Jurnal Indonesia : Manajemen Informatika dan Komunikasi
www.journal.amikindonesia.ac.id/jimik/ Vol 4 No 2, Mei (2023)
E-ISSN: 2723-7079, P-ISSN: 2776-8074

ANALISIS SENTIMEN ARTIKEL BERITA PEMILU BERBASIS METODE KLASIFIKASI

Fathir [1*], M. Amin Hariyadi [2], Yunifa Miftachul A [3]

[1*,2,3] Program Studi Magister Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Kota Malang, Provinsi Jawa Timur, Indonesia.

Email: fathirpuncak@gmail.com [1*], adyt2002@uin-malang.ac.id [2], yunif4@ti.uin-malang.ac.id [3]

Histori Artikel:
Dikirim 2 Maret 2023; *Diterima dalam bentuk revisi* 18 Maret 2023; *Diterima* 11 April 2023; *Diterbitkan* 10 Mei 2023.

Abstrak

Penyaluran informasi berupa berita online begitu masif di tengah masyarakat luas, sehingga sulit membedakan berupa berita haox ataupun berita positif. Sehingga dibutuhkan klasifikasi mengenai sentimen publik tentang pelaksanaan pemilu dengan menggunakan data artikel berita media mainstream yang menggunakan data uji 1064 dataset. Metode yang digunakan adalah pada penelitian ini adalah algoritma naive bayes, algoritma random forest, dan algoritma support vektor machine. Model uji coba menggunakan smote dimana hasil performa yang dilakukan oleh algortima yang digunakan dengan menggunakan smote dan tidak menggunakan smote, dimana random forest menghasilkan akurasi 91,88%, sedangkan tampa menggunakan smote support vektor machine menghasilkan akurasi 92,05%.

Kata Kunci: Analisis Sentimen; Naive Bayes; Support Vector Machine; Random Forest.

Abstract

The distribution of information in the form of online news is so massive in the wider community, that it is difficult to distinguish between haox news and positive news. So that a classification is needed regarding public sentiment about the implementation of elections using mainstream media news article data using 1064 dataset test data. The methods used in this study are the naive Bayes algorithm, the random forest algorithm, and the support vector machine algorithm. The test model uses smote where the performance results are carried out by the algorithm used using smote and not using smote, where random forest produces an accuracy of 91.88%, while without using a smote support vector machine it produces an accuracy of 92.05%.

Keyword: Sentiment Analysis; Naive Bayes; Support Vector Machine; Random Forest.

1. Pendahuluan

Dengan adanya perkembangan teknologi informasi dan komunikasi dewasa ini yang begitu pesat menjadikan berita online begerak begitu cepat dari di kalangan masyarakat luas [1]. Terutama pemberitaan tentang pelaksanaan pemilu 2024 yang akan datang, sehingga perlu ada pendidikan politik di tengah kalangan masyarakat untuk menghindari penyampaian berita online yang bersifat hoax mengakibatkan perpecahan sehingga dari beberapa topik pembahasan tersebut bisa diklasifikasi oleh algoritma machine learning berupa sentimen yang berbau positif dan negatif [2]. Konsumsi pemberitaan online oleh publik yang begitu pesat dan tidak terkontrol menjadikan asumsi publik tidak terkendali, baik pemberitaan yang bersifat positif maupun pemberitaan yang bersifat negatif, baik berupa fakta maupun berita hoax, sehingga mempengaruhi pemahaman masyarakat tentang politik terutama tentang pelaksanaan pemilu 2024, sehingga media penyampai informasi perlu menyampaikan pemberitaan yang independen dan juga di sampaikan dengan baik tampa ada unsur-unsur yang merugikan masyarakat [3].

