Implementasi algoritma classification and regression tree untuk mendeteksi multi-label hate speech dan abusive language pada twitter bahasa indonesia

Implementation of classification and regression tree algorithm for multi-label hate speech and abusive language detection in twitter indonesian language

Fauzi Ihsan, Surya Agustian[*)], Iwan Iskandar, Nazruddin Safaat H.

Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau
Jl. H.R. Soebrantas km 11.5 Simpang Baru Panam, Pekanbaru, Riau, Indonesia 28293

Cara sitasi: F. Ihsan and S. Agustian, I. Iskandar, N. Safaat H., "Impementasi algoritma classification and regression tree untuk mendeteksi multi-label hate speech dan abusive language pada twitter bahasa indonesia," *Jurnal Teknologi dan Sistem Komputer*, vol. x, no. x, pp. xx-xx, 202x. doi: 10.14710/jtsiskom.x.x.202x.xx-xx, [Online].

Abstract – Hate speech and abusive language are negative action that often occur in our environment. Moreover, with the advancement of technology, anyone can spread hate speech or abusive language to anyone he wants. Disputes often occur between the respective interested parties. One of them is through social media twitter. By tweeting hate speech and being re-tweeted by a group of others. However, it is almost indistinguishable that a tweet includes hate speech or abusive language. This research uses 13,126 twitter data. The dataset is divided into 90% training data, 10% testing data and 80% training data, 20% testing data. Weighting using Word2Vec FastText along 128 vectors. The application of features engineering in this study using the CART algorithm increases the average accuracy of hate speech labels, abusive, and levels from 70% to 71.28 for 90% training data and 10% test data, while for training data 80% and 20% accurate data test - previous expectations of 68.74% to 70.56%.

Keywords *– hate speech; abusive language; classification; twitter; CART*

Abstrak - Ujaran kebencian dan bahasa kasar adalah suatu tindakan negatif yang sering terjadi dilingkungan kita. Terlebih lagi dengan adanya teknologi yang semakin maju, siapa saja bisa melakukan penyebaran ujaran kebencian atau bahasa kasar kepada siapa saja yang ia kehendaki. Sering terjadi pertikaian antara masing-masing pihak yang berkepentingan. Salah satunya melalui media sosial twitter. Dengan melakukan sebuah* tweet *ujaran kebencian dan di* re-tweet *oleh sekelompok lainnya. Namun hampir tak dapat dibedakan sebuah* tweet *itu apakah termasuk ke dalam ujaran kebencian ataupun* bahasa kasar. *Penelitian ini menggunakan sebanyak 13.126 data twitter. Dataset dibagi menjadi 90 % data latih, 10 % data uji dan 80 % data latih, 20 % data uji. Pembobotan mengunakan Word2Vec FastText sepanjang 128 vector. Penerapan rekayasa fitur/fitur enggineering pada penelitian ini yang menggunakan algoritma CART meningkatkan akurasi rata-rata dari label hate speech, abusive, dan level dari sebelumnya 70% menjadi 71,28 untuk data latih 90 % dan data uji 10 %, sedangkan untuk data latih 80 % dan data uji 20 % akurasi rata-ratanya sebelumnya sebesar 68,74 % menjadi 70,56 %.

Kata kunci *– ujaran kebencian; bahasa kasar; klasifikasi; twitter; CART*

[*)] Penulis korespondensi (Surya Agustian)
Email: surya.agustian@uin-suska.ac.id

I. PENDAHULUAN

Ujaran kebencian (*hate speech*) merupakan suatu ungkapan langsung maupun tidak langsung yang menuju kepada individu ataupun kelompok yang mengandung kebencian berdasarkan suatu hal yang melekat pada individu atau kelompok tersebut yang menyerang agama, etnis, *gender*, dan orientasi seksual. Ujaran kebencian adalah suatu perkataan, perilaku, tulisan ataupun suatu pertujukan yang dilarang karena dapat memicu timbulnya tindakan kekerasan dan sikap prasangka, baik itu dari pihak pelaku yang memberikan pernyataan tersebut ataupun korban dari tindakan tersebut [1].

