Computing & Informatics*, (2017), pp. 1-7.
[32] Z. Zainol, M. T. H. Jaymes, & P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," *Journal of Physics: Conference Series*, Vol. 1018, No. 1, (2018), pp. 012011, available online: http://iopscience.iop.org/article/10.1088/1742-6596/1018/1/012011/meta, last visit: 28.07.2018