Di beberapa negara demokrasi menjadikan pemberitaan sangat diminati selain sebagai penyampaian informasi untuk menilai kebijakan politik ataupun kebijakan ekonomi sebuah pemerintahan kemudian untuk disampaikan oleh masyarakat luas, karena menyampaikan pemberitaan di negara demokrasi adalah bagian dari kebebasan berpendapat oleh masyarakat dan juga media pers sesuai undang-undang kebebasan pers pasal 28F dalam hal mencari, memperoleh, memiliki, menyimpan, mengolah, dan menyampaikan informasi, meskipun ada beberapa media yang melakukan hal sebaliknya untuk kepentingan beberapa lembaga tertentu [4][5][6].

Di beberapa penelitian sebelumnya mengenai *text mining* dan juga analisis sentimen tentang beberapa perusahaan besar seperti perusahaan penerbangan di amerika serikat, mengambil data dari komentar pengguna layanan penerbangan yang ada di amerika melalui media twitter, dan kemudian melakukan proses-proses yang umumnya dalam text mining yakni membersihkan tweet dan kemudian merepresentasikan tweet ini sebagai vektor menggunakan konsep pembelajaran. Metode klasifikasi yang digunakan dalam sentimen tersebut adalah random forest, support vektor machine, gaussian naive bayes. Pengklasifikasi dilatih menggunakan 80% data dan diuji menggunakan 20% data sisanya, hasil dari set penelitian adalah sentimen tweet berupa positif, negatif, dan juga netral. Berdasarkan hasil yang diperoleh, akurasi dihitung 80,01% untuk random forest [7][8].

Penelitian tentang sentimen pemilu juga pernah dilakukan oleh Sudiantoro di jawa barat, pada pelaksanaan pemilu gubernur jawa barat tersebut menilai sentimen publik pada data twitter sebanyak 300 dataset yang digunakan dengan menggunakan metode algoritma naive bayes, hasil pada penelitian tersebut mendapatkan akurasi sebesar 84%, dari 100 dataset yang digunakan sebagai data uji, 32 dataset 32 data bersentimen positif dan 68 dataset lainnya bersentimen negatif [9]. Penelitian yang melibatkan *text mining* dalam analisis sentimen pemilu di india juga dilakukan oleh Sharma menggunakan data twitter dalam bahasa hindi, penambangan data teks berjumlah 42.235 tweet yang dikumpulkan selama satu bulan yang merujuk pada lima partai politik nasional di india, metode yang digunakan *dictionary based, naive bayes,* dan *support vektor machine* untuk mengklasifikasi data uji sebagai positif, negatif dan netral, sehingga hasilnya support vektor machine memperkirakan peluang 78,4% [10],[7], [8], [11].

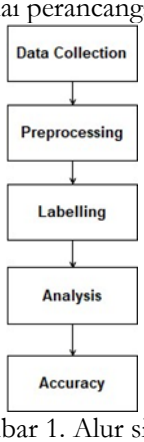
Beberapa penelitian yang mendukung dalam penelitian terutama tentang analisis sentimen adalah penggunaan algoritma *smote* dalam menyelesaikan masalah data-data yang tidak seimbang dalam jumlah kuantitas atau secara jumlah, peran smote tentu dalam rangka meningkatkan kinerja belajar machine learning misalnya penelitian Sarakit et al, memperbaiki kinerja machine learning dalam klasifikasi, pengenalan pola, kategorisasi teks, dan mengubah data minoritas menjadi jumlah instan dari kelas minoritas, dan hasilnya cukup signifikan mulai dari menggunakan algoritma *support vektor machine, random forest,* dan *naive bayes*, dari 80% sampai dengan 93% peningkatan dan teknik *SMOTE* dapat mengatasi masalah ketidakseimbangan data dan mendapatkan hasil klasifikasi yang lebih baik [12],[13].

