PENGESANAN BULI SIBER BAHASA MELAYU DI TWITTER MENGGUNAKAN PEMBELAJARAN MESIN

Nur Izaty Hana Binti Azmi
Assoc.Prof.Dr.Nazlia Binti Omar

Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Buli siber adalah masalah yang semakin meningkat yang boleh memberi impak buruk kepada generasi masa kini. Buli siber adalah aktiviti yang berlaku di dalam peranti digital seperti telefon bimbit, komputer, dan tablet. Buli siber boleh berlaku melalui laman rangkaian sosial dalam talian seperti media sosial, forum, atau permainan video di mana orang dapat melihat, menulis komen, atau berkongsi maklumat. Pada masa kini, terdapat pelbagai kajian yang memfokuskan kepada pengesanan perkataan buli di dalam bahasa Inggeris, namun perkataan buli di dalam bahasa Melayu adalah terhad untuk menangani isu ini. Kajian ini bertujuan untuk mengesan buli siber di Twitter menggunakan perkataan bahasa Melayu yang dikategorikan sebagai buli. Teknik yang digunakan adalah pembelajaran mesin yang memfokuskan kepada pengekstrakan ciri dan model pengelas Naïve Bayes yang membantu dalam mengesan twit yang dikategorikan sebagai buli. Data yang digunakan sebagai percubaan bagi pengelas Bayesian diperoleh daripada platform Twitter API yang menyediakan akses luas ke data Twitter awam yang telah dipilih pengguna untuk dikongsi. Analisis yang dijalankan di dalam kajian dijadikan sebagai panduan untuk pengkaji mengenal pasti ciri-ciri dalam mengesan buli siber dan memberikan model yang tepat bagi kajian ini. Hasil akhir, model pengelas berjaya memberikan keputusan yang baik dalam pengujian yang dijalankan. Selanjutnya, isu dan cabaran di dalam mengesan buli siber diketengahkan dan dibincangkan.

1 PENGENALAN

Buli siber adalah bentuk gangguan yang menggunakan media elektronik. Ia juga dikenali sebagai buli dalam talian. Buli siber berlaku apabila seorang individu atau berkumpulan mengganggu orang lain di Internet dan ruang digital terutama di laman media sosial dengan perbuatan mengugut, menyebarkan khabar angin yang memalukan mangsa, mendedahkan maklumat peribadi mangsa dan menulis kata-kata yang kesat. Buli siber juga merupakan tindakan berulang yang disengajakan, dilakukan sama ada individu atau berkumpulan ditujukan kepada sasaran yang tidak mampu mempertahankan diri mereka dan takut untuk melaporkan kepada pihak berkuasa apa yang terjadi kepada mereka. Disebabkan mangsa berdiam diri tanpa mengambil tindakan, ia mendorong pembuli untuk bertindak lebih agresif.

Pada zaman teknologi kini, media sosial menjadi satu platform yang penting bagi masyarakat untuk berkomunikasi antara satu sama lain tanpa batasan. Media sosial adalah gabungan daripada perkataan media dan sosial. Bagi memahami maksud media sosial dengan lebih mudah perlu terlebih dahulu mengetahui maksud media. Menurut Kamus Dewan Edisi Keempat, media merupakan alat atau perantara komunikasi seperti radio, televisyen dan akhbar yang dapat menyampaikan maklumat kepada orang ramai dalam masa yang singkat. Sosial pula membawa maksud segala hal yang berkaitan kemasyarakatan (Kamus Pelajar Edisi Kedua) seperti bergaul atau bercampur dengan masyarakat.

Media sosial didefinisikan sebagai sebuah kumpulan aplikasi berasaskan internet yang membangun atas dasar ideologi dan teknologi Web 2.0 dan yang membolehkan penciptaan dan pertukaran (Andreas Kaplan & Michael Haenlein, 2010). Ia meliputi pelbagai aplikasi termasuk Twitter, Instagram, Facebook, blog, dan forum. Media sosial memberi pengaruh yang besar kepada orang ramai dari pelbagai perspektif dan tujuannya digunakan kebanyakan pengguna yang menjadikan media sosial sebagai platform untuk meluahkan perasaan. Akan tetapi, tindakan ini telah diambil kesempatan oleh sesetengah pihak yang berniat buruk untuk mengaibkan atau mengugut. Hal seperti ini yang menyumbang kepada aktiviti pembulian siber di media sosial.

