Router : Jurnal Teknik Informatika dan Terapan
Volume. 1 No. 4 Desember 2023
e-ISSN: 3032-3312; dan p-ISSN: 3026-3611, Hal 20-32
DOI: https://doi.org/10.62951/router.v2i2.76
Tersedia: https://journal.aptii.or.id/index.php/Router

Studi Performa TF-IDF dan Word2Vec Pada Analisis Sentimen *Cyberbullying*

Ahmad Hilman Dani [1], Eva Yulia Puspaningrum [2], Retno Mumpuni [3]
[1,2,3] Universitas Pembangunan Nasional "Veteran" Jawa Timur
E-mail: ahmadhilmanlagi@gmail.com evapuspaningurm.if@upnjatim.ac.id retnomumpuni.if@upnjatim.ac.id

Abstract : *On August 14, 2023, Indonesia had approximately 228 million social media users, a number that is expected to continue growing to reach 267 million by 2028. Social media can be used to spread both positive and negative information, and one of the various negative effects is cyberbullying. Consequently, much research is conducted in the field of machine learning to develop sentiment analysis. One crucial step in sentiment analysis is word weighting. The two most common word weighting methods are TF-IDF and Word2Vec. These methods can be compared to determine which one produces better classification results, allowing cyberbullying sentiments on social media to be detected more accurately. Based on nine test scenarios, the final results showed that TF-IDF performed better than Word2Vec in this study, with an accuracy of 84%.*

Keywords*: Cyberbullying, Sentiment Analysis,SVM, TF-IDF, Word2Vec*

Abstrak : Indonesia memiliki sekitar 228 juta pengguna media sosial per 14 Agustus 2023, jumlah ini diperkirakan akan terus bertambah hingga mencapai 267 juta pada tahun 2028. Media sosial dapat digunakan untuk menyebarkan informasi positif maupun negatif, dan satu diantara berbagai efek negatifnya adalah cyberbullying. Oleh karena itu, banyak penelitian dilakukan dalam bidang *machine learning* untuk mengembangkan analisis sentimen. Salah satu langkah penting dalam analisis sentimen adalah pembobotan kata. Dua metode pembobotan kata yang paling umum adalah TF-IDF dan Word2Vec. Mereka dapat dibandingkan untuk menentukan mana yang memiliki hasil klasifikasi yang lebih baik, sehingga sentimen cyberbullying di media sosial dapat terdeteksi dengan lebih akurat. Berdasarkan Sembilan skenario uji, hasil akhir yang didapatkan adalah TF-IDF mampu bekerja lebih baik daripada Word2Vec pada penelitian ini dengan akurasi sebesar 84%.

Kata Kunci*: Cyberbullying, Sentimen Analisis,SVM, TF-IDF, Word2Vec.*

LATAR BELAKANG

Dewasa ini, ada banyak pengguna media sosial di Indonesia. Per 14 Agustus 2023, ada sekitar 228 juta pengguna media sosial, dan diperkirakan akan mencapai 267 juta pada tahun 2028 (Kemp, 2023). Bertambahnya pengguna media sosial ini mendorong pengembangan berbagai model analisis sentimen untuk berbagai tujuan, seperti mengetahui pendapat pengguna tentang produk tertentu, mengumpulkan opini masyarakat terkait suatu isu tertentu, dan lain-lain. Maka dari itu, pada penelitian ini dilakukan pembuatan sebuah model analisis sentimen *cyberbullying* dengan alasan peningkatan pengguna media sosial tidak selalu membawa dampak positif, melainkan juga dampak positif seperti perundungan verbal di dunia maya (*cyberbullying).*

Beberapa penelitian telah dilakukan untuk mengembangkan analisis sentimen. Penelitian yang dilakukan Tarissa Aura Azzahra, Nurul Anisa Sri Winarsih, Galuh Wilujeng Saraswati, Filmada Ocky Saputra, Muhammad Syaifur Rohman, Danny Oka

