Automatic Labelling of Malay Cyberbullying Twitter Corpus using Combinations of Sentiment, Emotion and Toxicity Polarities

Ruhaila, Maskat*
Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA Selangor, Malaysia
ruhaila@fskm.uitm.edu.my

Muhammad Faizzuddin, Zainal
Faculty of Computer & Mathematical Sciences Universiti Teknologi MARA Selangor, Malaysia
faizzainal97@gmail.com

Nurrissammimayantie, Ismail
Academy of Language Studies, Universiti Teknologi MARA Selangor, Malaysia
nurrissa@uitm.edu.my

Norizah, Ardi
Academy of Language Studies, Universiti Teknologi MARA Selangor, Malaysia
norizah@uitm.edu.my

Amirah, Ahmad
Academy of Language Studies, Universiti Teknologi MARA Selangor, Malaysia
amirah1275@uitm.edu.my

Noriza, Daud
Academy of Language Studies, Universiti Teknologi MARA Selangor, Malaysia
norizadaud@uitm.edu.my

ABSTRACT

Automatic labelling is essential in large corpuses. Engaging in human experts to label can be challenging. Semantic understanding can differ from one labeler to another based on individual's language ability. Platforms such as AmazonTurk are not able to ensure the quality of annotations in every domain. Extensive steps such as qualification and counter checking of labels may be implemented which will increase the cost of data annotation. Thus, the higher quality of labelled data expected, the greater the cost that needs to be expended. This scenario is made worse when the language is of low resource where in this work is the Malay language. Malay is a language used mostly in Malaysia, Indonesia, Singapore and Brunei. Unlike English which has large resources to tap into the semantics of sentences, making automatic labelling faster to mature, resources in Malay language are still limited. Further compounded is the use of social media data where the text is short, unnormalized and the inherent presence of code switching. The availability of qualified native Malay labelers is also scarce. To overcome this, we devised a method to automatically label a total of 219,444 Malay tweets by using a combination of sentiment, emotion and toxicity polarities. We extend the work from Arslan et al. who proposed the use of sentiment and emotion to identify cyberbullying text. Our work added toxicity polarity in the context of automatic labelling of cyberbully tweets in Malay. We were able to employ 5 experts with formal degrees in Malay language to label our training set. We applied this method to Malay cyberbullying corpus to determine "bully" and "not bully" labels. We have tested our method on 54,867 manually labelled data and achieved high accuracy.

*Corresponding author

ACAI 2020, December 24–26, 2020, Sanya, China

ACM ISBN 978-1-4503-8811-5/20/12…$15.00
https://doi.org/10.1145/3446132.3446412

CCS CONCEPTS

• **Computing methodologies**; • **Artificial Intelligence**; • **Natural language processing**; • **Information extraction**;

KEYWORDS

Automatic labelling, Malay language, Cyberbullying, Twitter

ACM Reference Format:
Ruhaila, Maskat, Muhammad Faizzuddin, Zainal, Nurrissammimayantie, Ismail, Norizah, Ardi, Amirah, Ahmad, and Noriza, Daud. 2020. Automatic Labelling of Malay Cyberbullying Twitter Corpus using Combinations of Sentiment, Emotion and Toxicity Polarities. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence (ACAI 2020), December 24–26, 2020, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3446132.3446412

1 INTRODUCTION

Labelling large corpuses can be a daunting task and worse, impractical. While existing data annotation platforms are readily available, such as the AmazonTurk which provides access to thousands of paid labelers, there is the important aspect of qualified labelers. Scarcity of qualified labelers in a particular domain or subject can deem these platforms useless. This is especially pronounced in low resource languages such as Malay. In contrast, English is the second most spoken language in the world, boasting a considerable number of potentially qualified labelers to choose from.

To produce high quality labelled data, a good number of language experts is required, however, when circumstances prohibit this, the qualification of the labelers is essential to ensure the production of good labels. It is common that there will be conflicts in perceiving labels across labelers, thus must be discussed and an agreed final label must be determined. This counter checking and the selection of highly qualified labelers can be costly.