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, Mei (2023), pp. 485-493
DOI : https://doi.org/10.35870/jimik.v4i2.220

Dari penjelasan tersebut kenapa perlu dilakukan penelitian ini adalah untuk mengklasifikasi beberapa artikel pemberitaan yang membahas tentang pemilu 2024 berbasis metode klasifikasi dengan menggunakan algoritma *meachine learning* seperti *support vektor meachine* [14], *random forest* dan *naive bayes* [15] kemudian menilai sentimen positif, negatif, dan netral dengan bantuan algoritma *smote* dalam menangani data-data yang tidak seimbang sehingga memberikan pendidikan politik kepada masyarakat dalam menilai topik pemberitaan media online.

2. Metode Penelitian

Pada bagian ini akan dijelaskan mengenai perancangan sistem dalam penelitian ini.

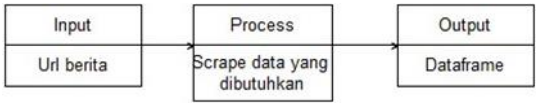


Gambar 1. Alur sistem

1) *Data Collection*

Pengumpulan data merupakan tahapan yang paling penting dalam melakukan sebuah penelitian. Pada penelitian ini pengumpulan data dilakukan dengan cara web scraping yakni mengambil data artikel berita, data artikel berita yang di scraping berjumlah 5900 data, dengan periode pengambilan mulai dari agustus 2020 sampai dengan november 2022, kemudian selanjutnya melakukan langkah-langkah preprocessing dan menganalisis dengan algoritma klasifikasi.

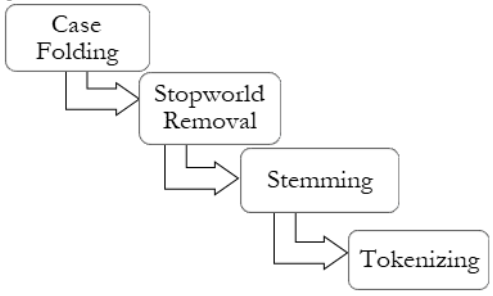
Metode scraping yang dilakukan menggunakan python dengan google colab, proses pengambilan data sebagai berikut:



Gambar 2. Tahapan scraping data

2) *Preprocessing*

Pada tahap ini ada beberapa proses yang dilakukan yakni teknik preprocessing atau pengolahan yang digunakan untuk mengolah data mentah menjadi mudah untuk dipahami [14][15]. yakni akan dijelaskan secara rinci sebagai berikut ini:



Gambar 3. Proses preprocessing

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, Mei (2023), pp. 485-493
DOI : https://doi.org/10.35870/jimik.v4i2.220

3) *Labelling*

Dilakukan labelisasi menggunakan algoritma *lexicon based*. Pada proses pelabelan didapatkan hasil bawah conten yang termasuk dalam class positif sebayak 5000 conten class negatif sebanyak 900 conten berita [16].

4) Teknik Analisa

Dari 3 algoritma yang digunakan menggunakan 2 eksperimen yakni dengan menggunakan algoritma *smote* dan tampa menggunakan algoritma *smote*. Teknik *oversampling minoritas sintetis* (*SMOTE*) adalah teknik statistik untuk meningkatkan jumlah kasus dalam himpunan data dengan cara yang seimbang. Komponen bekerja dengan menghasilkan instans baru dari kasus minoritas yang ada yang Anda berikan sebagai input. Implementasi smote ini tidak mengubah jumlah kasus mayoritas [12].

5) Evaluasi Performa

Evaluasi pengujian dilakukan sesuai tahap yang sudah dijelaskan pada bagian metode penelitian. Setelah dilakukan pre-processing, dan pelabelan selanjutnya dilakukan analisa perbandingan menggunakan algoritma smote dan tidak menggunakan algoritma smote, tentu dengan tiga algoritma naive bayes, algoritma random forest, dan algoritma support vektor machine, di mana alat ukur yang digunakan untuk perbandingan adalah nilai acuration, precision, recall dan f1-score.
Adapun rumus confusion matrix yang digunakan adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \quad (5)$$

TP (True Positive) jika keputusan yang dihasilkan oleh sistem dan GT menyatakan hal yang sama, TN (True Negatif) jika keputusan yang tidak dihasilkan oleh sistem dan GT menyatakan hal yang sama, FP (False Positif) jika keputusan yang tidak dihasilkan oleh sistem dan GT menyatakan hal yang berbeda, FN (False Negative) jika keputusan yang tidak dihasilkan oleh sistem adalah dan GT menyatakan hal yang berbeda.

