E-ISSN 2808-5841
P-ISSN 2808-7283

Analisis Sentimen Persepsi Publik Terhadap UPN "Veteran" Jawa Timur Menggunakan Metode SVM, Naïve Bayes, dan Multinomial Logistic Regression

Rahmatul Amanillah[1], Sahat Renaldi. S[2], Prismahardi Aji Riyantoko[3]
[1, 2, 3] *Program Studi Sains Data, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur*
[1] 20083010002@student.upnjatim.ac.id
[2] 20083010026@student.upnjatim.ac.id
[3] prismahardi.aji.ds@upnjatim.ac.id
Corresponding author email: 20083010002@student.upnjatim.ac.id

Abstract: *The development of information technology has influenced various aspects of human life. Through the social media platform Twitter, users can express their opinions or thoughts regarding real-world events. These opinions can be utilized to gain deeper insights, one of which is through sentiment analysis. This study conducts a public sentiment analysis on UPN "Veteran" East Java. The researcher utilizes data obtained through web crawling using Twint. The researcher employs feature extraction methods such as Bag of Words (BoW) and Term Inverse - Document Term Frequency (TF-IDF), as well as algorithms SVM, Naive Bayes, and Multinomial Logistic Regression to build sentiment classification models. Based on the results of the analysis, which includes sentiment labeling using Python TextBlob, it is found that 12.9% of the data is classified as negative sentiment, 55.2% as neutral sentiment, and 31.9% as positive sentiment. Additionally, it is observed that there is an increasing trend in the number of tweets related to UPN "Veteran" East Java in February 2023. The evaluation results of the classification models indicate that Logistic Regression with BoW achieves the best performance, with an accuracy of 0.75, precision of 0.74, and recall of 0.60. The BoW feature extraction method appears to work better across all models tested in this study.*
Keywords: *UPNVJT, Sentiment Analysis, Feature Extraction, Classification*
Abstrak: Perkembangan teknologi informasi mempengaruhi berbagai lini kehidupan manusia. Melalui media sosial *Twitter* pengguna dapat mengutarakan opini/pendapatnya terkait kejadian di dunia nyata. Opini/pendapat tersebut dapat dimanfaatkan untuk menggali *insight* lebih, salah satunya adalah melalui analisis sentimen. Penelitian ini melakukan analisis sentimen publik terhadap UPN "Veteran" Jawa Timur. Peneliti menggunakan data yang diambil dengan Teknik *crawling* dengan *Twint*. Peneliti menggunakan metode ekstraksi fitur *Back of Word* (BOW) dan *Term Invers – Document Term Frequency* (TF-IDF) dan algoritma SVM, *Naive Bayes*, dan *Multinomial Logistic Regression* untuk membangun model klasifikasi sentimen. Berdasarkan hasil dari proses analisis pada data yang telah dilakukan pelabelan sentimen dengan *Python TextBlob*, 12.9% dikelompokan dalam sentimen negatif, 55.2% sentimen netral, dan 31.9% sentimen positif. Selain itu, didapatkan informasi bahwa terjadi peningkatan trend jumlah *tweet* terkait UPN "Veteran" Jawa Timur bulan Februari 2023. Hasil uji dari model klasifikasi menunjukan bahwa *Logistic Regression* dengan BoW menghasilkan kinerja terbaik akurasi yang didapatkan sebesar 0.75, presisi sebesar 0.74 dan *recall* sebesar 0.60. Metode ekstraksi fitur BoW tampak bekerja lebih baik pada semua model yang diuji dalam penelitian ini.
Kata kunci: UPNVJT, Analisis Sentimen, Ekstraksi Fitur, Klasifikasi

I. PENDAHULUAN

Pengaruh teknologi dalam kehidupan manusia sekarang sangat cepat. Salah satunya dapat dilihat dari berbagai bentuk aktivitas kehidupan yang dilakukan melalui media sosial. *Twitter* adalah salah satu media sosial yang diminati sekarang ini. *Twitter* kerap menjadi pilihan penggunanya untuk mengutarakan opini/pendapat terkait hal-hal yang terjadi di dunia nyata. Melalui fitur *trending topic* dan *keyword search*, pengguna dapat mengetahui topik yang sedang menjadi pembicaraan dan juga dapat mengetahui opini/pendapat pengguna lainnya terkait hal tersebut. Berbicara mengenai opini/pendapat pengguna media sosial terkait suatu topik tertentu merupakan hal yang menarik untuk diteliti. Data teks terkait opini/pendapat tersebut bisa dimanfaatkan untuk menggali *insight* yang lebih, salah satunya adalah melalui analisis sentimen.