Received Mei 30, 2024; Accepted Juni 04, 2024; Published Juni 30, 2024

Ratmana, Ricardus Anggi Pramunendar, Guruh Fajar Shidik yang melakukan penelitian tentang analisis sentimen kebencaan. Data yang didapatkan berasal dari video YouTube dengan judul "[BREAKING NEWS] Jakarta Diguncang Gempa". Data yang didapatkan sebanyak 2.440 data komentar dari video tersebut. Pembobotan kata yang digunakan adalah TF-IDF dan terdapat dua metode klasifikasi yang digunakan untuk dibandingkan peformanya yakni SVM dan Naïve Bayes. Akurasi tertinggi yang didapatkan dari dua metode klasifikasi tersebut adalah metode SVM dengan nilai 92% (Aura Azzahra et al., 2024). Penelitian lain yang dilakukan oleh Jasmarizal, Rahmaddeni, Junadhi, M. Khairul Anam yang melakukan penelitian tentang analisis sentimen terkait produk *skincare* dengan data yang digunakan adalah data komentar ulasan produk MS Glow sebanyak 3.006 data. Pembobotan yang dipakai adalah TF-IDF dan metode klasifikasinya adalah SVM. Akurasi tertinggi yang didapatkan adalah 99.60% (Jasmarizal et al., n.d.) . Penelitian lain yang dilakukan oleh Lisyana Damayanti dan Kemas Muslim Lhaksmana yang melakukan penelitian tentang analisis sentimen tentang pemilihan umum presiden Indonesia tahun 2024. Data yang diambil berasal dari Twitter dengan cara crawling data dan mendapatkan 14.318 tweet. Pembobotan kata yang digunakan adalah Word2Vec dengan metode klasifikasi SVM. Penelitian ini mendapatkan hasil akurasi sebesar 90,43% (Damayanti & Lhaksmana, 2024). Penelitian lain yang dilakukan oleh Nurul Rezki, Sri Astuti Thamrin, Siswanto menghasilkan akurasi tertinggi terhadap analisis sentimen kebijakan kampus merdeka sebesar 89.87%. Penelitian ini menggunakan 10.000 data yang didapatkan dari Twitter dan dibobotkan menggunakan Word2Vec dengan metode klasifikasi SVM (Rezki et al., 2023).

Menurut empat penelitian yang telah dijelaskan pada paragraf sebelumnya, penelitian tersebut menggunakan metode klasifikasi yang sama, yaitu SVM. Namun, empat penelitian tersebut menggunakan dua metode pembobotan kata yang berbeda, yakni penelitian satu dan dua menggunakan TF-IDF dan penelitian tiga dan empat menggunakan Word2Vec. Proses pembobotan kata dilakukan karena ketidakmampuan mesin memproses data berupa huruf atau *string* secara langsung sehingga diperlukan perubahan dari *string* menjadi angka. TF-IDF merupakan metode untuk mempertimbangkan sepenting apa sebuah kata dalam dokumen yang merupakan bagian dari kumpulan dokumen (korpus) (Kim & Gil, 2019). Kelemahan dari TF-IDF adalah pembobotan kata yang tidak memperhatikan hubungan semantik antar kata. Sedangkan, Word2Vec merupakan metode untuk menghasilkan representasi vektor dari sebuah kata

dengan memperhatikan hubungan semantik antar kata melalui proses pelatihan menggunakan *neural network* (Mikolov et al., 2013). Padahal, penelitian yang dilakukan oleh Said A. Salloum, Rehan Khan, dan Khaled Shaalan menemukan bahwa semantik sangat penting untuk pengembangan proses pemrosesan bahasa alami (NLP) karena dengan adanya semantik, orang dapat mengatasi masalah seperti ambiguitas dan masalah lain yang terkait dengan NLP (Salloum et al., 2020). Sehingga, seharusnya peforma TF-IDF tidak lebih baik daripada Word2Vec. Namun, pada paparan sebelumnya, dua penelitian yang menggunakan TF-IDF menunjukkan bahwa TF-IDF bekerja lebih baik daripada Word2Vec.

KAJIAN TEORITIS

Analisis Sentimen

Analisis sentimen didefinisikan sebagai proses otomatisasi untuk menentukan sikap atau perasaan penulis terhadap suatu subjek, yang biasanya dikategorikan sebagai positif, negatif, atau netral (Medhat et al., 2014).

Cyberbullying

Cyberbullying adalah bentuk penindasan atau pelecehan yang terjadi melalui teknologi digital, terutama menggunakan internet dan perangkat komunikasi elektronik seperti ponsel, komputer, dan tablet (Patchin & Hinduja, 2006).