3. Hasil dan Pembahasan

Pada bagian hasil penulis akan menjelaskan hasil pengujian terhadap model usalan yang dibuat. Pengujian dilakukan sesuai tahap yang sudah dijelaskan pada bagian metode penelitian. Setelah dilakukan pre-processing, dan pelabelan selanjutnya dilakukan analisa perbandingan menggunakan algoritma smote dan tidak menggunakan algoritma smote, tentu dengan tiga algoritma naive bayes, algoritma random forest, dan algoritma support vektor machine, di mana alat ukur yang digunakan untuk perbandingan adalah nilai akuration, precision, recall dan f1-score.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, Mei (2023), pp. 485-493
DOI : https://doi.org/10.35870/jimik.v4i2.220

3.1 Pembuatan Model *Naive Bayes*

Tahap ini data yang sudah melalui preprocessing, pelabelan, dan selanjutnya akan dilakukan uji coba dengan menerapkan algoritma *naive bayes*. Pada proses ini dataset akan dilakukan pembagian data training dan data testing, yakni pembagian data sebanyak (80:20) [18]–[20]. Hasil evaluasi *Without Smote* akan di tampilkan sebagai berikut ini:

```
MultinomialNB Accuracy: 0.8444632290786137
MultinomialNB Precision: 0.4410479944354278
MultinomialNB Recall: 0.5353004069807547
MultinomialNB f1_score: 0.4645871754858845
confusion matrix:
 [[ 81   2  26]
 [  4   0   6]
 [142   4 918]]
================================================
```

Gambar 3. Hasil akurasi menggunakan *confusion matriks*

Dari hasil proses klasifikasi tentang uji coba menggunakan algoritma naive bayes 89,94%, kemudian hasil precision adalah 29,98%, dan nilai recall 33,33%, nilai f1-score adalah 31,56%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) dan false positive (FP). Hasil evaluasi *With Smote* akan di tampilkan sebagai berikut ini:

```
MultinomialNB Accuracy: 0.8994082840236687
MultinomialNB Precision: 0.2998027613412229
MultinomialNB Recall: 0.3333333333333333
MultinomialNB f1_score: 0.31568016614745587
confusion matrix:
[[  47    1   61]
 [   1    1    8]
 [  55    5 1004]]
|================================================
```

Gambar 4. Hasil akurasi menggunakan *confusion matriks*

Dari hasil gambar tersebut maka nilai akurasi dari algoritma naive bayes 0.84 (84,44%), kemudian hasil precision adalah 44,11%, dan nilai recall 53,53%, nilai f1-score adalah 46,45%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) adalah 81 data, dan nilai sebaliknya adalah (false) 28 data, dan nilai negatif true 918 dan nilai negatif false yakni 146 data.

3.2 Pembuatan Model *Random Forest*

Tahap ini data sudah melalui preprocessing dan pelabelan, akan dilakukan uji coba dengan menerapkan algoritma *random forest* dan akan melakukan pembagian data training dan data testing [13], [21], [22]. Hasil Without Smote akan di tampilkan sebagai berikut ini:

```
Random Forest Accuracy: 0.907861369399831
Random Forest Precision: 0.5535253654342219
Random Forest Recall: 0.3776384769262606
Random Forest f1_score: 0.39453779026256336
confusion matrix:
 [[  15    0   94]
 [   0    0   10]
 [   5    0 1059]]
================================================
```