2 PENYATAAN MASALAH

Penggunaan media sosial yang semakin meningkat dalam kehidupan kini menyumbang kepada perkembangan buli siber yang semakin mendapat perhatian. Berdasarkan sumber Ipsos, Malaysia menduduki tangga ke-6 dan ke-2 di Asia melalui tinjauan yang dilakukan terhadap orang dewasa berumur 16-84 tahun di 28 buah negara seluruh dunia pada tahun 2018-2020. Laman media sosial dan aplikasi merupakan ruang digital yang biasa digunakan untuk pembulian siber. Media sosial telah menjadisebahagian daripada kehidupan setiap orang dan tidak menghairankan itu digunakan sebagai medium untuk melakukan pembulian siber. Pengesanan buli siber dalam teks media sosial bahasa Melayu perlu dikaji dengan lebih mendalam terutamanya dalam mengesan buli siber di media sosial terutamanya Twitter. Hal ini kerana kajian sedia ada banyak memfokuskan kepada teks media sosial bahasa Inggeris sahaja.

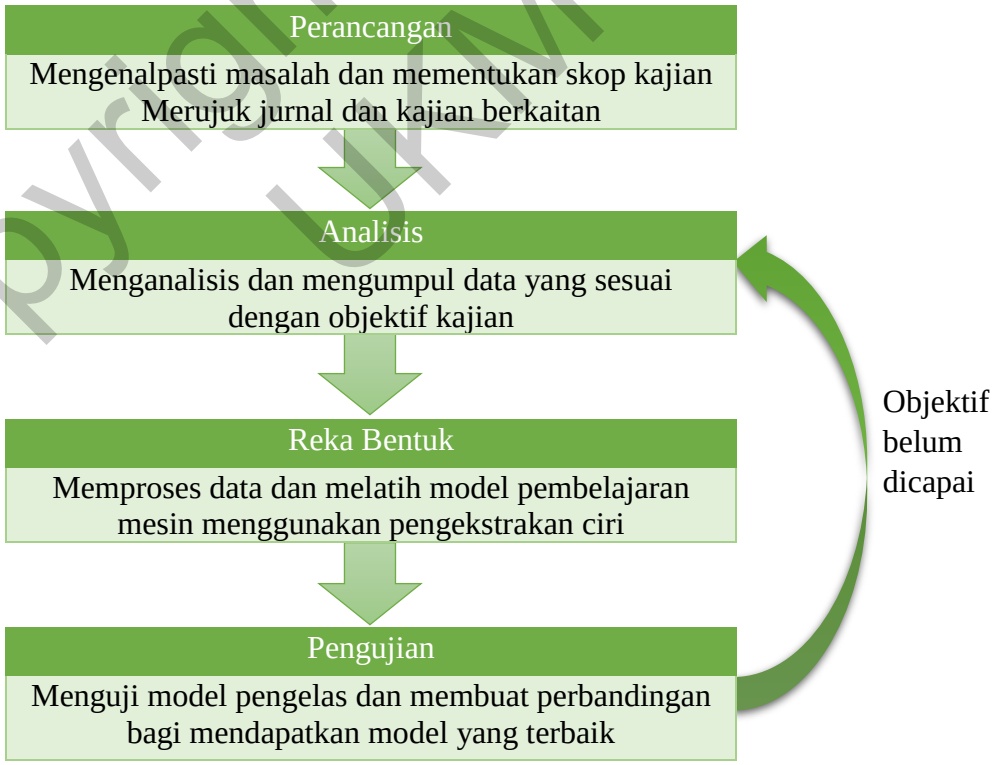
3 OBJEKTIF KAJIAN

Matlamat utama kajian ini bertujuan untuk mengesan twit Bahasa Melayu berunsur buli yang dimuatnaik pengguna media sosial Twitter menggunakan kaedah pembelajaran mesin. Objektif kajian ini adalah:

a. Mengenalpasti dan menggunakan pengekstrakan ciri bagi mengesan perkataan buli.
b. Membangunkan beberapa model pengelas bagi menilai ketepatan model berdasarkan pengekstrakan ciri.
c. Membuat perbandingan untuk mendapatkan model pengelas yang memberikan keputusan yang tepat.