Analisis sentimen ialah proses mengekstraksi, mengolah dan memahami data berupa teks yang tidak terstruktur secara otomatis guna mengambil informasi sentimen yang terdapat pada sebuah kalimat pendapat atau opini [1]. Analisis sentimen digunakan untuk menilai opini dan kecondongan suatu data terhadap suatu topik yang dikelompokkan pada kelas tertentu. Media sosial seperti *Twitter*

memberikan kesempatan bagi para peneliti untuk mencari *insight* dari data *tweet* pengguna sesuai kebutuhan penelitiannya.

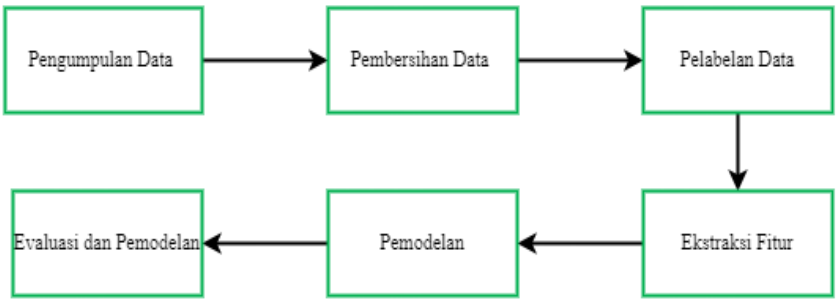
Dalam mengembangkan sistem untuk analisis sentimen, ada banyak metode yang bisa dilakukan. Penelitian yang dilakukan Arif, dkk. yaitu melakukan analisis sentimen terhadap maskapai penerbangan menggunakan metode *Naive Bayes* dengan seleksi fitur *information gain* menghasilkan rata-rata akurasi sebesar 80% [2]. Primandani, dkk. juga melakukan analisis sentimen yang berkaitan dengan adanya wacana pemindahan ibukota. Berdasarkan hasil pengujian terhadap *tweet* sentimen pemindahan ibukota dari media sosial *Twitter* sebanyak 1236 *tweet* (404 positif dan 832 negatif) menggunakan metode SVM diperoleh akurasi sebesar 96.68% [3]. Erna, dkk. dalam penelitiannya terkait analisis sentimen nasabah pada layanan perbankan didapatkan hasil bahwa model terbaik untuk mengklasifikasikan sentimen layanan pada BRI adalah *SMOTE-SVM* dengan Karnel RBF dengan akurasi mencapai 99.7% pada data *training* dan 99.2% pada data *testing*, sedangkan untuk Bank Mandiri model terbaik adalah dengan *SMOTE-Naive Bayes* yang berhasil mencapai nilai AUC sebesar 99.6% [4]. Pada tahun 2020, Imamah, dkk. melakukan penelitian sentimen analisis terhadap kasus Covid-19 menggunakan 355.384 data *tweest*. Dengan metode *Logistic Regression* dan ekstraksi fitur TF-IDF didapatkan akurasi sebesar 94.71% [5].