Data Preprocessing

Data preprocessing adalah tahap awal dalam alur kerja pembelajaran mesin dan analisis data di mana data mentah diubah atau diproses untuk meningkatkan kualitasnya sehingga dapat digunakan oleh model pembelajaran mesin (Kotsiantis et al., 2014).

TF-IDF

TF-IDF merupakan metode untuk mempertimbangkan sepenting apa sebuah kata dalam dokumen yang merupakan bagian dari kumpulan dokumen (korpus) (Kim & Gil, 2019). Secara matematis, rumus TF-IDF ialah sebagai berikut,

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t) \tag{1}$$

Dengan rumus tf(*t,d)* adalah sebagai berikut,

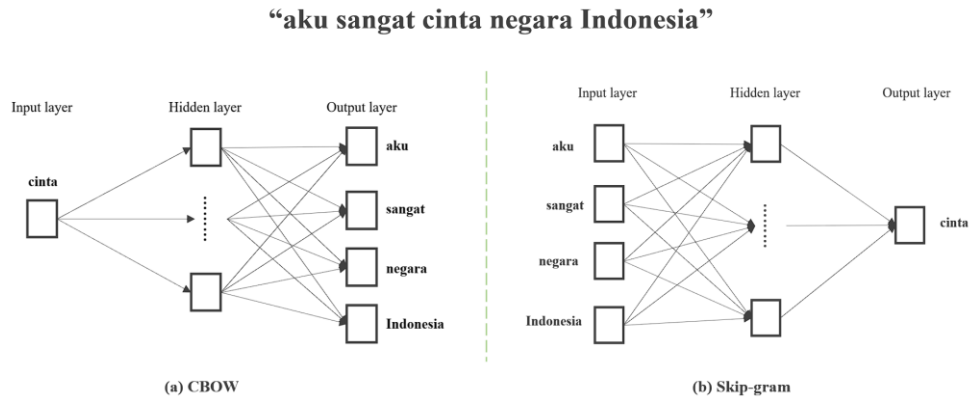
$$\text{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}, \tag{2}$$

Selanjutnya, rumus idf(*t)* adalah sebagai berikut,

$$\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1, \tag{3}$$

Word2Vec

Word2Vec merupakan satu diantara beberapa metode untuk menghasilkan representasi vektor dari sebuah kata melalui proses pelatihan menggunakan *neural network*. Terdapat dua arsitektur model yaitu CBOW dan Skip-gram. Perbedaan diantara keduanya adalah CBOW (*Continuous Bag of Words*) menghasilkan vektor kata dengan cara memprediksi sebuah kata berdasarkan kata disekitarnya, nantinya, kata yang berada disekitar kata yang diprediksi akan menjadi data di *input layer* pada saat pelatihan model. Sedangkan, Skip-gram menghasilkan representasi vektor dengan cara memprediksi kata yang ada disekitar sebuah kata, nantinya kata yang digunakan untuk memprediksi kata yang ada disekitarnya akan menjadi data di *input layer* pada saat pelatihan model (Mikolov et al., 2013). Berikut gambar 1 yang merupakan ilustrasi dari model CBOW dan Skip-gram,

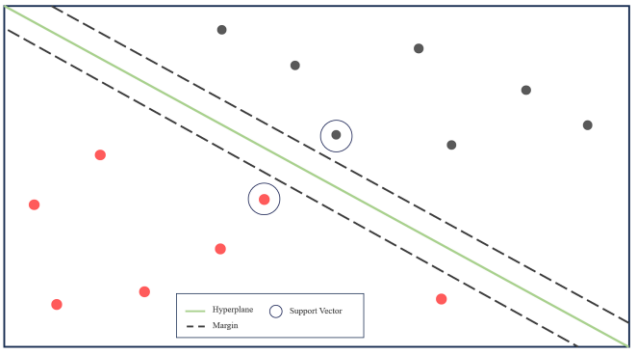


Gambar 1. Ilustrasi Word2Vec

Berdasarkan gambar 1 diatas, terdapat sebuah kalimat "aku sangat cinta negara Indonesia". Pada model CBOW, kata "cinta" akan digunakan di *input layer* untuk memprediksi kata "aku" sangat", "negara" "Indonesia". Sedangkan, pada model Skip-gram, kata "aku" sangat", "negara" "Indonesia" digunakan di *input layer* untuk memprediksi kata "cinta".