Dalam kehidupan sehari-hari, media sosial menjadi tempat/wadah bagi individu ataupun kelompok dalam melakukan penyebaran ujaran kebencian dan sering juga disertai dengan bahasa kasar (*abusive language*) [2]. Bahasa kasar dalam bahasa indonesia biasanya diucapkan dan dituliskan untuk menyerang pihak tertentu, mengungkapkan kekesalan, kekecewaan ataupun meluapkan emosi terhadap peristiwa tertentu. Salah satu pengungkapan kata kasar dapat diungkapkan dengan menyebutkan jenis hewan tertentu, seperti anjing, monyet dan sebagainya. Namun tidak semua kalimat yang

memuat jenis hewan anggap kedalam bahasa kasar [3].

Media sosial membuat komunikasi cepat tersampaikan, hal tersebut merupakan hal positif yang kita dapatkan dari penggunaan media sosial, namun bukan berarti kita bisa dengan bebas menggunakannya sesuai dengan keinginan kita. Media sosial juga memiliki peraturan yang diatur oleh negara agar tidak terjadi penyalahgunaan media sosial yang dapat merugikan orang lain. Meski telah diatur oleh negara, tak sedikit juga orang-orang yang tetap menyalahgunakan media sosial untuk kepentingan pribadi.

Bentuk penyalahgunaan media social diantaranya, penipuan, penyebaran ujaran kebencian (*hate specch*), bahasa kasar (*abusive language*), dan lain-lain. Hal-hal tersebut bisa saja terjadi kepada siapapun, oleh karena itu, perlunya pengetahuan mengenai penggunaan media sosial agar dapat terhindar dari orang-orang yang menyalahgunakan media sosial untuk kepentingan pribadi.

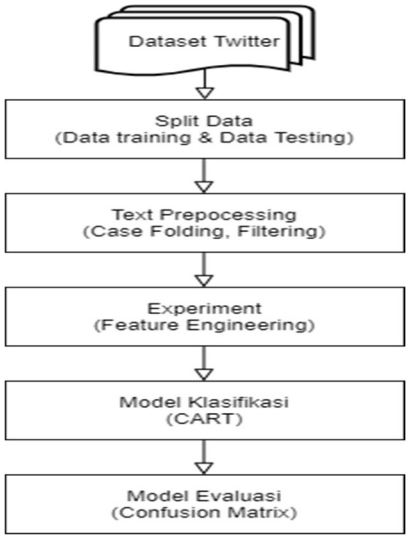
Pelanggaran yang dilakukan dengan mencemar nama baik orang lain ataupun memfitnah dan perbuatan tidak menyenangkan merupakan perbuatan yang melanggar hukum dikarenakan melanggar hak asasi orang lain. Meskipun perbuatan tersebut tidak secara langsung dimuka umum, namun akhir-akhir ini perbuatan tersebut sering dilakukan di dunia maya atau media sosial, karena di media sosial masyarakat merasa bebas memberikan pendapatnya ataupun mengkritik sesorang maupun kelompok, dan merasa tidak melanggar hukum karena tidak adanya kontak fisik dengan orang yang dikritik.

Kelebihan *twitter* dibanding dengan media sosial lainnya menurut [4] diantaranya adalah jangkauannya luas, tidak hanya teman, tetapi juga mampu menjangkau publik figur, potensi periklanan di masa mendatang lebih besar, komunikasi terjadi sangat cepat (*update*), *mulitilink* (terhubung dengan banyak jaringan) dan lebih terukur dari *facebook*. *Twitter* membantu penyebaran informasi secara lebih cepat yang kemudian akan menjadi sebuah topik yang dibahas oleh para penggunanya. Media massa seperti televisi, koran, majalah, tabloid pun menggunakan *twitter* sebagai penyampai berita-beritanya. Hal ini mempermudah masyarakat memperoleh informasi secara cepat dan *update* karena berita dapat di *update* setiap saat oleh media massa melalui *twitter*.

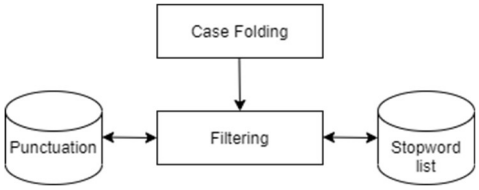
Namun algoritma yang digunakan pada penelitan tersebut diatas masih merupakan algoritma standar yang belum dioptimalkan untuk meningkatkan hasil akurasinya. Ada beberapa cara yang bisa dilakukan untuk meningkatkan akurasi sebuah algoritma, salah satunya adalah dengan menggunakan teknik *feature engineering*. Ketika dataset yang digunakan berskala besar, maka dapat digunakan algoritma CART untuk ditambah rekayasa fitur.

