Available online at https://ejournal.almaata.ac.id/index.php/IJUBI
Indonesian Journal of Business Intelligence
Volume 7 | Issue 2 | December (2024)
ISSN 2621-3915 (PRINT), ISSN 2621-3923 (ONLINE), Published by **Alma Ata University Press**
IJUBI
Indonesian Journal ---- of ---- Business Intelligence

DETEKSI KEMIRIPAN DOKUMEN MENGGUNAKAN COSINE SIMILARITY BERDASARKAN REPRESENTASI TEKS COUNT VECTORIZER DAN TF IDF

Musthofa Galih Pradana[1*], Nindy Irzavika[2], Nurhuda Maulana[3]
[1]Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta,
[2]Sains Data, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta,
[3]Informatika, Fakultas Ilmu Komputer, Universitas Pembangunan Nasional Veteran Jakarta.
*musthofagalihpradana@upnvj.ac.id
Jakarta, Indonesia

Article history: *Received: 2 Januari 2025; Revised: 2 Januari 2025; Accepted: 14 Januari 2025*

Abstract

The purpose of the thesis course or final project is to foster a culture of critical thinking so that students are able and show the ability to solve problems with logical constructions of research. However, from the many benefits, there are several problems that also arise because of this course. Plagiarism is a common problem. Taking someone else's work, including their own opinions, and making it look like their own is plagiarism. The first step in the use of technology is to detect document similarities early on. In this case, the documents that must be collected by students during the process of submitting their thesis title are abstract. When used, the cosine similarity algorithm is a computationally efficient algorithm because it is very easy to understand and can be used with large-scale data. This research was conducted with two approaches to text representation, namely by using TF-IDF and Count Vectorizer. The corpus data used in this study was 1600 data of student thesis abstract documents, with the test using 30 data to see the performance of the cosine similarity algorithm in detecting the similarity of abstract documents. The results show that the TF-IDF text representation approach gets the same at 7.72861, and the Count Vectorizer gets a result at 16.85541 or has a gap of 9.1268 with the advantage of Count Vectorizer. This is because the Count Vectorizer calculates the frequency of words without considering whether they are common or rare, so common words still contribute fully to similarity.

Keywords: *Cosine Similarity, Count Vectorizer, TF IDF, Abstract, Text.*

Abstrak

Tujuan mata kuliah skripsi atau tugas akhir menumbuhkan budaya berpikir kritis, dan menunjukan kemampuan untuk memecahkan permasalahan dengan konstruksi logis dari penelitian. Akan tetapi, dari banyaknya manfaat tersebut, ada beberapa permasalahan yang juga muncul dikarenakan mata kuliah ini. Plagiarisme adalah masalah umum. Mengambil karya orang lain, termasuk pendapat mereka sendiri, dan membuatnya seperti karya sendiri adalah plagiarisme. Langkah pertama dalam penggunaan teknologi adalah mendeteksi kesamaan dokumen sejak dini. Dalam hal ini, dokumen yang harus dikumpulkan oleh mahasiswa selama proses pengajuan judul skripsi mereka adalah abstrak. Ketika digunakan, algoritma cosine similarity adalah algoritma yang efisien secara komputasi karena sangat mudah dipahami dan dapat digunakan dengan data berskala besar. Penelitian ini dilakukan dengan dua pendekatan representasi teks yaitu dengan menggunakan TF-IDF dan Count Vectorizer. Data korpus yang digunakan dalam penelitian ini adalah 1600 data dokumen abstrak skripsi mahasiswa, dengan pengujian menggunakan 30 data untuk melihat kinerja algoritma cosine similarity dalam mendeteksi kesamaan dokumen abstrak. Hasil penelitian menunjukkan bahwa pendekatan representasi teks TF-IDF mendapatkan kesamaan di angka 7,72861 dan Count Vectorizer mendapatkan hasil di angka 16,85541 atau punya gap sebesar 9,1268 dengan keunggulan Count Vectorizer. Hal ini disebabkan Count Vectorizer menghitung frekuensi kata tanpa

mempertimbangkan apakah kata tersebut umum atau jarang, sehingga kata-kata umum tetap berkontribusi penuh terhadap similarity.