4 METOD KAJIAN

Pembangunan model pembangunan yang sesuai penting untuk memastikan projek berjalan dengan baik dan mendapatkan hasil keputusan yang tepat dan berkualiti. Model pembelajaran mesin melibatkan beberapa fasa dan ditambah dengan penggunaan perkakasan dan perisian. Pembelajaran mesin merupakan aplikasi kecerdasan buatan (AI) yang menyediakan sistem kemampuan untuk belajar dan memperbaiki secara automatik daripada pengalaman tanpa diprogram secara jelas. Pembelajaran mesin ditakrifkan sebagai kemampuan komputer untuk belajar sendiri bagaimana untuk membuat keputusan menggunakan data percubaan. Pembelajaran mesin diperlukan untuk tugas-tugas yang terlalu kompleks sehingga tidak realistik bagi manusia untuk dikodkan sebagai situasi dalam buli siber. Fasa pembangunan termasuk fasa perancangan, analisis, reka bentuk, dan pengujian. Rajah 1 menunjukkan metod pembangunan model pembelajaran mesin yang digunakan untuk mengesan twit bahasa melayu yang berunsur buli.



Rajah 1 Metod Pembangunan Model Pembelajaran Mesin

4.1 Fasa Perancangan

Fasa ini melibatkan pengumpulan dan analisis keperluan kajian seperti set data yang berkaitan dengan objektif kajian dan menganalisis sorotan kajian susastera yang melibatkan pengumpulan, pencarian dan pembacaan jurnal dan kajian lepas bagi mendapatkan idea serta dijadikan rujukan. Contoh topik yang dikaji adalah berkaitan pengesanan buli siber di media sosial menggunakan pembelajaran mesin sedia ada. Set data yang berkaitan kajian daripada Twitter dan *Mendeley Data* dikumpulkan.

4.2 Fasa Analisis

Fasa ini melibatkan cara pengumpulan data dan analisis terhadap data terkumpul. Ini bertujuan untuk memastikan data yang dikumpulkan sesuai dengan model yang dipilih dengan melalui pra-pemprosesan data dan pengekstrakan ciri. Sebanyak 1,646 set data berjaya dikumpul dan dilabelkan secara manual kepada dua kategori kelas iaitu buli dan bukan buli. Jumlah data bagi dua kategori kelas ini adalah sama iaitu sebanyak 823. Data yang digunakan bagi menjalankan kajian ini diambil daripada platform Twitter.

Set data daripada Twitter dikumpulkan melalui aplikasi Twitter *API.* Set data yang dikumpul merupakan data mentah (*raw data*) dan disimpan dalam dokumen berformat CSV. Seterusnya, analisis data dilakukan untuk memastikan data hanya dalam bahasa Melayu dan pelabelan bagi kelas data kepada buli dan bukan buli dilakukan secara manual. Ini bertujuan untuk memudahkan data dianalisis. Analisis data dijalankan bagi memastikan data hanya dalam bahasa Melayu, dilabelkan terlebih dahulu secara manual, tiada data yang hilang, data yang berulang, mengetahui jenis data, statistik set data dan visualisasi taburan data bagi mendapatkan set data yang berkualiti.

4.3 Fasa Reka Bentuk

Dalam fasa ini, kaedah pra pemprosesan digunakan kerana ia merupakan salah satu elemen penting dalam projek ini. Pra pemprosesan teks secara tradisional merupakan langkah penting untuk tugas Pemprosesan Bahasa Tabii (NLP). Ia mengubah teks menjadi bentuk yang lebih mudah diproses sehingga algoritma pembelajaran mesin dapat menunjukkan prestasi yang lebih baik. Antara pra pemprosesan yang biasa digunakan adalah menukar kepada huruf kecil, tokenisasi, menyingkirkan kata henti*(stopwords)*, menghapus tanda baca, dan membuang karakter atau simbol yang sinonim dengan data Twitter seperti nama pengguna (@), tanda pagar (#), pautan Pelokasi Sumber Seragam (URL), dan ulang twit (RT). Rajah 2 menunjukkan menunjukkan output data yang belum diproses dan selepas diproses. Setelah data dibersihkan, data kemudiannya dibahagikan kepada dua bahagian iaitu data latihan sebanyak 80% manakala data pengujian sebanyak 20%.