II. METODE PENELITIAN

Metodologi penelitian pada Gambar 1 menjelaskan alur penelitan ini. Ditarik dari garis besar terdapat 4 tahap yang paling penting pada penelitian ini, yaitu tahap *text preprocessing, training language model,* tahap *feature engineering* dan tahap klasifikasi. Dataset penelitian ini berasal dari penelitian yang dilakukan oleh [5], data diperoleh dengan menggunakan teknik *crawling,* kemudian pelabelan data dilakukan oleh 30 anotator dari berbagai latar belakang usia, pendidikan terakhir, pekerjaan, etnis dan agama. Dataset berupa teks



Gambar 1. Metodologi penelitian



Gambar 2. Tahap *text preprocessing*

Tabel 1. Persebaran dataset twitter

No	Label	Kelas	Jumlah	Total
1	Hate Speech	Hate Speech Tidak Hate Speech	5.552 7.574	13.126
2	Abusive	Abusive Tidak Abusive	5.034 8.092	13.126
3	Level	Lemah Sedang Kuat	3.375 1.704 473	5.552

cutian di twitter. Dataset terdiri dari 13.169 dengan 12 label data, namun terdapat data *noise* sebanyak 43 data, sehingga jumlah data yang digunakan pada penelitian ini 13.126 dan label yang digunakan terdiri dari tiga, yaitu *hate speech, abusive* dan level. Pada penelitian ini akan dilakukan pembagian dataset menjadi 80 % sebagai data latih 20 % sebagai data uji, dan 90 % sebagai data latih 10 % sebagai data uji.

Tabel 3. Rekayasa fitur (fitur tambahan yang akan dihitung frekuensinya)

Kategori Fitur	Fitur	Keterangan
Khusus Twitter	F1	Jumlah tag dalam cuitan
Tekstual	F2	Jumlah kata dalam cuitan
	F3	Jumlah tanda seru dalam cutian
	F4	Jumlah tanda tanya dalam cuitan
	F5	Jumlah kata huruf kapital dalam cuitan
	F6	Jumlah kata huruf kecil dalam cuitan
Lexicon	F7	Kata berkonotasi positif
	F8	Kata berkonotasi negatif
	F9	Kata yang mengandung *vocab abusive*

Tabel 2. Percobaan tanpa rekayasa fitur pada data latih 90 % dan data uji 10%

Case Folding	Stopword	Punctuation	Akurasi Hate Speech	Akurasi Abusive	Akurasi Level	Akurasi Rata-Rata
Ya	Ya	Ya	70,68 %	74,94 %	64,39 %	70 %
Ya	Ya	Tidak	68,62 %	74,18 %	61,87 %	68,22 %
Ya	Tidak	Ya	68,85 %	76,16 %	60,43 %	68,48 %
Ya	Tidak	Tidak	69,76 %	75,70 %	62,23 %	69,23 %
Tidak	Ya	Ya	67,94 %	72,35 %	59,35 %	66,55 %
Tidak	Ya	Tidak	69,23 %	73,95 %	62,59 %	68,59 %
Tidak	Tidak	Ya	69,38 %	73,72 %	60,43 %	67,84 %
Tidak	Tidak	Tidak	68,70 %	74,41 %	59,53%	67,55 %

Tabel 1. Percobaan tanpa rekayasa fitur pada data latih 80 % dan data uji 20 %

Case Folding	Stopword	Punctuation	Akurasi Hate Speech	Akurasi Abusive	Akurasi Level	Akurasi Rata-rata
Ya	Ya	Ya	69,38 %	72,51 %	63,37 %	68,42 %
Ya	Ya	Tidak	69,92 %	72,47 %	62,29 %	68,23 %
Ya	Tidak	Ya	69,12 %	73,76 %	61,75 %	68,21 %
Ya	Tidak	Tidak	71,13 %	72,85 %	61,30 %	68,43 %
Tidak	Ya	Ya	68,35 %	71,78 %	63,55 %	67,89 %
Tidak	Ya	Tidak	67,59 %	71,71 %	61,21 %	66,84 %
Tidak	Tidak	Ya	69,95 %	73,61 %	62,65 %	68,74 %
Tidak	Tidak	Tidak	70,18 %	72,51 %	61,31 %	68 %

Persebaran label dataset dapat dilihat pada **Error! Reference source not found.**.