Gambar 5. Hasil akurasi menggunakan *confusion matriks*

Dari hasil gambar tersebut maka nilai akurasi dari algoritma random forest adalah 91,88%, kemudian hasil precision adalah 57,62%, dan nilai recall 42,01%, nilai f1-score adalah 45,24%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) adalah 29 data, dan nilai sebaliknya adalah false positive (FP) 80 data, dan nilai true negative (TN) 1058 dan nilai false negative (FN) yakni 6 data.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, Mei (2023), pp. 485-493
DOI : https://doi.org/10.35870/jimik.v4i2.220

Hasil evaluasi *With Smote* akan di tampilkan sebagai berikut ini:

```
Random Forest Accuracy: 0.9188503803888419
Random Forest Precision: 0.5762555749466745
Random Forest Recall: 0.4201386493757329
Random Forest f1_score: 0.45248868778280543
confusion matrix:
 [[  29    1   79]
 [   1    0    9]
 [   6    0 1058]]
================================================
```

Gambar 6. Hasil akurasi menggunakan *confusion matriks*

Dari hasil gambar tersebut maka nilai akurasi dari algoritma random forest adalah 90,78%, kemudian hasil precision adalah 55,35%, dan nilai recall 37,76%, nilai f1-score adalah 39,45%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) adalah 15 data, dan nilai sebaliknya adalah false positive (FP) 94 data, dan nilai true negative (TN) 1059 dan nilai false negative (FN) yakni 5 data.

3.3 Pembuatan Model *Support Vektor Machine*

Tahap ini data sudah melalui preprocessing dan pelabelan, akan dilakukan uji coba dengan menerapkan algoritma support vektor machine dan akan melakukan pembagian data treaning dan data testing, yakni pembagian data sebanyak (80:20) [2]. Pengolahan data pada penelitian ini menggunakan google colab. Hasil evaluasi *Without Smote* akan di tampilkan sebagai berikut ini:

```
Support Vector Machine Accuracy: 0.9205409974640744
Support Vector Machine Precision: 0.5626234815749557
Support Vector Machine Recall: 0.43723414039686376
Support Vector Machine f1_score: 0.469786460699682
confusion matrix:
 [[  35    0   74]
 [   1    0    9]
 [  10    0 1054]]
================================================
```

Gambar 7. Hasil akurasi menggunakan *confusion matriks*

Dari hasil gambar tersebut maka nilai akurasi dari algoritma support vektor machine adalah 92,05%, kemudian hasil precision adalah 56,26%, dan nilai recall 43,72%, nilai f1-score adalah 46,97%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) adalah 35 data, dan nilai sebaliknya adalah false positive (FP) 74 data, dan nilai true negative (TN) 1054 dan nilai false negative (FN) yakni 10 data. Hasil evaluasi *With Smote* akan di tampilkan sebagai berikut ini:

```
Support Vector Machine Accuracy: 0.9070160608622148
Support Vector Machine Precision: 0.49679789265086155
Support Vector Machine Recall: 0.5090765905589661
Support Vector Machine f1_score: 0.502703453062641
confusion matrix:
 [[  63    2   44]
 [   3    0    7]
 [  51    3 1010]]
================================================
```

Gambar 8. Hasil akurasi menggunakan *confusion matriks*

Dari hasil gambar tersebut maka nilai akurasi dari algoritma support vektor machine adalah 90,70%, kemudian hasil precision adalah 49,67%, dan nilai recall 50,90%, nilai f1-score adalah 50,27%. Pada penelitian ini juga digambarkan dari matrik 3x3 di atas, dimana nilai true positive (TP) adalah 63 data, dan nilai sebaliknya adalah false positive (FP) 46 data, dan nilai true negative (TN) 1010 dan nilai false negative (FN) yakni 54 data.

