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

Adaptive Synthetic Support Vector Machine Multiclass untuk mengklasifikasikan *Imbalance data* pada Sentimen kenaikan Bahan Bakar Minyak

Ismatullah[1], Fatkhurokhman Fauzi[2], Indah Manfaati Nur[3],

[1,2,3]*Program Studi Statistika, Universitas Muhammadiyah Semarang*
Corresponding author email: fatkhurokhmanf@unimus.ac.id

Abstract: *The phenomenon of increasing fuel oil (BBM) has become a trending topic for people in Indonesia in September 2022. Indonesian people from various provinces have chosen their opinions regarding this phenomenon on social media, one of which is Twitter. Sentiment analysis is used to describe a person's opinion on social media about a phenomenon. In this study, we will look at public sentiment regarding the increase in fuel prices which is labeled into three categories, namely positive, neutral and negative. This study applies adaptive synthetics to overcome data imbalances caused by negative sentiment. The data used in this research is public opinion related to the increase in fuel prices in every province in Indonesia. Each province is limited to 100 opinions. The classification method applied to this research is the multiclass Support Vector Machine (SVM). The results obtained are that people in all provinces in Indonesia have a negative opinion regarding the increase in fuel prices. The results of the multiclass SVM classification show an average accuracy of 87.94%, with the highest accuracy of 95%.*
Keywords: *Adaptive Synthetic, BBM, Sentiment, SVM*
Abstrak: Fenomena kenaikan Bahan Bakar Minyak (BBM) menjadi *trending* topik masyarakat di Indonesia pada bulan September 2022. Masyarakat Indonesia dari berbagai provinsi menyuarakan pendapatnya terkait fenoemna tersebut di media sosial salah satunya *twitter*. Analisis sentimen digunakan untuk mengetahui gambaran pendapat seseorang di media sosial terhadap suatu fenomena. Pada penelitian ini akan melihat sentimen masyarakat terkait kenaikan BBM yang dilabelkan kedalam tiga kategori yaitu positif, netral, dan negatif. Penelitian ini menerapkan *adaptive synthetic* untuk menangani *imbalance data* yang disebabkan oleh sentimen negatif. Data yang digunakan pada penelitian ini adalah pendapat masyarkat terkait kenaikan BBM disetiap provinsi di Indonesia. Setiap provinsi dibatasi 100 pendapat. Metode klasifikasi yang diterapkan pada penelitian adalah Support Vector Machine (SVM) *multiclass*. Hasil yang didapat adalah masyarakat diseluruh provinsi di Indonesia berpendapat negatif terkait kenaikan BBM. Hasil klasifikasi SVM *multiclass* menunjukan rata-rata akurasi sebesar 87.94%, dengan akurasi tertinggi sebesar 95%.
Kata kunci: *Adaptive Synthetic*, BBM, Sentimen, SVM

I. PENDAHULUAN

Kenaikan harga Bahan Bakar Minyak (BBM) mengundang pro dan kontra di masyarakat Indonesia. Fenomena ini mengundang masyarakat berkomentar di media sosial tentang pandangannya terkait kenaikan BBM. Sebagai negara demokrasi Indonesia menjunjung tinggi kebebasan berpendapat. Pendapat tersebut dapat dijadikan sebagai acuan dalam evaluasi kebijakan pemerintah dalam kasus ini adalah kenaikan BBM.

Twitter menjadi salah satu *platform* media sosial yang digunakan masyakarat untuk menyuarakan pendapatnya [1]. Pendapat masyarakat yang terdapat di twitter dapat dianalisis, metode tersebut dinamakan analisis sentimen. Tujuan dari analisis sentimen yaitu mengklasifikiasi polaritas teks pada dokumen, kalimat, dan fitur serta menentukan apakah komentar/pendapat tersebut bersentimen positif, netral, atau negatif[2]. Analisis sentimen tentang kenaikan harga BBM perlu dilakukan untuk mengetahui respon pendapat/pandangan masyarakat khususnya pengguna twitter mengenai kenaikan harga BBM yang ditetapkan oleh pemerintah pada tanggal 3 September 2022. Sentimen dari masyarakat tentang kenaikan harga BBM merupakan faktor yang penting bagi pemerintah untuk menetapkan suatu kebijakan atau mengevaluasi suatu kebijakan.