Kata Kunci: *Cosine Similarity, Count Vectorizer, TF IDF, Abstrak, Teks.*

Pendahuluan

Proses pendidikan, penelitian, dan pengabdian, yang dikenal sebagai Tridharma Perguruan Tinggi adalah kewajiban [1] . Penelitian adalah bagian penting dari pendidikan tinggi karena dapat mendorong inovasi, pengembangan pengetahuan, dan pemecahan masalah kompleks dalam berbagai disiplin ilmu [2] . Keterlibatan sivitas akademika dalam penelitian tidak hanya tergantung pada dosen yang membantu menjalankan Tridharma Perguruan Tinggi, tetapi juga memerlukan mahasiswa untuk berpartisipasi dalam kegiatan penelitian. Menyelesaikan tugas akhir atau skripsi adalah cara untuk mendorong mahasiswa untuk terlibat aktif dalam kegiatan penelitian yang sudah menjadi tanggung jawab universitas. Saat ini percepatan dan kemajuan teknologi berbasis *artificial intelligence* membuat kemudahan dalam mengakses data dan informasi, namun kemudahan itu juga memiliki sisi lain yaitu masalah plagiarism. Plagiarisme adalah tindakan pengambilan karya bisa berupa pendapat dan sebagainya orang lain dan menjadikannya seolah-olah miliknya sendiri [3] . Tindakan plagiarisme harus dicegah, dan salah satu pendekatan yang banyak digunakan untuk memenuhi pencegahan ini adalah dengan menerapkan deteksi terhadap dokumen berbasis teks [4] . Penerapan metode untuk deteksi kemiripan salah satunya adalah dengan menggunakan *Cosine Similarity*.

Cosine Similarity merupakan teknik berbasis geometri yang bertujuan untuk menghitung tingkat kesamaan antara dua vektor teks dengan mengukur sudut kosinus di antara keduanya [5]. Teknik ini telah terbukti andal dalam berbagai aplikasi, termasuk dalam pengelompokan dokumen, sistem rekomendasi, dan deteksi plagiarisme. Pada dasarnya, untuk mengaplikasikan *cosine similarity*, dokumen teks perlu direpresentasikan dalam bentuk vektor numerik. Dua pendekatan yang sering digunakan untuk proses ini adalah *Count Vectorizer* dan *Term Frequency-Inverse Document Frequency* (TF-IDF). *Count Vectorizer* adalah metode dasar yang mengubah teks menjadi vektor berdasarkan frekuensi kemunculan setiap kata dalam dokumen. Meskipun sederhana, metode ini tidak mempertimbangkan pentingnya suatu kata dalam konteks keseluruhan korpus [6] . Sebaliknya, TF-IDF menawarkan pendekatan yang lebih canggih dengan memberikan bobot lebih pada kata-kata yang jarang muncul tetapi memiliki signifikansi tinggi dalam dokumen tertentu [7] . Dengan demikian, TF-IDF dapat menyaring pengaruh kata-kata umum yang sering muncul di banyak dokumen, seperti kata sambung atau kata fungsi. Studi ini bertujuan untuk membandingkan performa kedua pendekatan tersebut dalam mendeteksi kemiripan antar dokumen teks menggunakan *Cosine Similarity*. Pendekatan berbasis *Count Vectorizer* sering kali memberikan hasil yang lebih tinggi dalam hal deteksi kemiripan dikarenakan karakter dari *Count Vectorizer* dikarenakan mempertimbangkan semua kata secara setara tanpa membedakan relevansinya [8] . Sebaliknya, TF-IDF cenderung menghasilkan nilai kesamaan yang lebih rendah, tetapi lebih akurat dalam menggambarkan konteks semantik antar dokumen [9]. Melalui analisis yang mendalam, studi ini mengeksplorasi kelebihan dan kekurangan kedua metode ini, khususnya dalam korpus teks yang memiliki tingkat keragaman topik dan ukuran yang berbeda.

Penerapan metode *Cosine Similarity* dengan dua pendekatan ini memiliki relevansi yang luas, khususnya dalam dunia akademik, di mana sering kali diperlukan untuk mengevaluasi kemiripan abstrak skripsi, makalah penelitian, atau tugas akhir. Selain itu, pendekatan ini dapat diadaptasi untuk berbagai keperluan praktis, seperti penyaringan dokumen dalam sistem pencarian informasi