SVM

Support Vector Machine adalah *hyperplane* yang memisahkan data menjadi label yang berbeda dengan *margin* yang paling maksimal (Tong & Koller, 2001). Berikut gambar 2 yang merupakan ilustrasi SVM,

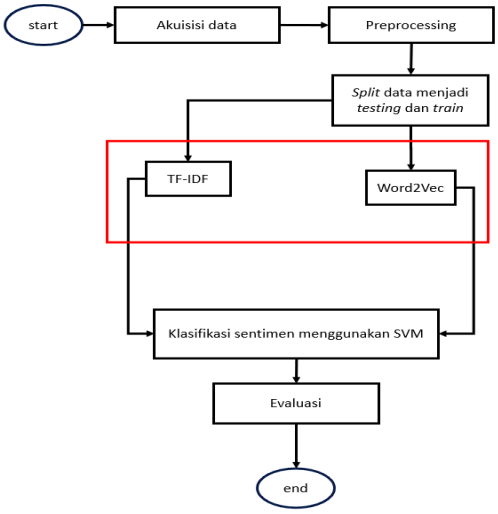


Gambar 2. Ilustrasi SVM

Berdasarkan gambar 2 diatas, titik yang diberi garis lingkaran bewarna merah disebut sebagai *Support vectors* yang merupakan titik terdekat dengan *hyperplane*, yang mempengaruhi posisi dan orientasi *hyperplane* tersebut. *Hyperplane* merupakan batas keputusan yang memisahkan dua kelas dalam ruang fitur. Pada gambar 2 diatas, *hyperplane* ditunjukan dengan garis hijau. Lalu, garis putus-putus hitam merupakan margin yang merupakan jarak antara *hyperplane* dan data *support vectotrs*.

METODE PENELITIAN

Berikut gambar 3 yang merupakan *flowchart* penelitian yang dilakukan,



Gambar 3. Flowchart Penelitian

Pada gambar 1 diatas, tahapan yang diberi kotak berwarna merah merupakan metode yang sedang dibahas pada penelitian ini.

Akuisisi Data

Data yang dipakai pada penelitian ini didapat dari kaggle.com pada tautan

https://www.kaggle.com/datasets/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text . Data yang didapatkan merupakan file .csv yang berisi tweet sebanyak 13.169 tweet (Ibrohim & Budi, 2019).

Data Preprocessing

Setelah data mendapatkan, dilakukan data preprocessing yang meliputi:

a. Menambahkan kolom 'is_bully'

Data yang didapatkan tidak memiliki kolom 'is_bully' karena data yang didapatkan adalah data sentimen yang mengandung unsur ujaran kebencian, kata kasar, atau tidak keduanya (netral). Maka dari itu, dibuatlah kolom 'is_bully' yang ketika sebuah sentimen terindikasi ujaran kebencian atau kalimat kasar, maka 'is_bully' bernilai 1 (*true*). Hal ini didasarkan pada Studi oleh Ni Nyoman Ayu Suciartini dan Ni Luh Putu Unix Sumartini menemukan bahwa perundungan secara verbal adalah pernyataan atau tulisan yang mengandung perendahan, pengejekan, fitnah, kritik kejam, pernyataan bernuansa pelecehan seksual, ancaman, atau gosip (Nyoman et al., 2018). Lalu, kolom yang tidak relevan akan dihapus.

b. *Case folding*

Case folding merupakan proses membuat seluruh huruf kapital menjadi *lowercase.* Proses ini dapat memastikan representasi teks yang konsisten dan koheren, yang penting untuk pengolahan dan analisis data teks (Ahmad Aliero et al., 2023).

c. *Cleaning*

Cleaning merupakan proses menghapus seluruh karakter atau huruf yang dapat mengganggu proses sentimen analisis. Pada penelitian ini, tahap *cleaning* meliputi penghapusan tanda baca, penghapusan angka, membuat karakter yang berulang menjadi tidak berulang seperti kata "iyaaaa" menjadi "iya". Proses ini dilakukan karena dapat membantu dalam mengurangi *noise* dalam data sehingga model dapat lebih fokus pada informasi yang relevan (Rahm & Do, 2000).