Besides that, the use of social media data brings along its own challenges. As a result of most social media platforms limiting the length of text to be written, data tend to be short and concise. The shortness of the allowable text means punctuations are often not exercised. Therefore, the context requires guesswork which could be pinpointed by someone who understands informal Malay slang

and possibly current issues due to social media having become the tool for expressing thoughts and opinions. Also, as a result of the limited allowable length of text, social media data is largely unnormalized. Shortcuts to spelling are widely practiced. This commonly involves the sacrificing of vowels in a word. As a consequence, a word can exhibit ambiguous meanings which relies on the underlying context, yet to know the underlying context requires identifying the word. This chicken and egg condition can often be resolved by a native Malay speaker. With English being the second most spoken language in most Asian countries, code switching takes place rampantly. This has produced informal languages known as Manglish in Malaysia and Singlish in Singapore. Such data necessitates a labeler to know the sentence construction of both Malay and English in order to be able to determine the switching that occurred and understand the context. This further complicates the automatic labelling work when social media data is used.

While English resources are abundant to tap into the semantics of sentences, allowing automatic labelling of English to mature at a quicker rate, Malay resources are far lesser. Hence, the slower maturity of Malay automatic labelling. Automatic labelling is a more cost-effective approach but can be less accurate. Misclassification percentage is expected to be larger than human labelling. However, with current big data, automatic labelling has to be considered as a more practical approach than human labelling. Recent works in automatic labelling show a widening adoption across domain. They include crisis response, image of the brain, spoken conversation, handwritings and facial emotion images. All these are big-sized and complex datasets.

According to [3] there is a strong correlation between "anger" emotion and "negative" sentiment labels to cyberbullying. In other words, cyberbullying words other than sarcasm tend to use sentiment of negativity and exhibit angry emotions. As defined by [21] cyberbullying is "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 easily defend him or herself". Cyberbullying can lead to a decline in mental health in vulnerable groups of people (e.g. teenagers) [3]. In this work, we extended this notion to include toxic polarity to label bullying tweets.

Cyberbullying is one of the cybercrimes in Malaysia where there is a high number of cases reported from 2014 to 2016 [23]. Malaysia is ranked 17th highest in cyberbullying among the twenty-five countries surveyed according to the Microsoft Global Youth Online Behaviour Survey [4]. Cyberbullying becomes serious when it leads to many mental problems to victims such as anxiety, depression and even worse, suicidal thoughts [20]. There are some challenges in detecting cyberbullying in social media even through machine learning. Three challenges were suggested by [17]. They are annotation problem, understanding the role in cyberbullying and rapid changes of language.

Malay is a low resource language. Malay can be found used in Malaysia, Singapore, Brunei and Indonesia with the two latter countries exhibiting their own variants of the Malay language. Malay language is much alike the formal Indonesian language. Due to its low resource nature, only small datasets related to cyberbullying in Malay and formal Indonesia could be found [16, 25]. Hence, previously there were no need of automatic labelling and thus a majority of work done focused on the prediction of cyberbullying. Unlike these earlier efforts, the problem that we are trying to address in our work is of a large corpus of Malay cyberbullying tweets that are too many to be manually labelled prior to the prediction of cyberbullying. With more people adopting a life with social media, large corpuses of unstructured social media data were able to be harnessed. It is a known fact that a good prediction performance could only come from a training set with high quality labels. However, to expect humans to be able to label this considerably large amount of data in timely manner is not feasible. Thus, it is imperative to propose a technique capable of automatically labelling Malay tweets yet preserve as much as possible the quality of the labels to be as close a manually labelled corpus.

This paper contributes the following:

- A corpus with a total of 219,444 tweets which consists of common Malay cyberbullying words and 54,867 tweets manually labelled by Malay language experts;
- A proposed method that automatically labels short text data by using a combination of sentiment, emotion and toxicity polarities;
- An evaluation of the proposed method.

2 RELATED WORKS

Proposal [11] of automatic text labelling can be found since 1989. More recent works on automatic text labelling show applications in the areas of software development reports [12], topic modelling [7, 8, 13], medical discussions [18] event extraction [9, 26], students' answer and learning performance [14, 22], crisis response [2], news article [19], user issues [15] and holy book [1]. Generally, underlying most automatic labelling strategies of large corpus is the ability to automatically classify a group of text items with the aim of constructing a training set which will later be used to help in the prediction of a larger unlabeled corpus. Incorrect labelling will produce incorrect prediction.