Klasifikasi teks atau kategorisasi teks merupakan proses yang secara otomatis menempatkan dokumen teks ke dalam suatu kategori berdasarkan isi dari teks tersebut. Klasifikasi teks berbasis statistik, klasifikasi teks berbasis koneksi, dan klasifikasi teks berbasis aturan adalah tiga kategori utama pendekatan klasifikasi teks telah digunakan dalam penelitian. Teknik klasifikasi berbasis statistik berkinerja lebih baik daripada yang lain [3].

Salah satu metode klasifikasi berbasis statisitk yang dapat digunakan untuk mengklasifikasi dokumen yaitu *Support Vector Machine* (SVM). Penelitian SVM *multiclass* pernah dilakukan oleh Irmanda dan Astriratma [4] dengan mengklasifikasikan jenis pantun. Hasil akurasi yang didapat oleh SVM multiclass sebesar 81.91%. Hal ini menunjukan bahwa metode SVM *multiclass* memiliki performasi yang sangat baik. Penelitian lainnya terkait metode SVM *multiclass*[5]–[7]

Sebagian besar studi menggunakan teknik klasifikasi standar yang mengasumsikan bahwa data *training* didistribusikan secara merata di semua kategori. Terlepas dari kenyataan bahwa ketidakseimbangan jumlah data *training* sering ditemui dalam praktiknya. Akan ada kelas minoritas dan kelas mayoritas di *dataset*. Kinerja pengklasifikasi teks seringkali mengalami penurunan ketika dihadapkan pada kondisi tersebut [8]. Untuk menyeimbangkan jumlah data *training* yang digunakan per estimasi, peneliti menggunakan metode "*Adaptive Synthetic*" untuk menyelesaikan permasalahan tersebut.

Penelitian ini bertujuan untuk mengklasifikasi *tweet* dari pengguna Twitter pada setiap provinsi di Indonesia terkait kenaikan harga BBM. *Dataset* dibagi menjadi *training set* dan *testing* set kemudian *testing set* diklasifikasikan berdasarkan model SVM *multiclass* pada *training set* yang telah dilakukan *preprocessing*. Hasil klasifikasi SVM *multiclass* akan dievaluasi berdasarkan nilai akurasi.

II. METODE PENELITIAN

Pengumpulan data dalam penelitian ini menggunakan metode *scraping*. *Web scraping* adalah teknik untuk mengekstraksi informasi secara otomatis dari suatu *website* atau beberapa *website* [9]. Teknik ini menghasilkan konten yang relevan berdasarkan kueri dan mengubah format tidak terstruktur menjadi representasi terstruktur [10]. Pada tahap *scraping*, peneliti mengambil data *tweet* dari pengguna Twitter terkait kenaikan harga BBM. Di bawah ini adalah contoh data yang dikumpulkan oleh peneliti.

Tabel 1. Sampel Data Hasil *Scraping*

Username	Tweet	User Location
@SilviaPutrii9	Penyesuaian bbm bertujuan untuk mengurangi beban subsidi besar akibat harga minyak dunia yang terus naik, kebijakan pemerintah sudah benar#BantuanBBMUntukRakyat #BLTBBMTepatSasaran #HematCermatBBM	Pontianak, Kalimantan Barat
@inthesky014	BBM naik itu biar yang sering bawa motor ke masjid pada jalan kaki, supaya pahalanya lebih banyak, subhanallah pemerintah kita ini sangat memperdulikan iman rakyatnya	Cianjur, Indonesia
@Kentringmanikk	@M45Broo_ Saya bisnis laundry dukung 1000% BBM naik. Bila perlu naikan terus, hilangkan subsidi BBM. Larang mobil pake BB fosil secepatnya. Subsidi bisa di pakai utk hal lain. Sudah tdk efisien negeri ini	Tanjung Emas, Indonesia