d. *Stemming*

Stemming merupakan proses mentrasnform kata menjadi kata dasar. Proses ini mampu mengurangi varian bentuk kata yang berbeda sehingga meningkatkan kemampuan sistem dalam mengenali dan memproses informasi dan mempercepat proses komputasi (Jivani et al., n.d.).

e. *Stopwords removal*

Stopwords removal merupakan proses menghapus seluruh kata yang tidak begitu penting namun terlalu sering muncul pada sebuah dokumen dalam korpus. Proses ini dilakukan karena dapat mengurangi dimensi data dan meningkatkan efisiensi komputasi (Al-Otaibi & Al-Rasheed, 2022).

f. Document selection

Sentiment selection merupakan proses memilih 3.000 dokumen untuk dijadikan bahan penelitian dari 13.169 dokumen yang didapatkan.

g. *Tokenization*

Tokenization merupakan membuat kamus berdasarkan korpus yang didapatkan dengan cara memecah setiap kata menjadi satu bagian tersendiri dari seluruh dokumen. Proses ini dilakukan karena membuat teks menjadi lebih mudah dianalisis oleh model machine learning (Toraman et al., 2022).

Split data menjadi *testing* dan *training*

Setelah tahap *data preprocessing* data dibagi menjadi *testing* dan *training* dengan pembagian 30% untuk data *testing* dan 70% untuk data *training*. Proses ini penting untuk dilakukan karena dapat membantu dalam mengevaluasi seberapa baik model dapat melakukan tugasnya di luar data *training*, sehingga dapat memberikan perkiraan realistis tentang kinerja model dalam aplikasi dunia nyata (Muraina, n.d.).

TF-IDF

Setelah data dibagi menjadi data *training* dan data *testing*, dilakukan proses pembobotan kata menggunakan TF-IDF menggunakan *librabry* dari sklearn.

Word2Vec

Selain dibobotkan menggunakan TF-IDF, data yang telah dibagi menjadi *testing* dan *training* juga akan dibobotkan menggunakan Word2Vec menggunakan librabry gensim.

SVM

Setelah itu, akan dilakukan klasifikasi menggunakan SVM dengan menggunakan library sklearn.

Evaluasi

Selanjutnya, dilakukan evaluasi menggunakan *confusion matrix* menggunakan library sklearn.

HASIL DAN PEMBAHASAN

Data yang didapatkan memiliki panjang 13.169 tweet. Data terdiri dari 12 kolom yakni:

1. Tweet: Tweet yang didapatkan
2. HS: kolom untuk label kalimat *hate-speech*
3. Abusive: kolom untuk label kalimat *abusive*
4. HS_Individual: label rinci dari ujaran kebencian yang ditujukan pada seseorang
5. HS_Group: label rinci dari ujaran kebencian yang ditujukan pada kelompok
6. HS_Religion: label rinci dari ujaran kebencian yang ditujukan pada agama
7. HS_Race: label rinci dari ujaran kebencian yang ditujukan pada ras
8. HS_Physical: label rinci dari ujaran kebencian yang ditujukan pada fisik
9. HS_Gender: label rinci dari ujaran kebencian yang ditujukan pada gender
10. HS_Weak: label rinci dari ujaran kebencian yang lemah
11. HS_Moderate: label rinci dari ujaran kebencian yang sedang
12. HS_Strong: label rinci dari ujaran kebencian yang kuat.

Sehingga, penelitian ini membuat kolom 'is_bully' dan menghapus kolom yang tidak relevan seperti yang telah dibahas pada sub-bab *data preprocessing*. Berikut merupakan hasil dari penambahan kolom dan penghapusan kolom yang tidak relevan. Berikut merupakan gambar 4 yang merupakan tampilan data setelah melalui proses ini,

Tweet	is_bully
Susilo Bambang Yudhoyono Presiden RI ke 6 meminta Khofifah Emil untuk dekat dengan masyarakat . No 1 Bisa ! #BaGusS	0
ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri:	0
tapi masih menikmati kebijakan yang dulu diusulkan PKI.. Iha yo opo...	1