Recent methods used for automatic labelling utilize machine learning classifiers (SVM, LDA, random forest, deep learning, decision tree and naïve bayes) with adaptations to each uniquely-proposed strategies. They include user summaries [10], text summaries [24] embeddings [7], raised issues [15] and distant supervision [2]. Although these classifiers can also be found widely used for the actual prediction task, labelling by classifying differs by being more stringent in achieving high performance. This is to ensure high quality training set is generated for later use down the analytical stream.

At present, proposals of classifying cyberbullying social media content explore notions of time and interaction (Hsin-Yu Chen et al., 2020), use of multiple information (e.g., image, video, comments, time) or multi-modal (Wang, 2020; Kumari, 2020; Alasadi, 2020), refinement of deep learning (Pradhan et al., 2020; Dadvar et al., 2020; Iwendi et al., 2020), multiple social media platforms (Bruwaene et al., 2020), tweaking of neural network (Lu et al., 2020), reinforcement learning (Aind et al., 2020) and elements of psychology such as emotion and personality. Unlike us, [6] holds the premise that cyberbullying actions are strongly linked to user personality. In their work, personality traits were extracted from The Big Five and Dark Triad models and used to classify social media users to

Attributes	Label = buli	Label = bukan buli	Emotion = an...	Emotion = sadness	Emotion = love	Emotion = fear	Emotion = joy	Emotion = surprise
Label = buli	1	-1	0.055	-0.025	-0.011	-0.020	-0.024	-0.043
Label = bukan buli	-1	1	-0.055	0.025	0.011	0.020	0.024	0.043
Emotion = anger	0.055	-0.055	1	-0.329	-0.577	-0.370	-0.371	-0.285
Emotion = sadness	-0.025	0.025	-0.329	1	-0.065	-0.042	-0.042	-0.032
Emotion = love	-0.011	0.011	-0.577	-0.065	1	-0.073	-0.073	-0.056
Emotion = fear	-0.020	0.020	-0.370	-0.042	-0.073	1	-0.047	-0.036
Emotion = joy	-0.024	0.024	-0.371	-0.042	-0.073	-0.047	1	-0.036
Emotion = surprise	-0.043	0.043	-0.285	-0.032	-0.056	-0.036	-0.036	1

Figure 1: Correlation Matrix of Cyberbullying and Emotion

be either a bully, an aggressor, a spammer or just plain normal. The underlying notion is "personality is a mental function that distinguishes one person from another and its ability in predicting consequential outcomes (e.g. psychological well-being) is suitable for detecting cyberbullying". Personalities of extraversion, agreeableness and neuroticism from the Big Five and psychopathy from the Dark Triad were discovered to have a considerable association with cyberbullying. Cyberbullying tend to be more apparent in individuals who scored high in extraversion, high in neuroticism, high in psychopathy and low in agreeableness. An improved version [5] of [6] shows the addition of psychological elements of sentiments and emotions into the equation apart from just personality and is more similar to ours. The results obtained an improved classification of cyberbullying when personalities and sentiments were used, however, it is not so with emotion. None of these proposals included toxicity into their works.

3 METHOD

We proposed a method extending the notion found in [3] which suggested that there is a strong correlation between "anger" emotion and "negative" sentiment polarity by including toxic polarity for the automatic labelling of Malay cyberbullying tweets. The initial hypothesis is there is a high correlation between tweets viewed by human experts to have cyberbullying content with "anger" emotion and "negative" sentiment. We extend this hypothesis to also include tweets with toxic content. To determine this hypothesis, we constructed correlation matrices between the labelled tweets with each of the categories. A high, positive correlation suggests the accuracy and feasibility of automatically labelling large number of tweets as cyberbullying or otherwise.

Based on [3], 6 emotions were used. They are anger, sadness, love, fear, joy and surprise. Our aim is to find out which of these emotions will have the highest positive correlation with cyberbullying tweets written in Malay. If there is, then we could automatically label any tweets that consists of that particular emotion as cyberbullying in nature. We used a third party API called Malaya [27]. Malaya has several models trained with its relevant dataset with 80% of training dataset and 20% of testing dataset. Bert-based model performed well in emotion analysis and sentiment analysis with score 0.87185 and 0.84132 respectively.