Jurnal Indonesia : Manajemen Informatika dan Komunikasi, Volume 4 No 2, Mei (2023), pp. 485-493
DOI : https://doi.org/10.35870/jimik.v4i2.220

3.4 Evaluasi

Berikut beberapa evaluasi menggunakan confusion matrix dari semua algoritma yang digunakan, dan menggunakan algoritma smote sebagai berikut ini:

Tabel 1. Evaluasi hasil perbandingan dengan smote

Algoritma	Performa			
	Accuracy	Precision	Recall	F1-score
Naïve Bayes	84,44%	44,10%	53,53%	46,45%
Random Forest	91,88%	57,62%	42,01%	45,24%
Support vektor machine	90,70%	49,67%	50,90%	50,27%.

Berikut beberapa evaluasi menggunakan confusion matrix dari semua algoritma yang digunakan, dan tampa menggunakan algoritma smote sebagai berikut ini:

Tabel 2. Evaluasi hasil perbandingan tampa smote

Algoritma	Performa			
	Accuracy	Precision	Recall	F1-score
Naïve Bayes	89,98%	29,98%	33,33%	31,56%
Random Forest	90,78%	55,35%	37,76%	39,45%
Support vektor machine	92,05%	56,26%	43,72%	46,97%

4. Kesimpulan

Berdasarkan hasil pengujian menggunakan algoritma naive bayes, random forest, dan support vektor machine dan evaluasi menggunakan confusion matriks dari tiga algoritma yang digunakan yakni dengan experimen pertama menggunakan algoritma smote mendapatkan hasil dari algoritma random forest mendapatkan akurasi tertinggi sebesar 91,88%, dan nilai dari presicion 57,62% dan nilai recall 42,01%, kemudian nilai f1-score 45,24%. Kemudian pada experimen kedua yakni tampa menggunakan algoritma smote algoritma yang menghasilkan akurasi tertinggi adalah support vektor machine dengan nilai akurasi sebesar 92,05%, dan nilai precision adalah 56,26%, recall sebesar 43,72%, dan nilai f1-score adalah 46,97%. Diharapkan penelitian selanjutnya memperluas dataset, dan menambahkna fitur ekstraksi data dengan word embeddings atau neural networks, dan juga melihat sentimen publik di bidang yang lain seperti kesehatan, pendidikan dan lingkungan hidup.

5. Daftar Pustaka

[1] Huda, I.A., 2020. Perkembangan teknologi informasi dan komunikasi (TIK) terhadap kualitas pembelajaran di sekolah dasar. *Jurnal Pendidikan Dan Konseling (JPDK)*, *2*(1), pp.121-125. DOI: https://doi.org/10.31004/jpdk.v1i2.622.

[2] Brzozowska-Woś, M., 2020. *Wpływ cyfrowej komunikacji marketingowej na angażowanie się w markę i współtworzenie jej wartości przez młodych konsumentów*. Wydawnictwo Politechniki Gdańskiej.

[3] Waworundeng, J.M.S., Sandag, G.A., Sahulata, R.A. and Rellely, G.D., 2022. Sentiment Analysis of Online Lectures Tweets using Naïve Bayes Classifier. *CogITo Smart Journal*, *8*(2), pp.371-384. DOI: https://doi.org/10.31154/cogito.v8i2.414.371-384.

[4] Sharma, P. and Moh, T.S., 2016, December. Prediction of Indian election using sentiment analysis on Hindi Twitter. In *2016 IEEE international conference on big data (big data)* (pp. 1966-1971). IEEE. DOI: https://doi.org/10.1109/BigData.2016.7840818.

[5] Yang, S. and Zhang, H., 2018. Text mining of Twitter data using a latent Dirichlet allocation topic model and sentiment analysis. *International Journal of Computer and Information Engineering*, *12*(7), pp.525-529.

[6] Younis, E.M., 2015. Sentiment analysis and text mining for social media microblogs using open source tools: an empirical study. *International Journal of Computer Applications*, *112*(5).