Gambar 4. Data setelah dilakukan penambahan dan penghapusan kolom

Setelah itu, dilakukan *case folding* dan menghasilkan sentimen seperti pada gambar 5 dibawah ini,

before	after
susilo bambang yudhoyono presiden ri ke 6 meminta khofifah emil untuk dekat dengan masyarakat . no 1 bisa ! #baguss	susilo bambang yudhoyono presiden ri ke meminta khofifah emil untuk dekat dengan masyarakat no bisa
ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri:	ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri
tapi masih menikmati kebijakan yang dulu diusulkan pki.. Iha yo opo...	tapi masih menikmati kebijakan yang dulu diusulkan pki Iha yo opo

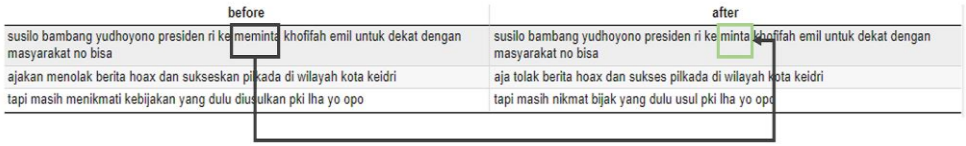
Gambar 5. Data setelah proses case folding

Setelah itu, dilakukan *cleaning* dan menghasilkan sentimen seperti pada gambar 6 dibawah ini,

before	after
susilo bambang yudhoyono presiden ri ke 6 meminta khofifah emil untuk dekat dengan masyarakat . no 1 bisa ! #baguss	susilo bambang yudhoyono presiden ri ke meminta khofifah emil untuk dekat dengan masyarakat no bisa
ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri:	ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri
tapi masih menikmati kebijakan yang dulu diusulkan pki.. Iha yo opo...	tapi masih menikmati kebijakan yang dulu diusulkan pki Iha yo opo

Gambar 6. Data setelah proses cleaning

Setelah itu, dilakukan *stemming* dan menghasilkan sentimen seperti pada gambar 7 dibawah ini,



before	after
susilo bambang yudhoyono presiden ri ke meminta khofifah emil untuk dekat dengan masyarakat no bisa	susilo bambang yudhoyono presiden ri ke minta khofifah emil untuk dekat dengan masyarakat no bisa
ajakan menolak berita hoax dan sukseskan pilkada di wilayah kota keidri	aja tolak berita hoax dan sukses pilkada di wilayah kota keidri
tapi masih menikmati kebijakan yang dulu diusulkan pki lha yo opo	tapi masih nikmat bijak yang dulu usul pki lha yo opo

Gambar 7. Data setelah proses stemming

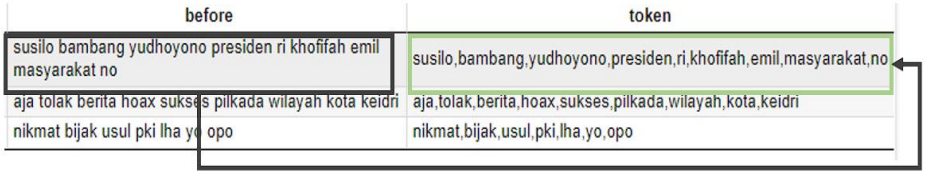
Setelah itu, dilakukan *stopwords removal* dan menghasilkan sentimen seperti pada gambar 8 dibawah ini,



before	after
susilo bambang yudhoyono presiden ri ke minta khofifah emil untuk deka dengan masyarakat no bisa	susilo bambang yudhoyono presiden ri khofifah emil masyarakat no
aja tolak berita hoax dan sukses pilkada di wilayah kota keidri	aja tolak berita hoax sukses pilkada wilayah kota keidri
tapi masih nikmat bijak yang dulu usul pki lha yo opo	nikmat bijak usul pki lha yo opo

Gambar 8. Data setelah proses Stopwords removal

Setelah itu, dilakukan *tokenization* dan menghasilkan sentimen seperti pada gambar 9 dibawah ini,



before	token
susilo bambang yudhoyono presiden ri khofifah emil masyarakat no	susilo,bambang,yudhoyono,presiden,ri,khofifah,emil,masyarakat,no
aja tolak berita hoax sukses pilkada wilayah kota keidri	aja,tolak,berita,hoax,sukses,pilkada,wilayah,kota,keidri
nikmat bijak usul pki lha yo opo	nikmat,bijak,usul,pki,lha,yo,opo