Figure 1 shows the correlation matrix between cyberbullying and emotion. We discovered that most anger emotion has the highest positive association with cyberbullying instances (0.055) while surprise emotion has the highest positive association with non-bullying instances (0.043). This shows that emotion can be used to help automate the labelling of Malay cyberbullying tweets.

Figure 2 displays the correlation matrix between cyberbullying and sentiment. The result shows negative sentiment has the highest positive association with bullying instances (0.075) while neutral sentiment has the highest positive association with non-bullying instances (0.061).

In Figure 3, the result shows toxicity = ['toxic, obscene, insult'] has the highest positive correlation with bullying instances (0.072) while toxicity = [] has the highest positive correlation with non-bullying instances (0.088). '[]' means the tweets have very little to no percentage of toxicity of all 6 toxicity types (toxic, severe toxic, obscene, threat, insult, identity hate).

The compounded detection of the presence and strength of these three categories provides a higher confidence that the labelled tweet is of cyberbullying in nature. Our method can be represented as the following.

Definition (Cyberbullying Automatic Labelling): An automatic labelling of cyberbullying is characterised with a given set of unlabelled Malay tweets $\{x_1,\ldots, x_n\}$. For each tweet, a positive detection of "anger" emotion $e \geq 0.5$ and a "negative" sentiment $s \geq 0.5$ and a toxicity $t \geq 0.5$ will result in the tweet receiving a label "cyberbully" where 1 represents very angry, very negative or very toxic.

$$Label(x) = (e_x \geq 0.5) \wedge (s_x \geq 0.5) \wedge (t_x \geq 0.5)$$

A low threshold may compromise the accuracy of the labels, resulting in an increased possibility of producing more false positives. In contrast, a high threshold may produce more false negatives. To receive a good quality training set, we placed a threshold for our automatic labelling of 0.5. During prediction, a type I error

Attributes	Label = buli	Label = bukan buli	Sentiment = negative	Sentiment = positive	Sentiment = neutral
Label = buli	1	-1	0.075	-0.043	-0.061
Label = bukan buli	-1	1	-0.075	0.043	0.061
Sentiment = negative	0.075	-0.075	1	-0.754	-0.579
Sentiment = positive	-0.043	0.043	-0.754	1	-0.098
Sentiment = neutral	-0.061	0.061	-0.579	-0.098	1

Figure 2: Correlation Matrix of Cyberbullying and Sentiment

Attributes	Label = buli	Label = bukan buli
Label = buli	1	-1
Label = bukan buli	-1	1
Toxicity = ['toxic']	0.038	-0.038
Toxicity = []	-0.088	0.088
Toxicity = ['toxic', 'identity_hate']	-0.004	0.004
Toxicity = ['toxic', 'obscene', 'insult']	0.072	-0.072
Toxicity = ['toxic', 'obscene']	-0.007	0.007
Toxicity = ['toxic', 'obscene', 'insult', 'identity_hate']	0.032	-0.032
Toxicity = ['toxic', 'obscene', 'identity_hate']	0.006	-0.006
Toxicity = ['toxic', 'insult']	0.048	-0.048
Toxicity = ['toxic', 'severe_toxic', 'obscene', 'insult']	0.011	-0.011
Toxicity = ['obscene']	0.002	-0.002
Toxicity = ['toxic', 'insult', 'identity_hate']	0.005	-0.005
Toxicity = ['toxic', 'severe_toxic', 'obscene', 'insult', 'identity_hate']	0.012	-0.012
Toxicity = ['insult']	-0.006	0.006
Toxicity = ['toxic', 'severe_toxic', 'obscene']	0.002	-0.002
Toxicity = ['identity_hate']	-0.001	0.001

Figure 3: Correlation Matrix of Cyberbullying and Toxicity

means more non-bullying tweets will be considered as cyberbullying, while type II error means actual cyberbullying tweets will be overlooked. The effect of the latter is more detrimental than the former and thus must be avoided at the expense of having lesser examples to later train the prediction classifier. Incorrect labelling in automatic labelling scenario can be problematic in large-scale corpuses.