[7] Rane, A. and Kumar, A., 2018, July. Sentiment classification system of twitter data for US airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE. DOI: https://doi.org/10.1109/COMPSAC.2018.00114.

[8] Demidova, L.A., Klyueva, I.A. and Pylkin, A.N., 2019. Hybrid approach to improving the results of the SVM classification using the random forest algorithm. *Procedia Computer Science*, *150*, pp.455-461. DOI: https://doi.org/10.1016/j.procs.2019.02.077.

[9] Rezapour, M., 2021. Sentiment classification of skewed shoppers' reviews using machine learning techniques, examining the textual features. *Engineering Reports*, *3*(1), p.e12280. DOI: https://doi.org/10.1002/eng2.12280.

[10] Bansal, B. and Srivastava, S., 2018. Sentiment classification of online consumer reviews using word vector representations. *Procedia computer science*, *132*, pp.1147-1153. DOI: https://doi.org/10.1016/j.procs.2018.05.029.

[11] Nayak, A. and Natarajan, D., 2016. Comparative study of naive Bayes, support vector machine and random forest classifiers in sentiment analysis of twitter feeds. *International Journal of Advance Studies in Computer Science and Engineering (IJASCSE)*, *5*(1), p.16.

[12] Wang, J., Xu, M., Wang, H. and Zhang, J., 2006, November. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In *2006 8th international Conference on Signal Processing* (Vol. 3). IEEE. DOI: https://doi.org/10.1109/ICOSP.2006.345752.

[13] Sarakit, P., Theeramunkong, T. and Haruechaiyasak, C., 2015, August. Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)* (pp. 1-5). IEEE. DOI: https://doi.org/10.1109/ICAICTA.2015.7335373.

[14] Ramanathan, V. and Meyyappan, T., 2019, January. Twitter text mining for sentiment analysis on people's feedback about Oman tourism. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-5). IEEE. DOI: https://doi.org/10.1109/ICBDSC.2019.8645596.

[15] Astuti, I.F., Widagdo, P.P., Tanro, M.L.R., Cahyadi, D. and Suntara, A.A., 2023, February. Sentiment analysis on land and forest fire management in Twitter using Naïve Bayes method. In *AIP Conference Proceedings* (Vol. 2482, No. 1, p. 140004). AIP Publishing LLC. DOI: https://doi.org/10.1063/5.0110588.

[16] Vapnik, V.N., 2009. Statistics the elements of statistical learning. *Math. Intell*, *27*(2), pp.83-85.

[17] Russell, M.A., 2013. *Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more*. " O'Reilly Media, Inc.".

[18] [Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B. and Wang, X., 2015. Word embedding composition for data imbalances in sentiment and emotion classification. *Cognitive Computation*, *7*, pp.226-240. DOI: https://doi.org/10.1007/s12559-015-9319-y.

[19] Song, J., Huang, X., Qin, S. and Song, Q., 2016, June. A bi-directional sampling based on K-means method for imbalance text classification. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (pp. 1-5). IEEE. DOI: https://doi.org/10.1109/ICIS.2016.7550920.

[20] Gata, W., Amsury, F., Wardhani, N.K., Sugiyarto, I., Sulistyowati, D.N. and Saputra, I., 2019, April. Informative tweet classification of the earthquake disaster situation in indonesia. In *2019 5th International Conference on Computing Engineering and Design (ICCED)* (pp. 1-6). IEEE. DOI: https://doi.org/10.1109/ICCED46541.2019.9161135.

[21] Mohasseb, A., Bader-El-Den, M., Cocea, M. and Liu, H., 2018, July. Improving imbalanced question classification using structured smote based approach. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)* (Vol. 2, pp. 593-597). IEEE. DOI: https://doi.org/10.1109/ICMLC.2018.8527028.

[22] Al-Azani, S. and El-Alfy, E.S.M., 2017. Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text. *Procedia Computer Science*, *109*, pp.359-366. DOI: https://doi.org/10.1016/j.procs.2017.05.365.