Gambar 9. Data setelah proses tokenization

Setelah proses *preprocessing* selesai, data dibagi menjadi data *training* dan *testing* dengan pembagian 70% untuk *training* dan 30% untuk *testing*. Setelah itu, dilakukan proses pembobotan kata menggunakan TF-IDF dengan potongan kode dibawah ini,

kode 1: *Training* TF-IDF

```
1  Tfidf_vect = TfidfVectorizer()
2  Tfidf_vect.fit(data_sentimen['token'])
3  Train_X_Count = Tfidf_vect.transform(Train_X)
4  Test_X_Count = Tfidf_vect.transform(Test_X)
```

Selain itu, dilakukan pula pembobotan kata menggunakan Word2Vec dengan potongan kode dibawah ini,

Code 2: *Training* Word2Vec

```
1  tokenized_tweet = data_sentimen['Tweet'].apply(lambda x: x.split())
```

```
model_w2v_cbow = gensim.models.Word2Vec(
        sentences = tokenized_tweet,
        window=5,
        min_count=2,
        sg = 1,
        hs = 0,
        negative = 10,
        workers= 32,
        seed = 34,
        epochs=30
)
```

Selanjutnya, dilakukan klasifikasi menggunakan SVM dengan tiga skenario uji yakni klasifikasi dengan SVM kernel linear dan pembobotan katanya TF-IDF, klasifikasi dengan SVM kernel linear dan pembobotan katanya Word2Vec CBOW, dan klasifikasi dengan SVM kernel linear dan pembobotan katanya Word2Vec Skip-gram.

Berdasarkan tiga skenario uji tersebut, hasil yang didapatkan melalui evaluasi menggunakan *confusion matrix* menggunakan library sklearn tersaji pada tabel 1 dibawah ini,

Tabel 1. Hasil *confusion matrix* dari tiga skenario uji

Skenario	non-bullying			bullying			Akurasi
	Precision	Recall	f1-Score	Precision	Recall	f1-Score	
TF-IDF, SVM kernel linear	83%	83%	83%	83%	83%	83%	83%
CBOW, SVM kernel linear	79%	75%	77%	76%	80%	78%	77%
Skip-gram, SVM kernel linear	80%	76%	78%	77%	80%	78%	78%

Berdasarkan tabel 1 tersebut, didapatkan analisis yakni,

TF-IDF dengan SVM Kernel Linear

Model pada skenario 1 terbilang efektif dalam melakukan analisis sentimen dilihat dari presisi, akurasi, dan f1-score yang didapatkan pada skenario ini.

Hasil yang sangat seimbang antara precision dan recall untuk kedua kelas (non-bullying dan bullying). Selain itu, f1-Score yang tinggi menunjukkan bahwa metode ini efektif dalam mengidentifikasi kedua jenis sentimen tersebut.

Word2Vec CBOW dengan SVM Kernel Linear:

Model pada skenario 2 terbilang cukup efektif dalam melakukan analisis sentimen dilihat dari presisi, akurasi, dan f1-score yang didapatkan pada skenario ini namun tidak dapat mengalahkan model pada skenario 1 apabila dilihat dari hasil yang didapatkan dari precision dan recall yang sedikit lebih rendah dibandingkan dengan TF-IDF. Terutama pada kelas non-bullying, recall lebih rendah (77%), yang berarti beberapa kasus non-bullying tidak teridentifikasi dengan baik. Selain itu, nilai akurasi juga lebih rendah (77%) daripada akurasi yang didapatkan di skenario 1 (82%).

Word2Vec Skip-gram dengan SVM Kernel Linear:

Model pada skenario 3 terbilang cukup efektif dalam melakukan analisis sentimen dilihat dari preisis, akurasi, f1-score yang didapatkan dan lebih baik daripada model pada skenario 2 namun tidak lebih baik pada skenario 1.