Pada penelitian ini, peneliti melakukan analisis sentimen publik terhadap UPN "Veteran" Jawa Timur di media sosial *Twitter*. UPN "Veteran" Jawa Timur merupakan perguruan tinggi negeri terakreditasi A di Jawa Timur. Untuk meningkatkan kualitas dan mutu universitas, UPN "Veteran" Jawa Timur perlu mengetahui opini publik terkait reputasi dan pelayanan serta pengabdian yang dilakukan selama ini. Opini publik dapat menjadi tolak ukur keberhasilan pelayanan yang ada di UPN "Veteran" Jawa Timur. Bila terdapat lebih banyak sentimen negatif, maka UPN "Veteran" Jawa Timur harus berbenah untuk meningkatkan reputasinya di masa depan sehingga tidak mempengaruhi tingkat minat masyarakat yang terhadap universitas. Peneliti mengambil data dari media sosial *Twitter* dengan *keyword search* "upnjatim" dan "upnvjt" yang terhitung sejak tanggal 1 Oktober 2022 sampai 25 Maret 2023. Total terdapat 909 baris dan 36 kolom data yang didapatkan dari proses *crawling*. Data mentah ini kemudian akan melewati proses pembersihan sebelum masuk dalam tahap pemodelan. Penelitian ini dilakukan untuk membangun model analisis sentimen untuk memberikan gambaran dan masukan terhadap UPN "Veteran" Jawa Timur supaya dapat meningkatkan kualitas dan mutu dengan menggunakan analisis sentimen sebagai tolak ukur penilaian terhadap institusi tersebut.

Berdasarkan tinjauan literatur dari penelitian sebelumnya, metode SVM, *Naive Bayes* dan *Multinomial Logistic Regression* memiliki kemampuan yang baik dalam mengklasifikasi sentimen. Oleh karena itu, peneliti mengusulkan metode SVM, *Naive Bayes* dan *Multinomial Logistic Regression* untuk diterapkan pada *tweet* terkait *keyword search* yang diambil dalam penelitian ini. Peneliti juga melakukan ekstraksi fitur dengan metode *Back of Word* (BOW) dan *Term Frequency - Invers Document Frequency* (TFIDF) serta membandingkan kinerja dari ketiga model.

II. METODE PENELITIAN

Metodologi pada suatu penelitian merujuk pada alur atau langkah-langkah sistematis untuk mencapai tujuan penelitian sesuai dengan yang direncanakan. Metodologi penelitian dirancang dengan alur yang terstruktur dan ilmiah. Dalam penelitian ini, alur penelitian dimulai dari proses pengumpulan data yang akan digunakan, pembersihan data, pengelompokan sentimen, ekstraksi fitur untuk memperkecil dimensi data, pemodelan, sampai dengan evaluasi dan validasi model. Peneliti menggunakan *Confusion Matrix* dan *k-Fold cross validation* untuk mengukur kebaikan model.



Gambar 1. Alur Penelitian

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini didapatkan dari hasil proses *crawling* data *Twitter* menggunakan *Twint*. *Keyword* yang digunakan untuk mengambil data adalah "upn jatim" dan "upnvjt". Penelitian ini mengambil data *tweet* yang terhitung sejak tanggal 1 Oktober 2022 sampai 25 Maret 2023. Total terdapat 909 baris dan 36 kolom data mentah yang dihasilkan dari proses *crawling*.

3.2 Pembersihan Data

Data yang diperoleh dari proses crawling merupakan data mentah yang masih mengandung unsur-unsur yang tidak diperlukan. Pembersihan dilakukan untuk mempersiapkan data sebelum melakukan pemodelan. Pada data sebanyak 909 baris dan 36 kolom akan dilakukan pembersihan mulai dari penghapusan kolom yang tidak diperlukan, tautan (*URLs*), *mention*, konversi tagar (*hashtag*) menjadi kata, penghapusan simbol dan angka, serta pola-pola lainnya. Proses dilanjutkan dengan *case folding* atau mengubah setiap kata ke dalam huruf kecil, *stopword* atau pembuangan *term* yang tidak memiliki arti/tidak relevan, *stemming* atau penghapusan imbuhan suatu kata, penghapusan *emoticon*, *tokenizing* atau pemotongan *string* input berdasarkan kata yang menyusunnya, konversi slang/kata-kata informal menjadi formal, penghapusan tanda baca, baris yang kosong, dan data duplikat.

3.3 Pengelompokan Data

Data *tweet* dikategorikan dalam 3 label berdasarkan sentimen yaitu label -1 (negatif), 0 (netral), dan 1 (positif). Pengelompokan ini dilakukan dengan menggunakan *TextBlob*. TextBlob adalah sebuah pustaka di Python untuk memproses data teks [6]. Analisis sentimen menggunakan TextBlob hanya tersedia dalam bahasa Inggris, sehingg peneliti menerjemahkan data dari hasil pra-pemrosesan ke dalam bahasa Inggris sebelum dilakukan analisis sentimen.