Jumlah data yang diambil pada setiap provinsi sebanyak 100 *tweet* dengan total *tweet* yang digunakan dalam penelitian ini sebanyak 3400. Data yang diambil adalah *tweet* dalam bahasa

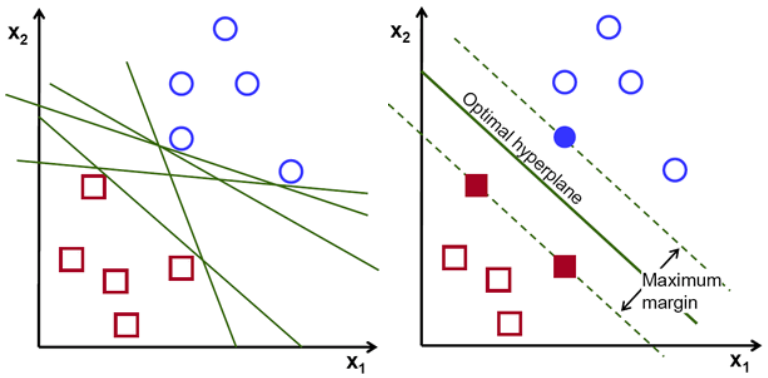
Indonesia. Kategori sentimen sebanyak tiga kategori yaitu sentimen positif, sentimen netral, dan sentimen negatif.

Langkah pertama pada penelitian ini adalah *text preprocessing*. *Text preprocessing* dilakukan untuk mengelola data teks sebelum membuat model *machine learning* [11]. Misalnya, menghapus tagar, URL, *stopword*, tanda baca, *@username*, dan karakter duplikat dalam sebuah kata [12]. Selain itu, *case folding*, normalisasi kata, *cleansing*, pemfilteran/penghapusan kata berhenti, *stemming*, dan tokenisasi juga dilakukan dalam *preprocessing* teks [13].

a) Case Folding
   Data diubah menjadi huruf kecil sehingga huruf besar dan huruf kecil dengan arti yang sama tidak diperlakukan berbeda [14].
b) Word Normalization
   Normalisasi kata digunakan untuk mengubah kata yang tidak baku atau disingkat menjadi kata baku dalam Bahasa Indonesia.
c) Cleansing
   Cleansing digunakan untuk membersihkan kata-kata yang tidak diperlukan seperti *hashtag* (#), alamat *website*, *username* (@username), angka, emoji dan *email*.
d) Filtering/Stopwords Removal
   Stopwords removal adalah langkah utama dalam *preprocessing* teks di *Natural Language Preprocessing* (NLP). Tahap ini menyaring kata-kata yang mengandung sedikit informasi atau kata yang tidak ada makna semantik dari teks yang diberikan [15].
e) Stemming
   Langkah ini dilakukan untuk menemukan akar kata dengan mengahapus awalan atau akhiran.
f) Tokenization
   Tokenisasi mengacu pada proses mengubah teks apa pun menjadi serangkaian token, masing-masing berbeda dan tidak bergantung satu sama lain.

Tahap kedua adalah penerapan metode *Adaptive Synthetic* (ADASYN) *Imbalanced class data* adalah kondisi data yang tidak seimbang antar kelas data [16]. Kondisi data yang tidak seimbang merupakan masalah dalam klasifikasi karena *learning classifier* akan cenderung memprediksi kelas data mayoritas dibandingkan dengan kelas minoritas. Akibatnya, akurasi prediksi yang dihasilkan menguntungkan untuk sebagian besar kelas data *training*, sedangkan untuk kelas minoritas akan menghasilkan akurasi prediksi yang buruk (Chawla, 2003).Kami mengusulkan metode adaptif untuk memfasilitasi pembelajaran dengan *imbalanced class data*. Tujuannya di sini adalah dua, yaitu: mengurangi bias dan belajar secara adaptif. Algoritma yang diusulkan untuk masalah klasifikasi *multiclass* dijelaskan dalam algoritma ADASYN [17].