	Twit	Twit_bersih
0	tak kisah la panggil awek masih macam bodoh. a...	kisah la panggil awek bodoh pilihan memanggil ...
1	Ibai bernasib baik saya bangun bodoh jika tida...	ibai bernasib bangun bodoh ah
2	VIDEO - cara untuk menginspirasi pelanggan unt...	video menginspirasi pelanggan tuli maklum bala...
3	oi pintu masuk begitu besar masih mahu memerah...	oi pintu masuk mahu memerah menolak bodoh eh sume
4	Siapkan tanah simi lanjiao lah LHL tanah sudah...	siapkan tanah simi lanjiao lhl tanah hadapan t...
5	Hshdjshs bodoh lah. Nak caj sejam ke camne?	hshdjsh bodoh nak caj sejam camn
6	saya adalah junior selama 4 tahun. kakak perem...	junior kakak perempuan tingkatan bodoh harfiah...
7	apa yang membuat anda berfikir bahawa menyindi...	berfikir bahawa menyindir okay bodoh
8	perkara seterusnya saya tahu dia berada di tan...	perkara tanah menumbuknya henti dirotan perhim...
9	Saya terkejut orang-orang ini masih menggunak...	terkejut orang orang teknik bodoh main jiwa
10	Kami jatuh cinta pada usia muda. Masa tu bodoh...	jatuh cinta usia muda tu bodoh gaduh pasal ben...
11	Mengapa perlu memperkosa gadis apabila anda bo...	memperkosa gadi puki palsu murah mencari gadi ...
12	setelah tweet ini saya google menerjemahkannya...	tweet googl menerjemahkannya menyedari bahawa ...
13	Bukan saya katakan "kejap ya puan, saya vaksin...	kejap ya puan vaksin bilik kejap maksudkan ber...
14	Adakah saya perlu mengikuti kelas pengurusan k...	adakah mengikuti kela pengurusan kemarahan men...
15	Atau seseorang harus memberinya tamparan kera...	memberinya tamparan kera kuat senang hati terp...
16	pernah (berkali-kali tbh) bermain badminton de...	berkali kali tbh bermain badminton dengannya g...
17	ini di sini secara terbuka membolehkan dan mem...	terbuka membolehkan membenarkan perkara bermas...
18	Memata-matai unta hanya dibenarkan untuk orang...	memata matai unta dibenarkan orang bodoh a bodoh
19	Dalam episod Rahsia CEO ini, Vinny Ribas dan T...	episod rahsia ceo vinni riba toni bodoh membin...

Rajah 2 Output data sebelum dan selepas diproses

Selain itu, kajian ini juga memerlukan keperluan perkakasan dan perisian untuk membangunkan algoritma dengan lancar. Jadual 1 dan 2 merupakan senarai perkakasan dan perisian yang diperlukan bagi kajian ini.

Jadual 1 Spesifikasi Keperluan Perkakasan

Peranti	**Komputer riba HP**
Sistem Operasi(OS)	Windows 10 (64-bit)
Pemproses	
Jenis Pemproses	Intel
Memori	
Saiz Memori (RAM)	4.00GB
SSD	512GB

Jadual 2 Spesifikasi Keperluan Perisian

Perisian	**Penerangan**
Google Chrome	Mencari maklumat
Microsoft Excel Office 2019	Menyimpan data dan maklumat dari Twitter dalam format fail .csv
Python	Bahasa pengaturcaraan utama
Google Colaboratory	Melaksanakan kod

Seterusnya, fasa ini juga merangkumi pengekstrakan ciri yang digunakan bagi melatih model pengelas untuk mengesan perkataan buli bahasa Melayu di dalam twit menggunakan data latihan. Set data input daripada twit yang diproses akan diubah menjadi satu set ciri untuk menjalankan tugas yang dikehendaki iaitu mengenalpasti perkataan bahasa Melayu berunsur buli yang terdapat dalam twit tersebut. Pengekstrakan ciri yang digunakan dalam kajian ini adalah *Bag of Words* (BoW) dan *Term Frequency-Inverse Document Frequency*(TF-IDF).