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang dilakukan, dapat disimpulkan bahwa metode TF-IDF dengan SVM kernel linear memberikan performa terbaik dalam analisis sentimen cyberbullying, dengan akurasi, precision, recall, dan F1-score yang lebih tinggi dibandingkan dengan model berbasis Word2Vec (CBOW dan Skip-gram). Meskipun hipotesis awal menyatakan bahwa Word2Vec seharusnya bekerja lebih baik karena mampu menangkap hubungan semantik antar kata, hasil menunjukkan bahwa TF-IDF masih sangat efektif untuk tugas ini. Saran untuk penelitian selanjutnya adalah untuk mengeksplorasi teknik *preprocessing data* lebih lanjut seperti percobaan Word2Vec tanpa dilakukan stemming atau stopwords removal mengingat proses pembobotan Word2Vec sangat bergantung pada kata-kata yang ada disekitarnya. Selain itu, *tuning* parameter lebih lanjut dan penggunaan arsitektur neural network yang lebih kompleks dapat dipertimbangkan untuk meningkatkan performa model berbasis Word2Vec. Penelitian juga dapat diperluas dengan menggunakan dataset yang lebih besar dan beragam untuk melihat apakah hasil yang sama berlaku secara umum.

DAFTAR REFERENSI

Ahmad Aliero, A., Dankolo, N., Sulaimon Adebayo, B., Olanrewaju Aliyu, H., Gogo Tafida, A., Umar Kangiwa, B., & Muhammad Dankolo, N. (2023). Systematic review on text normalization techniques and its approach to non-standard words. *International Journal of Computer Applications*, 185(33). https://www.researchgate.net/publication/374166354

Al-Otaibi, S., & Al-Rasheed, A. (2022). A review and comparative analysis of sentiment analysis techniques. *Informatica (Slovenia)*, 46(6), 33–44. https://doi.org/10.31449/inf.v46i6.3991

Aura Azzahra, T., Anisa Sri Winarsih, N., Wilujeng Saraswati, G., Ocky Saputra, F., Syaifur Rohman, M., Oka Ratmana, D., Anggi Pramunendar, R., & Fajar Shidik, G. (2024). Perbandingan efektivitas Naïve Bayes dan SVM dalam menganalisis sentimen kebencanaan di YouTube. *Jurnal Media Informatika Budidarma*. https://doi.org/10.30865/mib.v8i1.7186

Damayanti, L., & Lhaksmana, K. M. (2024). Sentiment analysis of the 2024 Indonesia presidential election on Twitter. *Jurnal Dan Penelitian Teknik Informatika*, 8(2). https://doi.org/10.33395/v8i2.13379

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. *Komnas HAM*. https://www.komnasham.go.id/index.php/

Jivani, A. G., Anjali, M., & Jivani, G. (n.d.). A comparative study of stemming algorithms. https://www.researchgate.net/publication/284038938

Kemp, S. (2023, January 26). Digital 2023: Global overview report. *DataReportal*. https://datareportal.com/reports/digital-2023-global-overview-report

Jasmarizal, Rahmaddeni, Junadhi, & Khairul Anam, M. (n.d.). Penerapan metode support vector machine untuk analisis sentimen terhadap. *Indonesian Journal of Computer Science Attribution*, 13(1), 2024–1438.

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, 9(1). https://doi.org/10.1186/s13673-019-0192-7

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. E. (2014). Data preprocessing for supervised learning. https://www.researchgate.net/publication/228084519

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. http://arxiv.org/abs/1301.3781

Muraina, I. O. (n.d.). Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts. https://www.researchgate.net/publication/358284895

Nyoman, N. I., Suciartini, A., Luh, N. I., & Sumartini, P. U. (2018). 2 pertama diterima: 4 Agustus.

Patchin, J. W., & Hinduja, S. (2006). Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2), 148–169. https://doi.org/10.1177/1541204006286288

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. https://www.researchgate.net/publication/220282831

Rezki, N., Thamrin, S. A., & Siswanto, S. (2023). Sentiment analysis of Merdeka Belajar Kampus Merdeka policy using support vector machine with Word2Vec. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 17(1), 0481–0486. https://doi.org/10.30598/barekengvol17iss1pp0481-0486

Salloum, S. A., Khan, R., & Shaalan, K. (2020). A survey of semantic analysis approaches. *Advances in Intelligent Systems and Computing*, 1153, 61–70. https://doi.org/10.1007/978-3-030-44289-7_6

Tong, S., & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*.

Toraman, C., Yilmaz, E. H., Şahinuç, F., & Ozcelik, O. (2022). Impact of tokenization on language models: An analysis for Turkish. https://doi.org/10.1145/3578707