Tahap ketiga adalah klasifikasi dengan metode *Support Vector Machine* (SVM) *multiclass*. SVM merupakan metode klasifikasi yang memiliki tingkat akurasi yang baik. Tujuan dari algoritma SVM adalah untuk menemukan *hyperplane* dalam ruang berdimensi-N (N — jumlah fitur) yang secara jelas mengklasifikasikan titik data [18].



Gambar 1. *Support Vector Machine* (SVM)

Untuk memisahkan dua kelas titik data tersebut, ada banyak kemungkinan hyperplane yang bisa dipilih. Tujuan kita adalah menemukan bidang yang memiliki margin maksimum, yaitu jarak maksimum antara titik data dari kedua kelas. Memaksimalkan jarak margin memberikan penguatan sehingga titik data di masa mendatang dapat diklasifikasikan dengan lebih baik[19].

Hyperplanes adalah batas keputusan yang membantu mengklasifikasikan titik data. Poin data yang jatuh di kedua sisi hyperplane dapat dikaitkan dengan kelas yang berbeda. Juga, dimensi hyperplane bergantung pada jumlah fitur. Jika jumlah fitur input adalah 2, maka hyperplane hanya berupa garis. Jika jumlah fitur masukan adalah 3, maka *hyperplane* menjadi bidang dua dimensi. Sulit membayangkan ketika jumlah fitur melebihi 3[20]. Didalam algoritma SVM, dilihat batas maksimal antara poin data dan *hyperplane*. *Los function* membantu memaksimumkan batas adalah *hinge loss*[21]. Bidang pembatas pertama menjadi batas kelas pertama sedangkan bidang pembatas kedua adalah batas dari kelas kedua, sehingga diperoleh Persamaan 1.

$$xi\,.w\,+\,b\,\geq\,+1\;foryi\,=\,+1 \qquad (1)$$
$$xi\,.w\,+\,b\,\leq -1\;foryi\,=\,-1$$

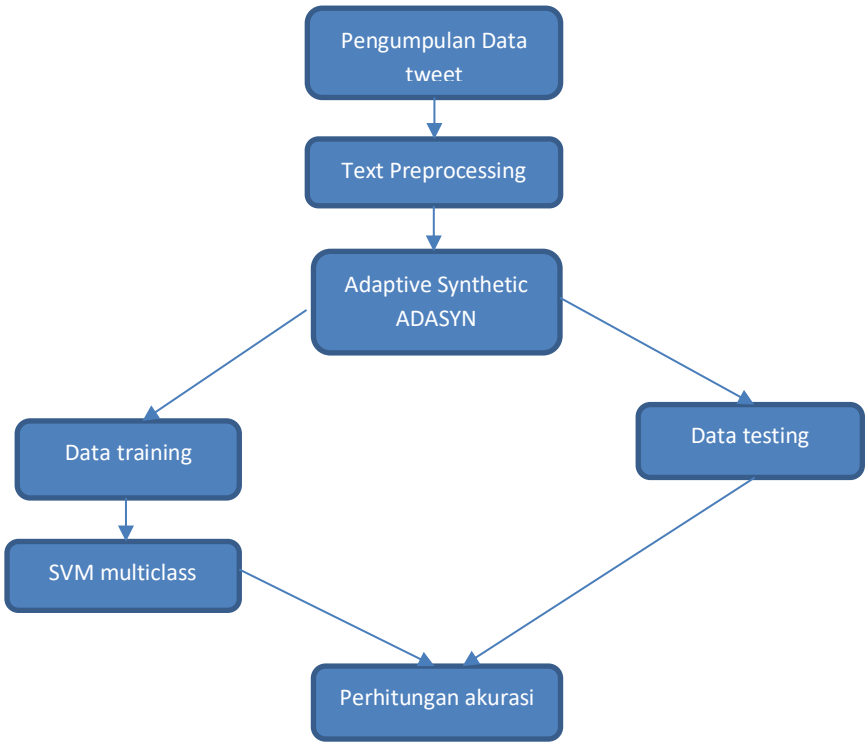
Keterangan:

w : Normal bidang

b : Posisi bidang relatif terhadap pusat koordinat

Secara umum, cara kerja dari SVM adalah menemukan jarak terjauh dari hyperplane dengan kedua kelas. Proses penentuan jarak terjauh dilakukan berulang kali hingga menemukan hyperplane terbaik. Untuk itulah diperlukan optimasi pada SVM untuk menemukan jarak maksimum hyperplane dengan kedua kelas tersebut. Dalam pembangunan SVM, terdapat dua bentuk optimasi yang digunakan untuk menemukan hyperplane. Bentuk optimasi pertama yaitu Primal Form SVM dan yang kedua adalah Dual Form SVM. Primal Form tidak dapat digunakan dalam penelitian ini karena tidak akan pernah memenuhi konstrain.

Langkah terakhir pada penelitian ini adalah menghitung akurasi pada setiap provinsi di Indonesia. Perhitungan akurasi digunakan untuk mengetahui seberapa baik metode ADASYN dan SVM dalam mengklasifikasian sentimen kenaikan BBM di seluruh provinsi di Indonesia. Berikut bagan penelitian disajikan pada Gambar 2.



Gambar 2. Diagram penelitian

III. HASIL DAN PEMBAHASAN

Pada bagian ini akan diperlihatkan gambaran komentar masyarakat Indonesia terhadap kenaikan Bahan Bakar Minyak (BBM) melalui bentuk *wordlcoud*. Sentimen masyarakat Indonesia terhadap kenaikan BBM dibagi menjadi tiga label yaitu positif (pro terhadap kenaikan BBM), netral (tidak

memihak), dan negative (kontra terhadap kenaikan BBM). Sentimen yang dihasilkan lebih banyakan kontra terhadap kenaikan BBM, sehingga mengakibatkan *imbalance* data. Pada penelitian ini menerapkan metode *Adaptive Synthetic* (ADASYN) untuk mengatasi permasalahan tersebut. Hasil ADASYN akan diklasifikasian menggunakan metode *Support Vector Machine* (SVM).

Pembahasan pertama pada penelitian dimulai dari visulasisasi sentimen masyarakat Indonesia menggunakan *wordcloud*. Visualisasi dilakukan untuk setiap provinsi di Indonesia. Semakin besar ukuran kata pada *worldcloud* maka topik tersebut sering dibicarakan oleh masyarakat. Berikut merupakan sampel hasil *worldcloud* setiap provinsi.

Gambar 3. Sampel *Wordcloud* Sentimen Kenaikan Harga BBM (a) Positif Provinsi Aceh (b) Netral Provinsi Aceh (c) Negatif Provinsi Aceh (d) Positif Provinsi Kalimantan Timur (e) Netral Provinsi Kalimantan Timur (f) Negatif Kalimantan Timur (g) Positif Provinsi Jawa Tengah (h) Netral Provinsi Jawa Tengah (i) Negatif Provinsi Jawa Tengah (j) Positif Provinsi DKI Jakarta (k) Netral Provinsi DKI Jakarta (l) Negatif Provinsi DKI Jakarta.

Berdasarkan Gambar 3 terlihat bahwa sentimen positif di setiap provinsi frekeunesinya hampir merata dan tidak ada kata yang menonjol mendukung kenaikan BBM. Kata-kata yang muncul pada sentimen positif lebih menekankan pada apa yang harus dilakukan oleh pemerintah seperti "subsidi" dan "sesuai harga". Hal yang sama terjadi pada sentimen netral frekunesi kata terbanyak adalah "harga", "turun", "naik". Sedangkan kata yang sering muncul pada sentimen negatif adalah "harga", "demo", "bakar", dan "naik". Kata tersebut menggambarkan luapan amarah dari masyarakat Indonesia terkait kenaikan BBM.