4.4 Fasa Pengujian

Dalam fasa ini, model pengelas yang dipilih bagi kajian ini dibangunkan. Terdapat empat model pengelas yang digunakan bagi kajian ini iaitu *Naïve Bayes, Logistic Regression, Random Forest* dan *Support Vector Machine* (*SVM*). Prestasi model-model ini diukur melalui keputusan ketepatan yang dijalankan dengan menggunakan data latihan sebanyak 80% dan data pengujian sebanyak 20%. Jadual 3 menunjukkan hasil keputusan kajian yang dibahagikan kepada ketepatan model bagi data latihan yang menggunakan pengekstrakan ciri dan ketepatan model bagi data pengujian tanpa menggunakan pengekstrakan ciri dalam mengesan twit bahasa Melayu berunsur buli.

Jadual 3 Hasil Analisis Model

Model \ Kejituan	**Naive Bayes**	**Logistic Regression**	**Random Forest**	**SVM**
Training Accuracy	72.24%	83.03%	95.30%	85.26%
Testing Accuracy	59.21%	76.32%	75.99%	72.04%

Model-model pembelajaran mesin yang digunakan dalam kajian ini dianalisis dengan menggunakan data yang dilatih melalui pengekstrakan ciri dan data pengujian tanpa melalui proses pengekstrakan ciri. Bagi data yang dilatih, *Random Forest* mencapai ketepatan yang paling tinggi iaitu 95.30%. Nilai ketepatan ini juga bergantung kepada pengekstrakan ciri yang digunakan. Bagi data pengujian, *Logistic Regression* mencapai ketepatan yang paling tinggi iaitu 76.32% di mana ia memberi prestasi yang baik tanpa bantuan pengekstrakan ciri. Setiap model pengelas juga mempunyai cara dan nilai yang berbeza dalam menguji data mengikut situasi.

5 HASIL KAJIAN

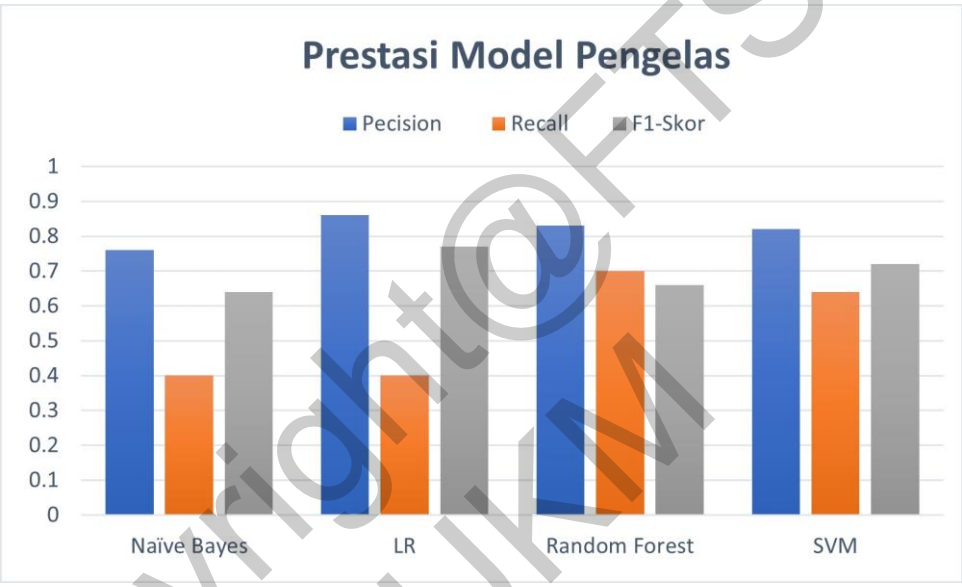
Melalui pengekstrakan ciri yang dipilih iaitu TF-IDF, senarai perkataan buli bahasa Melayu yang terdapat di dalam set data dapat dikesan. Rajah 3 menunjukkan *Word Cloud* perkataan buli bahasa Melayu yang dikesan menggunakan TF-IDF.