4 EXPERIMENT

4.1 Corpus

Tweets were scraped from Twitter using Python through Twitter-API library. Scraping of tweets used two ways, random-based and keyword-based. A list of common cyberbullying terms received from our Malay language experts were used. Time range is from January 2018 to February 2019. To maximize the gain of cyberbullying instances, we did a random scrape for 2 days without using any cyberbullying words and discovered there is a peak of negative tweets between 12pm to 1pm. This is no surprise considering that this is lunch time in Malaysia where users get the chance to check their social media accounts and post tweets. Refer Figure 4

We then used the given list of keywords to scrape tweets within the peak hour, resulting in 219,444 tweets collected over the 13 months period. The limitation of our corpus is the number of cyberbullying words used is constrained to common words identified by our language experts whereas from time to time newer words are introduced. From the collected tweets, 54,867 were successfully labelled by the language experts. The experts are qualified in Malay

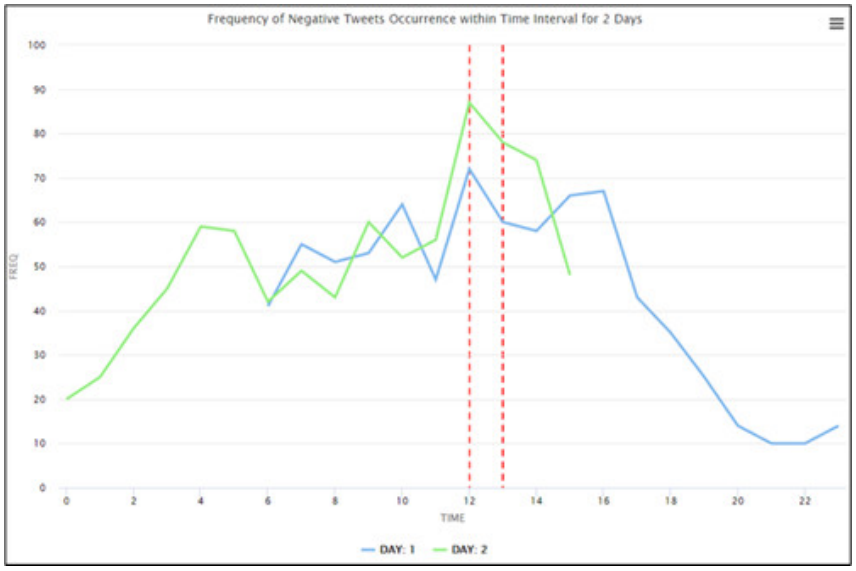

Figure 4: Peak of Negative Tweets within Time Interval of 2 Days

Table 1: Automatic vs. Manual Labels

	Not Cyberbully (Automatic)	Cyberbully (Automatic)
Not Cyberbully (Manual)	120	1,411
Cyberbully (Manual)	57	53,279

Table 2: Sample 1

	Not Cyberbully (Automatic)	Cyberbully (Automatic)
Not Cyberbully (Manual)	2,173	10
Cyberbully (Manual)	7	159
Total	2,180	169

language at a university level, therefore, are quite proficient with the different forms of cyberbullying text.

4.2 Evaluation Procedure

In order to determine the quality of our proposed method, we conducted two phases of evaluation. The assumption is if our method is sound enough against the ground truth, it is more likely to perform well when applied to large unlabelled data. However, reality is the amount of unlabelled data is too enormous to be labelled thus we could only resort to evaluating our generated automatic labels by manually checking samples.

During the first phase, we ran our method on the 54,867 labelled tweets to generate automatic labels. These labels were then checked against the manual labels and the accuracy was calculated. Our method produced 97.3% accuracy (Table 1). We could see that our method is better at identifying cyberbully tweets than non cyberbully ones. The mislabelling rate of not cyberbully tweets labelled as cyberbully is 92% as compared to the mislabelling rate of cyberbully tweets labelled as not cyberbullying equals to 8%. This reduced the possibility of excluding too many tweets that are cyberbully and the adverse consequences as a result of not identifying cyberbully tweets. This indicates that our method is able to correctly label cyberbully data automatically on our unlabelled corpus to a certain degree of correctness. This is a preferable trait since it is more important to identify cyberbully tweets than non cyberbully tweets.

In the next phase, we apply our method to our unlabelled cyber bully corpus. Since the amount is too enormous to be labelled, therefore we could only evaluate the correctness of our automatically generated labels through manual checking of randomly selected samples. Two samples were used to total 4,534. In both samples (Table 2 and Table 3), not cyberbully tweets were serendipitously more. Here our method performed well in identifying cyberbully tweets as well as non cyberbully tweets.