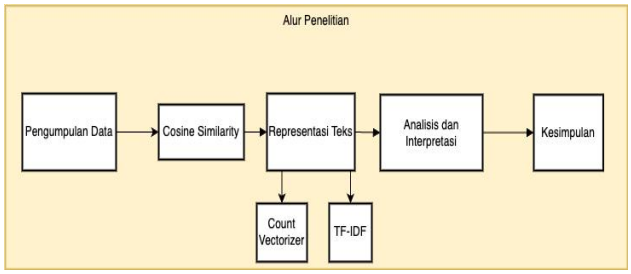
atau mendeteksi plagiarisme pada karya tulis. Dengan semakin berkembangnya teknologi, evaluasi terhadap performa metode berbasis *Count Vectorizer* dan TF-IDF menjadi semakin penting untuk memahami bagaimana metode-metode ini dapat dioptimalkan dalam konteks nyata.

Hasil penelitian ini tidak hanya bermanfaat dalam konteks akademik tetapi juga memiliki aplikasi praktis yang luas dalam berbagai bidang seperti ilmu perpustakaan, e-commerce, dan teknologi informasi. Dengan demikian, studi ini menjadi landasan penting bagi pengembangan lebih lanjut metode analisis teks berbasis *cosine similarity* yang mampu memenuhi kebutuhan pengolahan data teks yang semakin kompleks.

Penelitian relevan mengenai penerapan *Count Vectorizer* dan TF-IDF dalam proses teks mining diantaranya dalam melakukan sentimen untuk review film dengan hasil dengan penerapan hyperparameter dalam eksperimennya [10]. Dalam klasifikasi teks lain TF-IDF dan *Count Vectorizer* juga memiliki hasil yang bersaing [11] dan dalam klasifikasi sentimen [12] . TF-IDF memiliki keunggulan dalam beberapa penelitian seperti pada klasifikasi hate speech [13] dan klasifikasi emoticon [14] atau klasifikasi sentiment [15] . Keunggulan juga terkadang terjadi pada *Count Vectorizer* seperti pada Category Detection [16] dan juga Reddit flair detection [17] atau kombinasi kedua representasi teks ini dengan algoritma klasifikasi *Support Vector* Machine [18].

Metode

Pendekatan metode atau alur penelitian ditunjukan pada Gambar 1.



Gambar 1. Alur Penelitian

Alur pada metode di atas akan dijabarkan menjadi lebih detail pada bagian pembahasan hasil.

Pembahasan dan Hasil

Pengumpulan Data

Data yang digunakan adalah data dari dokumen abstrak mahasiswa pada salah satu perguruan tinggi. Data yang digunakan adalah data abstrak yang akan digunakan sebagai data training dan testing pada 2 tahun terakhir. Data ini diambil dari 4 program studi dalam 1 fakultas di sebuah perguruan tinggi. Adapun gambaran dari data corpus yang digunakan ditunjukan pada Tabel 1.

Tabel 1. Raw Data

Corpus
penelitian ini dilakukan untuk merancang sistem untuk mengrasipkan dokumen-dokumen pada pt federal international finance atau yang lebih dikenal pt. fif. dengan banyaknya dokumen setiap harinya sekitar 108.000 dokumen yang harus disimpan, tentulah menjadi sukar dalam pencarian dokumen, mudah terjadi kerusakan fisik dan cakupan dokumen. topik pembahasan penulisan skripsi ini adalah pembuatan alat sistem manajemen dokumen, sedangkan batas masalahnya adalah pengarsapan dokumen kredit. metode penelitian yang

dilakukan meliputi pengumpulan data yaitu dengan teknik wawancara langsung kepada staf yang berwenang. dengan mengarsipkan dokumen secara digital juga dapat diatur siapa saja yang berhak melihat atau mendownload dokumen tersebut. sehingga menghindari orang yang tidak bertanggung jawab atas dokumen.

Tokenization

Tahap ini akan menguraikan deskripsi yang semula berupa kalimat menjadi kata. Proses ini berguna untuk membagi kata ke dalam token untuk mempermudah proses identifikasi kata, detailnya seperti ditunjukan pada Tabel 2.