3.4 Ekstaksi Fitur

Setelah dilakukan proses pembersihan data, data *tweet* kemudian diekstraksi untuk memperkecil dimensi data. Ekstraksi fitur merupakan proses penting pada klasifikasi teks untuk mengubah format tekstual yang tidak terstruktur menjadi terstruktur sehingga dapat diproses oleh algoritma *machine learning* untuk mengklasifikasikan ke *class* yang telah ditentukan [7]. Dalam penelitian ini, peneliti menggunakan metode ekstraksi fitur *Back of Word* (BoW) dan *Term Frequency - Invers Document Frequency* (TF-IDF). Metode ekstraksi fitur dengan *Back of Word* (BoW) menggambarkan jumlah kemunculan suatu kata di dalam dokumen. Gambar 2 merepresentasikan bagaimana *Back of Word* (BoW) bekerja mengekstraksi sebuah fitur.

E-ISSN 2808-5841
P-ISSN 2808-7283

Dataset

Doc1	upn kampus bela negara
Doc2	fakultas kedolteran upn jatim
Doc3	kampus upn bersih

Vocabulary (sorted) :
["bela", "bersih", "fakultas", "jatim", "kampus", "kedokteran", "negara", "upn"]

Bag of Word (BoW) :

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Doc1	1	0	0	0	1	0	1	1
Doc2	0	0	1	1	0	1	0	1
Doc3	0	1	0	0	1	0	0	1

Gambar 2. Ilustrasi Cara Kerja Metode Back of Word (BoW)

Sedangkan, metode *Term Frequency - Invers Document Frequency* (TF-IDF) adalah suatu proses untuk memberikan bobot relasi suatu kata (*term*) dengan dokumen. Proses TF-IDF mengkombinasikan dua skema perhitungan bobot, yakni frekuensi kemunculan suatu kata di dalam suatu dokumen dan *inverse* frekuensi dokumen yang memuat kata tersebut. Metode TF-IDF menentukan frekuensi relatif kata-kata dalam dokumen tertentu melalui *invers* proporsi kata di seluruh korpus dokumen. *Invers Document Frequency* adalah matrik untuk menentukan seberapa jarang suatu kata didasarkan pada suatu dokumen [8]. Berikut adalah persamaan untuk menghitung nilai IDF:

$$IDFt = log\ log\ 10\left(\frac{D}{dfi}\right) \quad (1)$$

$$Wd,t = tf\ d,t * IDFt \quad (2)$$

Mengacu pada (1) nilai D merupakan jumlah dokumen yang berisi *term* (t) dan nilai dfi merupakan jumlah kemunculan kata terhadap D. Adapun formula yang digunakan untuk menghitung bobot (w) masing-masing dokumen ditunjukkan pada persamaan (2) dengan Wd, t berarti bobot dokumen ke-d pada kata ke-t, dan tf d, t adalah frekuensi kata.

3.5 Pemodelan

Dalam penelitian ini, peneliti menggunakan metode klasifikasi SVM (*Support Vector Machine*), *Naive Bayes*, dan *Multinomial Logistic Regression*. Metode pengklasifikasian dengan SVM dilakukan pada pengaturan *kernel* RBF (*Radial Basis Function*). Peneliti menggunakan jenis *kernel* ini karena berdasarkan studi literatur yang telah dilakukan, RBF *kernel* memiliki kinerja yang baik untuk kasus *multiclass classification*. Penelitian yang dilakukan oleh Harun, dkk. yang membandingkan *kernel* SVM dalam proses klasifikasi *multiclass Human Development Index* (HDI). Hasil klasifikasi *Human Development Index* (HDI) dengan menggunakan *kernel* RBF merupakan kernel terbaik untuk mengatasi masalah HDI [9].

Untuk melakukan proses pemodelan diperlukan hasil dari data *tweet* yang telah diekstraksi fitur. Peneliti membandingkan hasil antara dua metode ekstraksi fitur yang digunakan. Persentase jumlah data yang digunakan sebagai data *training* adalah sebesar 80%. Setelah dilakukan *training process*, maka akan dilakukan pengujian terhadap data *testing*.