Data yang didapat dari hasil *crawling* merupakan data mentah yang perlu dilakukan tahap *text processing*. Sebelum dilakukan proses klasifikasi pada data. Tahap *text preprocessing* yang dilakukan pada penelitian ini antara lain *case folding, tokenizing, punctuation,* dan *filtering* seperti pada Gambar 2 yang pernah juga dilakukan oleh [6] menggunakan data *facebook.* Namun, pada penelitian ini, peneliti tertantang untuk melakukan berbagai *experiment* terhadap data untuk mengklasifikasi data tanpa melakukan *text preprocessing* dikarenakan huruf besar, kecil dan tanda baca juga berpengaruh terhadap maksud dan tujuan dari cuitan seseorang.

Setelah data dibagi mejadi data latih dan uji, langkah selanjutnya adalah *training language model,* pada tahap ini dilakukan proses *word embedding* menggunakan *library FastText. FastText* merupakan *library* yang dikeluarkan oleh *facebook* dan *Fastext* merupakan pengembangan dari Word2Vec seperti yang pernah diteliti oleh [7], namun Word2Vec tidak mampu menangani kata yang tidak ada didalam *corpus* [8], sedangkan *FastText* mampu menangani kata yang tidak pernah dijumpai sebelumnya [9]. Pada tahap ini kata akan diubah kedalam *vector* sepanjang 128.

Hasil klasifikasi algoritma CART akan dilakukan *experiment* selanjutnya dengan merekayasa fitur baru atau disebut dengan *feature engineering. Feature engineering* merupakan suatu proses yang menggunakan pengetahuan untuk memilih *features* atau membuat *features* baru. *Feature engineering* yang dilakukan dapat dilihat pada Tabel 3.

Evaluasi model pada penelitan ini menggunakan *confusion matrix* guna untuk mengukur keakuratan algoritma dalam melakukan klasifikasi. *Confusion matrix* digunakan untuk mendapatkan nilai akurasi,

precision, dan *recall*. Implementasi penelitian ini menggunakan bahasa pemprograman python versi 3 dan menggunakan tool jupyter notebook.

III. HASIL DAN PEMBAHASAN

Metode standar CART menghasilkan rata-rata tertinggi dengan berbagai perpaduan percobaan *case folding, punctuation, stopword* pada 90 % data latih dan 10% data uji seperti yang ditunjukkan pada Tabel 2. Nilai akurasi tertinggi yang didapatkan yaitu sebesar 70 % dengan penggunaan *case folding, stopword,* dan *punctuation.*

Tabel 1 menunjukkan hasil akurasi rata-rata pada data latih 80% dan data uji 20%. Akurasi tertinggi sebesar 68,74 % tanpa proses *case folding, stopword,* dan menggunakan proses *punctuation.*

Selanjutnya dilakukan rekayasa fitur dengan melakukan pembahan panjang *vector FastText* dengan menghitung frekuensi fitur khusus pada twitter, fitur tekstual dan *lexicon*. Hasil untuk data latih 90% dan data uji 10% disajikan pada Tabel 4, Tabel 5, dan Tabel 6. Pada rekayasa fitur khusus twitter menghasilkan akurasi rata-rata sebesar 69,36 %, presisi sebesar 65,92 %, dan *recall* sebesar 60,94 %, pada rekayasa fitur tekstual diperoleh akurasi rata-rata sebesar 69,37 %, presisi sebesar 65,69 %, dan recall sebesar 61,62 %, sedangkan pada rekayasa fitur tekstual diperoleh akurasi sebesar 70,30 %, presisi sebesar 68,73 %, dan recall sebesar 59,87 %.