Tabel 2. Tokenization

Corpus
"penelitian", "ini", "dilakukan", "untuk", "merancang", "sistem", "untuk", "mengarsipkan", "dokumen", "dokumen", "pada", "pt", "federal", "international", "finance", "atau", "yang", "lebih", "dikenal", "pt", "fif", "dengan", "banyaknya", "dokumen", "setiap", "harinya", "sekitar", "108.000", "dokumen", "yang", "harus", "disimpan", "tentulah", "menjadi", "sukar", "dalam", "pencarian", "dokumen", "mudah", "terjadi", "kerusakan", "fisik", "dan", "cakupan", "dokumen", "topik", "pembahasan", "penulisan", "skripsi", "ini", "adalah", "pembuatan", "alat", "sistem", "manajemen", "dokumen", "sedangkan", "batas", "masalahnya", "adalah", "pengarsapan", "dokumen", "kredit", "metode", "penelitian" "yang", "dilakukan", "meliputi", "pengumpulan", "data", "yaitu", "dengan", "teknik", "wawancara", "langsung", "kepada", "staf", "yang", "berwenang", "dengan", "mengarsipkan", "dokumen", "secara", "digital", "juga", "dapat", "diatur", "siapa", "saja", "yang", "berhak", "melihat", "atau", "mendownload", "dokumen", "tersebut", "sehingga", "menghindari", "orang", "yang", "tidak", "bertanggung", "jawab", "atas", "dokumen"

Filtering

Tahapan ini melakukan filtering untuk menghapus kata-kata yang tidak relevan atau stop words, hal ini akan berpengaruh terhadap hasil keseluruhan hasil analisis karena fungsinya untuk meminimalisir penggunaan kata yang kurang berimpact, gambaran kata yang dihilangkan seperti pada Tabel 3

Tabel 3. Filtering

Corpus
untuk, atau, yang, lebih, dengan, menjadi, masalahnya, tentulah, dalam, terjadi, ini, sedangkan, yaitu, dilakuman, secara, juga, dapat, saja, atas, siapa, adalah, pembuatan.

Stemming

Tahap stemming adalah mengubah kata-kata menjadi bentuk dasarnya. Tujuan utamanya adalah mereduksi jumlah variasi dalam representasi dari sebuah kata. Hasil stem ditunjukan pada Tabel 4.

Tabel 4. Stemming

Corpus
penelitian merancang sistem mengrasipkan dokumen-dokumen pt federal international finance dikenal pt fif banyaknya dokumen harinya 108.000 dokumen disimpan sukar pencarian dokumen mudah kerusakan fisik cakupan dokumen topik pembahasan penulisan skripsi pembuatan alat sistem manajemen dokumen batas pengarsapan dokumen kredit metode penelitian meliputi pengumpulandata teknik wawancara langsung staf berwenang mengarsipkan dokumen digital diatur berhak mendownload dokumen menghindari orang bertanggung dokumen

Cosine Similarity

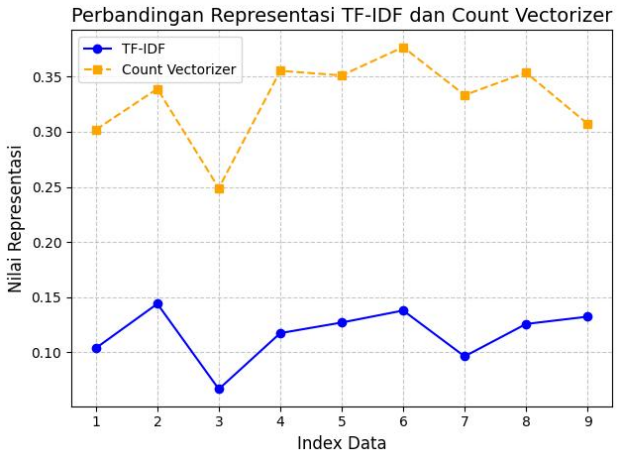
Model yang digunakan adalah menggunakan cosine similarity dengan implementasi pendekatan representasi teks yang berbeda yakni menggunakan TF IDF dan menggunakan Count Vectorizer. Hasil kedua representasi teks ini akan dibandingkan satu sama lain.

Representasi Teks

Representasi teks yang digunakan adalah dua model, yakni menggunakan TF-IDF dan menggunakan Count Vectorizer. Hasil kedua model ini dibandingkan dan di analisis hasil dari keduanya. Hasil dari representasi teks diuji dalam 3 skenario dengan masing-masing scenario menguji 10 data corpus.