3.6 Evaluasi dan Validasi

Evaluasi dan validasi adalah langkah untuk mengetahui performa atau kinerja dari model yang telah dibangun. Dalam kasus ini digunakan *Confusion Matrix* dan *k-Fold cross validation* untuk

mengukur kebaikan model. Setelah dilakukan pengujian pada data *testing*, maka akan dilakukan perbandingan hasil klasifikasi model dengan hasil klasifikasi yang sebelumnya sudah didefinisikan. Peneliti juga membandingkan hasilnya dengan data *tweet* yang telah disintesis dengan SMOTE (*Synthetic Minority Oversampling Technique*). Model terbaik berdasarkan hasil pada *confusion matrix* kemudian akan dilakukan validasi dengan *k-Fold cross validation* untuk bisa lebih detail melihat kemampuan model. *K-Fold cross validation* bekerja dengan membagi data ke dalam sejumlah nilai dan melakukan iterasi sebanyak k.

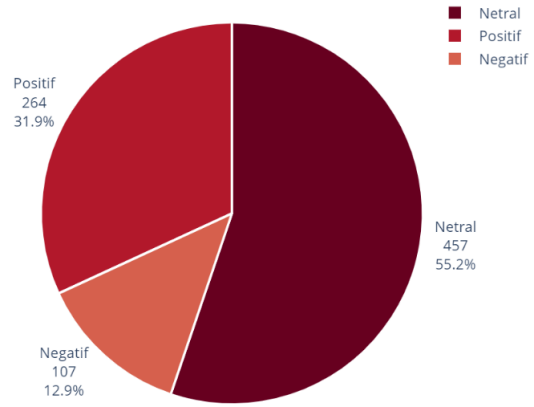
III. HASIL DAN PEMBAHASAN

4.1 Pembersihan Data

Peneliti menggunakan modul NLTK yang tersedia di *python* sebagai salah satu alat untuk melakukan pembersihan terhadap data *tweet*. Data awal yang didapatkan dari proses *crawling* terdiri dari 909 baris dan 36 kolom direduksi menjadi 828 baris dan 4 kolom. Adapun kolom yang digunakan untuk proses analisis berikutnya adalah kolom date (tanggal *tweet* dikirim), *tweet* (data text), tweet_eng (data text dalam bahasa Inggris), dan *sentiment* (label).

4.2 Sentimen

Proses pengelompokan sentimen dilakukan secara otomatis dengan bantuan *package* TextBlob pada data kolom tweet_eng. Dari total 828 *tweet*, terdapat 107 *tweet* yang masuk dalam sentimen negatif, 457 *tweet* netral, dan sisanya sebesar 164 merupakan *tweet* positif.

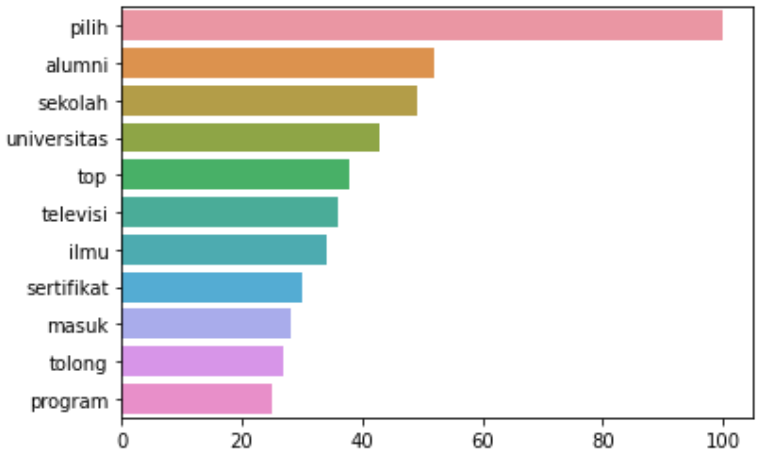


Gambar 3. Sentimen

4.3 Explorasi Data Analisis (EDA)

Peneliti melakukan *exploratory* data analisis untuk mendapatkan *insight* tambahan dari data yang telah dibersihkan. Peneliti menggali informasi terkait kata-kata yang sering muncul pada sentimen positif, melihat sentimen negatif dengan visualisasi *wordcloud*, time series analisis banyaknya *tweet* terkait UPN "Veteran" Jawa Timur dari hari ke hari, serta melihat banyaknya sentimen negatif, netral, dan positif di setiap bulannya dengan visualisasi *stacked bar chart*.