Nilai akurasi rata-rata setelah dilakukan rekayasa fitur, pada fitur *lexicon* lebih tinggi dibandingkan dengan rekayasa fitur lainnya dan lebih tinggi dibandingkan tanpa rekayasa fitur. Akurasi sebelum dilakukan rekayasa fitur sebesar 70 % lihat pada Tabel 2, sedangkan setelah dilakukan rekayasa fitur *lexicon* akurasi menjadi 70,30 % lihat Tabel 6. Akurasi pada label *hate speech* dan *abusive* mengalami kenaikan, sedangkan pada label level mengalami penurunan, namun secara rata-rata akurasi bertambah sekitar 0,30 %

Tabel 7, Tabel 8, dan Tabel 9 menunjukkan hasil akurasi untuk data latih 80% dan data uji 20% dengan menghitung frekuensi fitur khusus, fitur tekstual dan fitur *lexicon* pada cuitan. Pada rekayasa fitur khusus diperoleh akurasi sebesar 68,82 %, presisi sebesar 64,57 %, dan *recall* 61,81 %, pada rekayasa fitur tekstual diperoleh akurasi sebesar 68,81 %, presisi sebesar 64,45 %, dan *recall* 62,08 %, sedangkan pada rekayasa fitur *lexicon* diperoleh akurasi rata-rata sebesar 70,31 %, presisi sebesar 66,48 % dan *recall* sebesar 65,24 %.

Nilai akurasi rata-rata setelah dilakukan rekayasa fitur *lexicon* menunjukkan lebih tinggi daripada rekayasa fitur khusus dan tekstual, dan akurasinya lebih tinggi dari tanpa penggunaan rekayasa fitur. Akurasi rata-rata sebelum dilakukan rekayasa fitur sebesar 68,74.% lihat Tabel 1, sedangkan setelah dilakukan rekayasa fitur lexicon akurasinya menjadi 70,31 % lihat Tabel 9. Akurasi pada label hate speech, abusive dan level mengalami kenaikan.

Tabel 4. Rekayasa fitur khusus pada data latih 90 % dan data uji 10 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	71,13 %	74,18 %	62,77 %	69,36 %
Presisi	69,21 %	65,78 %	62,77 %	65,92 %
Recall	58,21 %	61,83 %	62,77 %	60,94%

Tabel 5. Rekayasa fitur tekstual pada data latih 90 % dan data uji 10 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	71,13 %	74,03 %	62,95 %	69,37 %
Presisi	69,21 %	64,90 %	62,95 %	65,69 %
Recall	58,21 %	63,69 %	62,95 %	61,62 %

Tabel 6. Rekayasa fitur *lexicon* pada data latih 90 % dan data uji 10 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	71,13 %	77,91 %	61,87 %	70,30 %
Presisi	69,21 %	75,12 %	61,87 %	68,73 %
Recall	58,21 %	59,54 %	61,87 %	59,87 %

Tabel 7. Rekayasa fitur khusus pada data latih 80 % dan data uji 20 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	70,75 %	72,62 %	63,10 %	68,82 %
Presisi	65,24 %	65,37 %	63,10 %	64,57 %
Recall	65,83 %	56,49 %	63,10 %	61,81 %

Tabel 8. Rekayasa fitur tekstual pada data latih 80 % dan data uji 20 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	70,49 %	72,66 %	63,28 %	68,81 %
Presisi	64,67 %	65,41 %	63,28	64,45

Recall	66,37 %	56,59 %	% 63,28 %	% 62,08 %

Setelah didapat fitur *lexicon* memiliki akurasi tertinggi, maka pada penelitan ini lakukan pengujian terhadap ketiga fitur pada *lexicon* yaitu kata berkonotasi positif, kata berkonotasi negatif dan kata yang mengandung *vocab abusive, vocab abusive* didapat dari penelitian sebelumnya [5]. Diperoleh hasil tertinggi pada penggunaan rekayasa fitur *vocab abusive* dengan akurasi rata-rata sebesar 71,28 %, presisi sebesar 67,80 %, dan *recall* sebesar 64,83 % disajikan pada Tabel 10. Sedangkan pada data latih 80 % dan data uji 20 % diperoleh hasil akurasi tertinggi pada rekayasa *vocab abusive* dengan akurasi rata-rata sebesar 70,53 %, presisi sebesar 67,04 %, dan *recall* sebesar 64,27 % percobaan ini terdapat pada Tabel 11.