Skenario 1

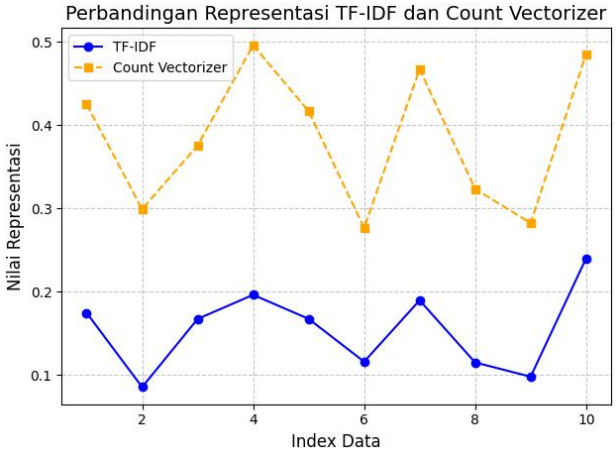
Hasil pada skenario 1 dengan 10 data menunjukan representasi Count Vectorizer lebih unggul daripada TF-IDF. Hasil grafik pada skenario 1 ditunjukan pada Gambar 2.



Gambar 2. Hasil Uji Skenario 1

Skenario 2

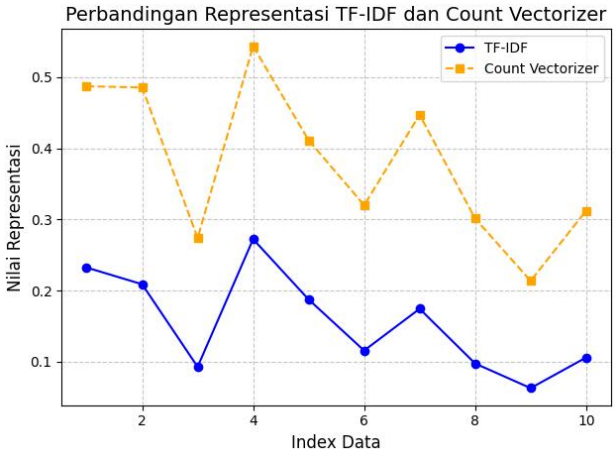
Hasil pada skenario 2 dengan 10 data juga menunjukan representasi Count Vectorizer lebih unggul daripada TF-IDF. Hasil grafik pada skenario 2 ditunjukan pada Gambar 3.



Gambar 3. Hasil Uji Skenario 2

Skenario 3

Hasil pada skenario 3 dengan 10 data juga masih menunjukan representasi Count Vectorizer lebih unggul daripada TF-IDF. Hasil grafik pada skenario 3 ditunjukan pada Gambar 4.



Gambar 4. Hasil Uji Skenario 3

Analisis dan Interpretasi

Hasil perbandingan kedua model dari tiga kali scenario percobaan jenis representasi teks TF-IDF dan menggunakan *Count Vectorizer* menunjukan hasil yang memiliki perbedaan. Deteksi kemiripan menggunakan Cosine Similarity dan pendekatan representasi teks TF-IDF mendapatkan kesamaan rata-rata di angka 7,72861. Deteksi kemiripan menggunakan Cosine Similarity dan pendekatan representasi teks Count Vectorizer mendapatkan hasil rata-rata di angka 16,85541. Gap deteksi kemiripan dokumen menggunakan Cosine Similarity antara TF-IDF dan Count Vectorizer memiliki nilai sebesar 9,1268. Keunggulan Count Vectorizer dibandingkan TF IDF disebabkan Count Vectorizer menghitung frekuensi kata tanpa mempertimbangkan apakah kata tersebut umum atau jarang, sehingga kata-kata umum tetap berkontribusi penuh terhadap similarity.

Kesimpulan dan Saran

Berdasarkan pembahasan sebelumnya, dapat disimpulkan sebagai berikut:

1. Deteksi kemiripan menggunakan *Cosine Similarity* dan pendekatan representasi teks TF-IDF mendapatkan kesamaan di angka 7,72861.
2. Deteksi kemiripan menggunakan *Cosine Similarity* dan pendekatan representasi teks *Count Vectorizer* mendapatkan hasil di angka 16,85541.
3. Gap deteksi kemiripan dokumen menggunakan *Cosine Similarity* antara TF-IDF dan *Count Vectorizer* memiliki nilai sebesar 9,1268.
4. Keunggulan *Count Vectorizer* dibandingkan TF IDF disebabkan *Count Vectorizer* menghitung frekuensi kata tanpa mempertimbangkan apakah kata tersebut umum atau jarang, sehingga kata-kata umum tetap berkontribusi penuh terhadap similarity.