E-ISSN 2808-5841
P-ISSN 2808-7283



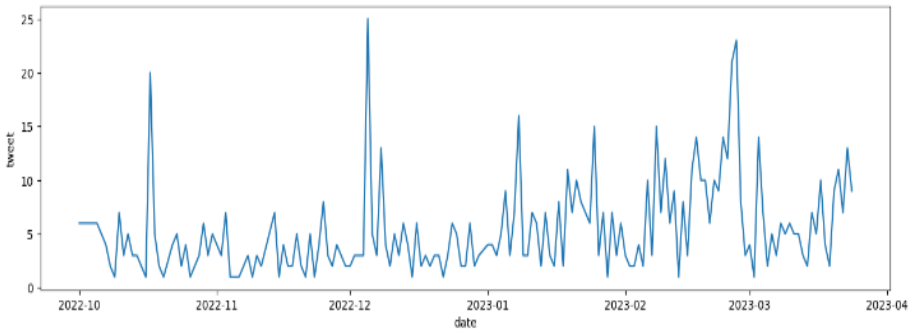
Gambar 4. Kata yang Sering Muncul terkait Sentimen Positif

Pada Gambar 4 dapat dilihat bahwa kata "pilih", "alumni", "sekolah", "universitas", "top", merupakan lima kata teratas yang paling sering muncul pada *tweet* sentimen positif. Kata-kata tersebut merepresentasikan ketertarikan terhadap UPN "Veteran" Jawa Timur. Kata "universitas" dan "top" mengindikasikan tanggapan public bahwa UPN "Veteran" Jawa Timur merupakan salah satu universitas yang dikenal baik di kalangan masyarakat. Dilanjutkan dengan adanya kata "pilih", "alumni", dan "sekolah" memberikan gambaran bahwa UPN "Veteran" Jawa Timur merupakan salah satu pilihan bagi banyak siswa dari berbagai sekolah untuk melanjutkan pendidikan.

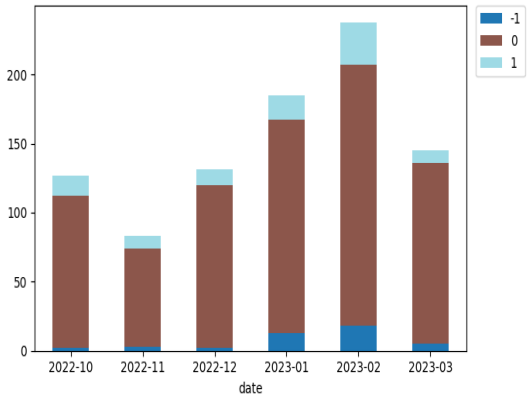
Gambar 5. *Wordcloud* Sentimen Positif

Dari *wordcloud* yang ditampilkan pada Gambar 5 memuat kata "ketat", "akreditasi", "snbp" yang sebelumnya tidak diketahui. Kata "ketat" dan "snbp" dapat mengindikasikan bahwa ada aturan atau kebijakan yang ketat dalam hal seleksi untuk bisa diterima di UPN "Veteran" Jawa. Selain itu, tampak juga kata "akreditasi" yang mungkin menjadi salah satu fokus utama yang dipertimbangkan banyak orang dalam memilih universitas.

E-ISSN 2808-5841
P-ISSN 2808-7283



Gambar 6. Jumlah Tweet per Hari



Gambar 7. Jumlah Tweet per Bulan

Pada Gambar 6 dan Gambar 7 dapat dilihat bahwa terjadi peningkatan jumlah tweet pada periode bulan Februari 2023. Peneliti mengindikasikan peningkatan ini berkaitan dengan fenomena yang terjadi pada bulan ini yaitu penerimaan mahasiswa baru melalui jalur SNBP (Seleksi Nasional Berbasis Prestasi).