Rajah 3. Perkataan buli bahasa Melayu

Selain itu, hasil kajian ini juga adalah membuat perbandingan dan mendapatkan model pengelas dengan prestasi yang baik bagi kajian ini. Penilaian terhadap prestasi bagi setiap pengelas dibentangkan di dalam jadual dan dibincangkan. Jadual 4 merupakan hasil penilaian dan perbandingan bagi empat pengelas yang digunakan untuk mengesan twit berunsur buli yang diukur dari penilaian *Precision, Recall,* dan F1-Skor. Rajah 4 menunjukkan visualisasi bagi prestasi model.

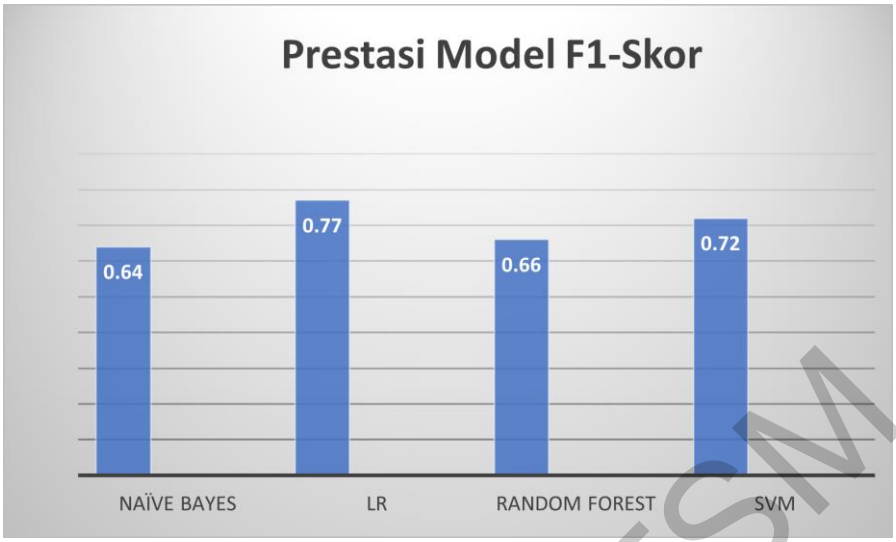
Jadual 4 Hasil perbandingan prestasi model pengelas

	Precision	***Recall***	**F1-Skor**
Naïve Bayes	0.76	0.40	0.64
Logistic Regression	0.86	0.70	0.77
Random Forest	0.83	0.70	0.66
SVM	0.82	0.64	0.72



Rajah 4. Prestasi model pengelas

F1-skor adalah salah satu penilaian metrik yang paling efektif jika dibandingkan dengan ketepatan dalam mendapatkan model pengelas yang terbaik. Ini kerana F1-skor sangat membantu jika taburan kelas bagi set data adalah tidak seimbang. Untuk mendapatkan F1-skor yang tinggi, pemilihan model pengelas amat penting bagi kajian berkaitan. Rajah 5 menunjukkan visualisasi prestasi model pengelas berdasarkan nilai F1-skor. *Logistic Regression* memberikan prestasi yang paling baik iaitu dengan nilai 0.77 bagi masalah klasifikasi binari dan berfungsi lebih baik dengan bertambahnya bilangan set data.



Rajah 5. Prestasi model berdasarkan F1-Skor

6 KESIMPULAN

Kesimpulannya, dokumen ini menjelaskan secara ringkas pengesanan buli siber bahasa Melayu di Twitter menggunakan pembelajaran mesin. Penggunaan ciri pengekstrakan dapat membantu meningkatkan ketepatan model pengelas mengesan perkataan buli bahasa Melayu yang terdapat di dalam twit dan mengurangkan data yang berlebihan. Perbandingan yang dijalankan bagi model-model pengelas adalah bertujuan untuk mendapatkan model terbaik bagi menyelesaikan masalah dalam kajian ini dan membantu para pengkaji untuk memilih model pengelas yang efektif bagi kajian mereka untuk mendapatkan keputusan yang bagus dan tepat. Kajian bagi pembulian siber dalam bahasa Melayu adalah amat terhad berbanding bahasa Inggeris. Oleh itu, kajian dalam bahasa Melayu perlu diperbanyakkan lagi supaya dapat menambahbaik kajian ini dan kajian yang sedia ada bagi membendung masalah buli siber dengan lebih berkesan.

7 RUJUKAN

Masoom Patel, M. H. (2020). *Bully Identification with Machine Learning Algorithms.* Jurnal of Critical Value.