Saran untuk penelitian lanjutan bisa dikembangan dengan model transformer atau dengan pre-trained model untuk membandingkan hasil akurasi dari model.

Referensi

[1] Pemerintah Indonesia, "Undang-Undang Nomor 4 Tahun 2014 Tentang Penyelenggaraan Pendidikan Tinggi dan Pengelolaan Perguruan Tinggi," *Standar Nasional Pendidikan*, p. 37, 2014, [Online]. Available: https://peraturan.bpk.go.id/Home/Details/5441/pp-no-4-tahun-2014

[2] Kementerian Pendidikan dan Kebudayaan, *Permendikbud Nomor 3 Tahun 2020*. www.kemdikbud.go.id, 2020.

[3] A. Kleebayoon and V. Wiwanitkit, "Artificial Intelligence, Chatbots, Plagiarism and Basic Honesty: Comment," *Cell Mol Bioeng*, vol. 16, no. 2, pp. 173–174, Apr. 2023, doi: 10.1007/s12195-023-00759-x.

[4] V. Chandere, S. Satish, and R. Lakshminarayanan, "Online plagiarism detection tools in the digital age: A review," *Ann Rom Soc Cell Biol*, vol. 25, no. 1, pp. 7110–7119, 2021, [Online]. Available: https://annalsofrscb.ro/index.php/journal/article/view/881

[5] K. W. G. A. P. P. H. S. D. P. W. D. H. R. S. K. N. M. A. P. P. Musthofa Galih Pradana, *Information Retrieval*. Penamuda, 2024.

[6] A. Kulkarni and A. Shivananda, *Natural Language Processing Recipes*. 2021. doi: 10.1007/978-1-4842-7351-7.

[7] Raymond S. T. Lee, *Natural Language Processing: A Textbook with Python Implementation*. Springer, 2023.

[8] Thushan Ganegedara, *Natural Language Processing with TensorFlow - Second Edition*. Packt Publishing, 2022.

[9] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information (Switzerland)*, vol. 11, no. 9, pp. 1–17, 2020, doi: 10.3390/info11090421.

[10] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer," *Soc Netw Anal Min*, vol. 14, no. 1, p. 87, Apr. 2024, doi: 10.1007/s13278-024-01250-9.

[11] A. Wendland, M. Zenere, and J. Niemann, "Introduction to Text Classification: Impact of Stemming and Comparing TF-IDF and Count Vectorization as Feature Extraction Technique," 2021, pp. 289–300. doi: 10.1007/978-3-030-85521-5_19.

[12] G. M. Raza, Z. S. Butt, S. Latif, and A. Wahid, "Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models," in *2021 International Conference on Digital Futures*

and Transformative Technologies (ICoDT2), IEEE, May 2021, pp. 1–6. doi: 10.1109/ICoDT252288.2021.9441508.

[13] K. M. Suryaningrum, "Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 5, no. 2, pp. 79–83, May 2023, doi: 10.21512/emacsjournal.v5i2.9978.

[14] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. Gulzar Hussain, "Bangla Text Emotion Classification using LR, MNB and MLP with TF-IDF & CountVectorizer," in *2022 26th International Computer Science and Engineering Conference (ICSEC)*, IEEE, Dec. 2022, pp. 275–280. doi: 10.1109/ICSEC56337.2022.10049341.

[15] H. D. Abubakar and M. Umar, "Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec," *SLU Journal of Science and Technology*, vol. 4, no. 1 & 2, pp. 27–33, Aug. 2022, doi: 10.56471/slujst.v4i.266.

[16] A. Gupta and U. Sharma, "Machine Learning Based Aspect Category Detection for Hindi Data Using TF-IDF and Count Vectorization," in *2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, IEEE, Mar. 2024, pp. 39–44. doi: 10.1109/DICCT61038.2024.10532960.

[17] M. Singhal, N. Singhal, S. Khera, A. Upmanyu, and P. Nagrath, "Improvisation of Reddit flair detection using TF-IDF and countvectorizer," 2023, p. 020003. doi: 10.1063/5.0181369.

[18] Sajid Khan, Mehmoon Anwar, Huma Qayyum, Farooq Ali, and Marriam Nawaz, "Fake News Classification using Machine Learning: Count Vectorizer and Support Vector Machine," *Journal of Computing & Biomedical Informatics*, vol. 4, no. 01, Jan. 2023, doi: 10.56979/401/2022/85.