4.4 Ekstraksi Fitur

Setelah sebelumnya telah melalui tahap preprocessing, *tweet* yang tersisa kemudian dilakukan ekstraksi fitur untuk memperkecil dimensi data. Proses ekstraksi dilakukan dengan memberi pembobotan pada setiap vektor kata untuk setiap kalimat. Metode ekstraksi fitur yang digunakan adalah *Back of Word* (BoW) dan *Term Frequency - Inverse Document Frequency* (TF-IDF).

Tabel 1. Tweet Mentah dan Tweet Bersih

Ekstraksi Fitur	Train (80%)	Test (20%)
Data awal	662	166
BoWs	8167	1881
TF-IDF	8167	1881

4.5 Pemodelan

Penelitian ini mengimplementasikan algoritma SVM, *Naive Bayes*, dan *Logistic Regression* pada sentimen di media sosial *Twitter* terkait UPN "Veteran" Jawa Timur. Perbandingan data *training* dan data *testing* adalah 8 : 2. Metode pengklasifikasian dengan SVM dilakukan pada pengaturan kernel RBF (*Radial Basis Function*) dengan random *state* 42, serta pada *Logistic Regression* dengan pengaturan *multiclass multinomial* dengan random *state* yang sama. Setelahnya dilakukan ekstraksi fitur dengan menjalankan model pada data *testing* untuk melihat nilai akurasi, presisi, dan *recall*.

Selain itu, peneliti juga membandingkan hasil pemodelan terhadap data yang disintesis dengan SMOTE. *Synthetic Minority Over sampling Technique* (SMOTE) merupakan salah satu metode yang dapat diterapkan untuk menangani ketidakseimbangan sentimen/label. Peneliti mengimplementasikan SMOTE pada data *training*.

Tabel 2. Akurasi, Presisi, *Recall* tanpa SMOTE

Metode	BOW			TF-IDF		
	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall
SVM	0.66	0.44	0.44	0.67	0.68	0.46
Naive Bayes	0.72	0.64	0.56	0.62	0.46	0.39
Multinomial Logistic Regression	**0.75**	**0.74**	**0.60**	0.70	0.84	0.50

Tabel 3. Akurasi, Presisi, *Recall* dengan SMOTE

Metode	BOW			TF-IDF		
	Akurasi	Presisi	Recall	Akurasi	Presisi	Recall
SVM	0.59	0.49	0.48	0.70	0.79	0.52
Naive Bayes	0.55	0.52	0.53	0.58	0.51	0.53
Multinomial Logistic Regression	0.58	0.54	0.59	0.73	0.69	0.61

Hasil pengujian data tanpa SMOTE pada tabel 2 terlihat bahwa metode *Logistic Regression* bekerja lebih baik saat dikombinasikan dengan BoW. Nilai akurasi yang didapatkan sebesar 0.75, presisi sebesar 0.74 dan *recall* sebesar 0.60. Hasil tersebut merupakan hasil yang paling tinggi jika dibandingkan dengan metode *Support Vector Machine* (SVM) dan *Naive Bayes* yang dikombinasikan dengan ekstraksi fitur BoW maupun TF-IDF. Namun, antara metode ekstraksi fitur BoW dengan TF-IDF, metode BoW cenderung memberikan hasil yang lebih baik.

Pada data yang disintesis dengan SMOTE, hasil akurasi, presisi, *recall* tidak lebih baik dibandingkan tanpa SMOTE. Data yang diekstraksi dengan TF-IDF dan disintesis dengan SMOTE, menghasilkan skor yang lebih tinggi dibandingkan jika data hasil SMOTE tersebut diekstraksi dengan BoW di setiap model yang diuji.

4.6 Evaluasi dan Validasi

Pemodelan data *tweet* terkait topik UPN "Veteran" Jawa Timur dengan metode *Logistic Regression* dan BoW menunjukkan performa paling baik dibandingkan dengan yang lain, Untuk memastikan lebih detail terkait kebaikan model *Logistic Regression* dengan kombinasi BoW ini, maka akan dilakukan validasi dengan *k-Fold cross validation*. Pada tahap validasi *dengan k-Fold cross validation*, data *training* dibagi secara random ke dalam k bagian dengan perbandingan yang sama. Proses validasi pada penelitian ini menggunakan nilai k = 5.