5 CONCLUSION

With large corpuses of text, manual labelling is no longer a feasible option. As a conclusion, using anger emotion, negative sensitivity and toxicity polarities can help in the automatic labelling of

Table 3: Sample 2

	Not Cyberbully (Automatic)	Cyberbully (Automatic)
Not Cyberbully (Manual)	2,018	7
Cyberbully (Manual)	6	154
Total	2,024	161

Malay tweets. In this work, we have constructed a corpus of 219,444 tweets of common Malay cyberbullying words and 54,867 manually labelled by Malay language experts. We have also proposed a method that automatically labels short text data by using a combination of sentiment, emotion and toxicity polarities. Besides these, we have conducted an evaluation of the proposed method which yield high accuracy value, indicating that our method can be a suitable solution to automatic labelling of Malay tweets for cyberbullying context. Future work would include adding more Malay cyberbullying words.

ACKNOWLEDGMENTS

We are extremely grateful to the reviewers who have taken the time to give constructive comments and useful suggestions in the improvement of this article. This work is supported by the Malaysia Government under Fundamental Research Grant scheme (FRGS) at Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia (FRGS/1/2018/SSI01/UITM/03/1).

REFERENCES

[1] Abdullah Adeleke, Noor Azah Samsudin, Aida Mustapha, Nazri Mohd Nawi, Abdullahi O Adeleke #a, Noor A Samsudin #b, and Nazri M Nawi. 2017. Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses. *Int. J. Adv. Sci. Eng. Inf. Technol.*7, 4 (2017). DOI:https://doi.org/10.18517/ijaseit.7.4.2198

[2] Reem Alrashdi and Simon O'Keefe. 2020. Automatic Labeling of Tweets for Crisis Response Using Distant Supervision. In *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*, 418–425. DOI:https://doi.org/10.1145/3366424.3383757

[3] Pinar Arslan, Michele Corazza, Elena Cabrio, and Serena Villata. 2019. Overwhelmed by negative emotions? Maybe You Are Being Cyber-bullied! *Proc. ACM Symp. Appl. Comput.*Part F1477, (2019), 1061–1063. DOI:https://doi.org/10.1145/3297280.3297573

[4] Vimala Balakrishnan. 2015. Cyberbullying among young adults in Malaysia: The roles of gender, age and Internet frequency. *Comput. Human Behav.*46, (2015), 149–157. DOI:https://doi.org/10.1016/j.chb.2015.01.021

[5] Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia. 2020. Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Comput. Secur.*90, (2020), 101710. DOI:https://doi.org/10.1016/j.cose.2019.101710

[6] Vimala Balakrishnan, Shahzaib Khan, Terence Fernandez, and Hamid R. Arabnia. 2019. Cyberbullying detection on twitter using Big Five and Dark Triad features. *Pers. Individ. Dif.*141, September 2018 (2019), 252–257. DOI:https://doi.org/10.1016/j.paid.2019.01.024

[7] Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2016. Automatic labelling of topics with neural embeddings. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, Association for Computational Linguistics, ACL Anthology, 953–963. Retrieved September 18, 2020 from https://github.com/attardi/wikiextractor/

[8] Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. 2014. Automatic labelling of topic models learned from Twitter by summarisation. In*52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, Association for Computational Linguistics, 618–624. DOI:https://doi.org/10.3115/v1/p14-2101

[9] Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*1, (2017), 409–419. DOI:https://doi.org/10.18653/v1/P17-1038

[10] Lishan Cui, Xiuzhen Zhang, Amanda Kimpton, and Daryl D'Souza. 2016. Automatic labelling of topics via analysis of user summaries. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*9877 LNCS, (2016), 295–307. DOI:https://doi.org/10.1007/978-3-319-46922-5_23

[11] E Dermatas and G Kokkinakis Speech. 1989. A system for automatic text labelling. In *First European Conference on Speech Communication and Technology*. Retrieved September 18, 2020 from https://www.isca-speech.org/archive/eurospeech_1989/e89_1382.html