Tabel 9. Rekayasa fitur *lexicon* pada data latih 80 % dan data uji 20 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	70,53 %	77,65 %	62,74 %	70,31 %
Presisi	64,11 %	72,58 %	62,74 %	66,48 %
Recall	68,62 %	64,35 %	62,74 %	65,24 %

Tabel 10. Rekayasa fitur dengan menghitung frekuensi *vocab abusive* mendapatkan hasil akurasi tertinggi untuk data latih 90 % dan data uji 10 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	71,29 %	77,99 %	64,57 %	71,28 %
Presisi	67,70 %	71,12 %	64,57 %	67,80 %
Recall	62,50 %	67,43 %	64,57 %	64,83 %

Tabel 11. Rekayasa fitur dengan menghitung frekuensi *vocab abusive* mendapatkan hasil akurasi tertinggi untuk data latih 20 % dan data uji 20 %

Parameter	Hate Speech	Abusive	Level	Rata-rata
Akurasi	70,56 %	77,76 %	63,37 %	70,56 %
Presisi	65,16 %	74,74 %	63,37 %	67,76 %
Recall	65,10 %	59,54 %	63,37 %	62,67 %

Tabel 12. *Confusion matrix* untuk rekayasa fitur abusive count dengan hasil terbaik pada data 90 % latih dan 10 % data uji

Prediksi	Positif	Negatif	Total
Abusive	287	195	482
Tidak Abusive	97	734	831
Total	384	839	**1313**

Peningkatan menggunakan rekayasa fitur yang dilakukan berhasil untuk menaikkan akurasi meskipun hasil yang didapat tidak terlalu besar. Hal ini disebabkan pada algoritma CART jumlah data latih mempengaruhi hasil akurasi semakin besar data latih semakin besar akurasi yang didapatkan. Selain itu jumlah kelas juga mempengaruhi kinerja algoritma CART dalam mengklasifikasikan teks.

VI. KESIMPULAN

Penerapan rekayasa fitur dapat diterapkan sehingga meningkatkan akurasi algoritma CART dalam melakukan klasifikasi teks twitter dengan memberikan nilai akurasi terbaik dibandingkan tanpa rekayasa fitur. Model yang terbentuk pada penelitian ini dapat dipertimbangkan dalam kasus klasifikasi twitter guna untuk mendeteksi tweet yang mengandung ujaran kebencian, bahasa kasar dan mengukur level dari ujaran kebencian.

V. DAFTAR PUSTAKA

[1] M. Febriyani, "Analisis faktor penyebab pelaku melakukan ujaran kebencian (hate speech) dalam media sosial," *J. Linguist.*, vol. 3, no. 2, pp. 139–157, 2018.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language ∗," 2013.

[3] A. F. Hidayatullah, A. A. Fadila, K. P. Juwairi, and R. A. Nayoan, "Identifikasi Konten Kasar Pada Tweet Bahasa Indonesia," *J. Linguist. Komputasional*, vol. 2, no. 1, p. 1, 2019.

[4] E. D. Putra, *Menguak Jejaring Sosial*. 2014.

[5] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," pp. 46–57, 2019.

[6] A. Kresna, B. Arda, M. A. Fauzi, and B. D. Setiawan, "Identifikasi Ujaran Kebencian Pada Facebook Dengan Metode Ensemble Feature Dan Support Vector Machine," vol. 2, no. 12, 2018.

[7] K. Antariksa, Y. S. P. Wp, and D. Ernawati, "Klasifikasi Ujaran Kebencian pada Cuitan

dalam Bahasa Indonesia," vol. 10, pp. 164–171, 2019.

[8] J. Santoso, A. Dewa, B. Soetiono, E. Setyati, and E. M. Yuniarno, "Self-Training Naive Bayes Berbasis Word2Vec untuk Kategorisasi Berita Bahasa Indonesia," vol. 7, no. 2, pp. 158–166, 2018.

[9] Z. A. Arliyanti Nurdin, Bernadus Anggo Seno Aji, Anugrayani Bustamin, "Perbandingan kinerja Word Embedding Word2Vec, Glove dan FastText pada klasifikasi teks," *J. TEKNOKOMPAK*, vol. 14, no. 2, pp. 74--79, 2020.