Samaneh Nadali. (2013). *A Review of Cyberbulllying Detectiona: An Overview*. Conference Paper

Frotunas.M (2020). *Combining Textual Features to Detect Cyberbullying in Social Media Postst*. Procedia Computer Science, 612-621.

Mercado R. N.M (2018). *Automatic Cyberbullying Detection in Spansh-language Socisl Network using Sentiment Analysis Techniques*. International Journal of Advanced Computer Science an Applications, Vol.9.

Alharbi B.Y(2019). *Automatic Cyberbullying Detection in Arabic Social Media.* International Journal of Engineering Research and Technology. Vol.12.

Azianura Hani Shaari (2019). *Buli Siber: Ketidaksamaan Bahasa dan Etika Media Sosial Dalam Kalangan Remaja Malaysia*. Journal of Social Sciences and Humanities Vol.16.

Tan Jie Mei, (n.p). *ANALISIS SENTIMEN MENGGUNAKAN TEKS AGRESIF DALAM PENGESANAN BULI SIBER.*

B.Sri Nandhini, J.I.Sheeba (2015). *Online Social Network Bullying Detection Using Intelligent Techniques.*

Manpeet Singh, Maninder Kaur (2019). *Content-based Cybercrime Detection: A Concise Review*. International Journal of Innovative Technology and Exploring Engineering, Vol.8.

Khaerunnisa, D. (6 Jan, 2019). *Analisis Menggunakan NaiveBayes dengan Python pada Data Perawatan Kutil dengan Cryotherapy*. Retrieved from medium.com: https://medium.com/@16611055/analisis-menggunakan-naivebayes-dengan-python-pada-data-perawatan-kutil-dengan-cryotherapy-20ee5bc90561

Kharwal, A. (24 December, 2020). *What is Sentiment Analysis?* Retrieved from the clever programmer: https://thecleverprogrammer.com/2020/12/24/what-is-sentiment-analysis/

ML | Label Encoding of datasets in Python. (7 August, 2019). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

MonkeyLearn (2020). *Guide to Text Classification with Machine Learning & NLP*. Retrieved from MonkeyLearn: https://monkeylearn.com/text-classification/

MonkeyLearn(2020). *The Beginner's Guide to Text Vectorization*. Retrieved from MonkeyLearn: https://monkeylearn.com/blog/beginners-guide-text-vectorization/

Stecanella, B. (11 May, 2019). *What is TF-IDF?* Retrieved from monkeylearn: https://monkeylearn.com/blog/what-is-tf-idf/

JAIN, K. (2015, January 5). *Analytics Vidhya*. Retrieved from Scikit-learn(sklearn) in Python – the most important Machine Learning tool I learnt last year!: https://www.analyticsvidhya.com/blog/2015/01/scikit-learn-python-machine-learning-tool/

Swarnkar, N. (2020, May 21). *Quant Insti*. Retrieved from VADER Sentiment Analysis in Algorithmic Trading: https://blog.quantinsti.com/vader-sentiment/

Kshirsagar.V(Dec 25, 2019). *Detecting Hate tweets-Twitter Sentiment Analysis.* Retrieved from towardsdatascience: https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6

Raschka, S. (Oct 4, 2014). *Naïve Bayes and Text Classification- Introduction and Theory*. Retrieved from article: https://sebastianraschka.com/Articles/2014_naive_bayes_1.html#n-grams

Barua, J. (Aug 5, 2020). *Word Embeddings Versus Bag-of-Words: The Curious Case of Recommender Systems*. Retrieved from medium: https://medium.com/swlh/word-embeddings-versus-bag-of-words-the-curious-case-of-recommender-systems-6ac1604d4424

Hani J, Nashaat M, Ahmed M, Emad Z & Amer E (2019). *Social Media Cyberbulyying Detection using Machine Learning*. International Journal of Advanced Computer Science and Applications Vol.10.

Nakul L.(2020). *Hate Speech Detection in Scoial Media using Python*. https://github.com/NakulLakhotia/Hate-Speech-Detection-in-Social-Media-using-Python/blob/master/final_customization.ipynb.(2021)

dhavalpotdar(2019). *Cyberbullying Detection*. https://github.com/dhavalpotdar/cyberbullying-detection