[12] Hassan Fazayeli, Sharifah Mashita Syed-Mohamad, and Nur Shazwani Md Akhir. 2019. Towards auto-labelling issue reports for pull-based software development using text mining approach. In *Procedia Computer Science*, 585–592. DOI:https://doi.org/10.1016/j.procs.2019.11.160

[13] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1536–1545. Retrieved September 18, 2020 from http://opennlp.sourceforge.net/

[14] Jesus Gerardo Alvarado Mantecon, Hadi Abdi Ghavidel, Amal Zouaq, Jelena Jovanovic, and Jenny McDonald. 2018. A Comparison of Features for the Automatic Labeling of Student Answers to Open-ended Questions. In Proceedings of the 11th International Conference on Educational Data Mining, EDM 2018. Retrieved from http://dbpedia.org/page/Aorta

[15] Stuart McIlroy, Nasir Ali, Hammad Khalid, and Ahmed E. Hassan. 2016. Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. Empir. Softw. Eng. 21, 3 (2016), 1067–1106. DOI:https://doi.org/10.1007/s10664-015-9375-7

[16] Muhammad Okky Ibrohim, Erryan Sazany, and Indra Budi. 2019. Identify abusive and offensive language in indonesian twitter using deep learning approach. In Journal of Physics: Conference Series, Institute of Physics Publishing. DOI:https://doi.org/10.1088/1742-6596/1196/1/012041

[17] Elaheh Raisi and Bert Huang. 2016. Cyberbullying Identification Using Participant-Vocabulary Consistency. ICML Work. #Data4Good Mach. Learn. Soc. Good Appl. (June 2016). Retrieved September 19, 2020 from http://arxiv.org/abs/1606.08084

[18] Harsh Ranjan, Sumit Agarwal, Amit Prakash, and Sujan Kumar Saha. 2018. Automatic labelling of important terms and phrases from medical discussions. In 2017 Conference on Information and Communication Technology, CICT 2017, 1–5. DOI:https://doi.org/10.1109/INFOCOMTECH.2017.8340644

[19] Taishi Saito and Osamu Uchida. 2018. Automatic Labeling for News Article Classification Based on Paragraph Vector. 9th International Conference on Information Technology and Electrical Engineering, ICITEE 2017 2018-Janua

[20] Erica Sizemore. 2015. Youth Bullying: From Traditional Bullying Perpetration to Cyberbullying Perpetration and the Role of Gender. Electronic Theses and Dissertations. Retrieved September 19, 2020 from https://dc.etsu.edu/etd/2543

[21] Robert Slonje and Peter K. Smith. 2008. Cyberbullying: Another main type of bullying?: Personality and Social Sciences. Scand. J. Psychol. 49, 2 (2008), 147–154. DOI:https://doi.org/10.1111/j.1467-9450.2007.00611.x

[22] S Supraja, Kevin Hartman, Sivanagaraja Tatinati, and Andy W.H. Khong. 2017. Toward the automatic labeling of course questions for ensuring their alignment with learning outcomes. In Proceedings of the 10th International Conference on Educational Data Mining, EDM 2017, 56–63

[23] Tan Sri Lee Lam Thye. 2017. On the alert for cyberbullying. Star Online (2017), 1–3. Retrieved September 19, 2020 from https://www.thestar.com.my/opinion/letters/2017/04/11/on-the-alert-for-cyberbullying/

[24] Xiaojun Wan and Tianming Wang. 2016. Automatic labeling of topic models using text summaries. 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap. 4, (2016), 2297–2305. DOI:https://doi.org/10.18653/v1/p16-1217

[25] Zuraini Zainol, Sharyar Wani, Puteri Nohuddin, Wan Noormanshah, and Syahaneim Marzukhi. 2018. Association Analysis of Cyberbullying on Social Media using Apriori Algorithm. Int. J. Eng. Technol. 7, December (2018), 72–75. DOI:https://doi.org/10.14419/ijet.v7i4.29.21847

[26] Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao. 2018. Scale up event extraction learning via automatic training data generation. 32nd AAAI Conf. Artif. Intell. AAAI 2018 (2018), 6045–6052

[27] H. Zolkepli. Malaya, Natural-Language-Toolkit library for bahasa Malaysia, powered by Deep Learning Tensorflow. Retrieved from https://github.com/huseinzol05/malaya