Tabel 4. Nilai Akurasi, Presisi, *Recall*, dengan Validasi k-Fold

k-Fold	Akurasi	Presisi	Recall
1	0.74	0.77	0.61
2	0.69	0.71	0.58
3	0.82	0.88	0.74
4	0.77	0.76	0.64
5	0.76	0.80	0.63

Pengujian dengan menggunakan *k-Fold cross validation*, nilai akurasi, presisi, dan *recall* tertinggi adalah pada saat nilai k adalah 3. Nilai akurasi saat k = 3 adalah 0.82, presisi sebesar 0.88, dan recall 0.74.

IV. KESIMPULAN

Berdasarkan hasil analisis data, didapatkan 12.9% dikelompokan dalam sentimen negatif, 55.2% sentimen netral, dan 31.9% sentimen positif. Melalui proses eksplorasi data analisis diketahui terdapat peningkatan jumlah *tweet* terkait UPN "Veteran" Jawa timur pada bulan Februari 2023. Hasil dari pengujian model klasifikasi sentiment dari beberapa model didapatkan metode *Logistic Regression* bekerja lebih baik saat dikombinasikan dengan BoW. Pengujian pada data tanpa SMOTE dengan *Logistic Regression* dan BoW menghasilkan akurasi sebesar 0.75, presisi sebesar 0.74 dan *recall* sebesar 0.60. Pengujian dengan menggunakan *k-Fold cross validation*, akurasi, presisi, dan *recall* tertinggi adalah pada saat nilai k adalah 3. Metode ekstraksi fitur BoW tampak bekerja lebih baik pada semua model yang diuji dalam penelitian ini.

UCAPAN TERIMA KASIH

REFERENSI

1. B. Brahimi, M. Touahria, dan A. Tari, "Improving sentiment analysis in Arabic: A combined approach," J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 10, hal. 1242–1250, 2021, doi: 10.1016/j.jksuci.2019.07.011.
2. A. B. P. Negara, H. Muhardi, dan I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 3, hal. 599, 2020, doi: 10.25126/jtiik.2020711947.
3. P. Arsi dan R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," J. Teknol. Inf. dan Ilmu Komput., vol. 8, no. 1, hal. 147, 2021, doi: 10.25126/jtiik.0813944.
4. E. D. N. Sari dan I. Irhamah, "Analisis Sentimen Nasabah pada Layanan Perbankan Menggunakan Metode Regresi Logistik Biner, Naïve Bayes Classifier (NBC), dan Support Vector Machine (SVM)," J. Sains dan Seni ITS, vol. 8, no. 2, 2020, doi: 10.12962/j23373520.v8i2.44565.
5. Imamah dan F. H. Rachman, "Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regresion," Proceeding - 6th Inf. Technol. Int. Semin. ITIS 2020, hal. 238–242, 2020, doi: 10.1109/ITIS50118.2020.9320958.
6. D. Hazarika, G. Konwar, S. Deb, dan D. J. Bora, "Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing," Proc. Int. Conf. Res. Manag. Technovation 2020, vol. 24, hal. 63–67, 2020, doi: 10.15439/2020km20.
7. M. R. Faisal dan D. T. Nugrahadi, "2517-5933-1-Pb," vol. 8, no. 1, hal. 62–69, 2020.
8. A. S. Neogi, K. A. Garg, R. K. Mishra, dan Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data," Int. J. Inf. Manag. Data Insights, vol. 1, no. 2, hal. 100019, 2021, doi: 10.1016/j.jjimei.2021.100019.
9. H. Al Azies, D. Trishnanti, dan E. Mustikawati P.H, "Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI)," IPTEK J. Proc. Ser., vol. 0, no. 6, hal. 53, 2019, doi: 10.12962/j23546026.y2019i6.6339.
10. Z. Drus dan H. Khalid, "Sentiment analysis in social media and its application: Systematic literature review," Procedia Comput. Sci., vol. 161, hal. 707–714, 2019, doi:

10.1016/j.procs.2019.11.174.Arni, U. D. (2018). *Apa Itu Text Mining?*. Tersedia dari: Www.Garudacyber.Co.Id. https://garudacyber.co.id/artikel/1254-apa-itu-text-mining