Untuk mengetahui lebih detail mengenai kata yang sering diucapkan oleh masyarakat terkait kenaikan BBM disajikan pada Tabel 2. Tabel 2 menyajikan frekuensi kemunculan kata terbanyak untuk setiap sentimen positif, netral, dan negatif di setiap provinsi akan digunakan untuk merepresentasikan kata-kata yang paling sering digunakan oleh masyarakat terkait dengan kenaikan

harga BBM. Kata-kata yang paling sering digunakan oleh masyarakat umum pada setiap kategori sentimen di setiap provinsi terdapat pada tabel di bawah ini.

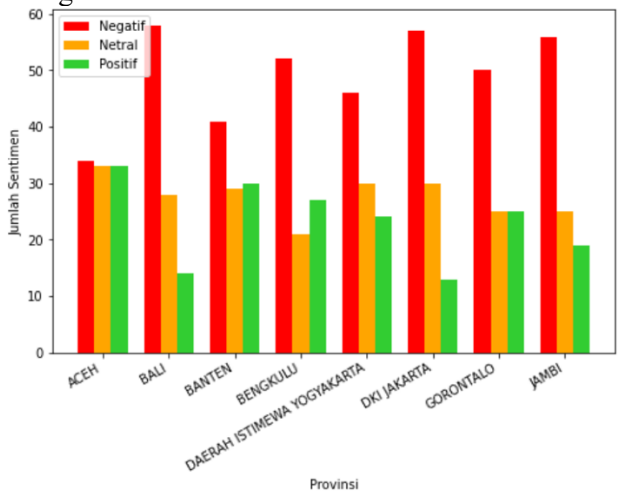
Tabel 2. Frekuensi Kata yang Sering Digunakan

No	Provinsi	Kata yang Sering Digunakan (%)		
		Negatif	Netral	Positif
1	Aceh	Harga (5.97 %)	Harga (9.71%)	Harga (14.72 %)
2	Bali	Harga (9.88 %)	Harga (9.42 %)	Harga (10.47 %)
3	Banten	Harga (13.64 %)	Harga (5.61 %)	Harga (13.92 %)
4	Bengkulu	Harga (11.11 %)	BLT (7.35%)	Harga (13.22 %)
5	DI Yogyakarta	Harga (6.71 %)	Harga (6.36 %)	Harga (14.16 %)
6	DKI Jakarta	Bakar (5.70 %)	Bakar (5.32 %)	Bakar (5.87 %)
7	Gorontalo	Harga (14.29 %)	Harga (11.97 %)	Harga (14.46 %)
8	Jambi	Harga (11.76 %)	Harga (10 %)	Harga (14.30 %)
9	Jawa Barat	Harga (7.19 %)	Harga (7.46 %)	Harga (11.32 %)
10	Jawa Tengah	Harga (8.01 %)	Harga (6.56 %)	Harga (12.88 %)
11	Jawa Timur	Harga (8.41 %)	Harga (7.41 %)	Harga (12.69 %)
12	Kalimantan Timur	Harga (10.26 %)	Sepakat (10.04 %)	Harga (12.62 %)
13	Kalimantan Selatan	Harga (10.29 %)	Harga (8.14 %)	Kalimantan (10.36 %)
14	Kalimantan Tengah	Naik (5.86 %)	Harga (5.02 %)	Harga (12.54 %)
15	Kalimantan Barat	Moga (14.29)	Salah (3.39 %)	Sesuai (9.77 %)
16	Kalimantan Utara	Harga (13.33 %)	Harga (11.11 %)	Harga (10.53 %)
17	Bangka Belitung	Harga (7.14 %)	Harga (8.22 %)	Pulau (13.61 %)
18	Riau	Blitar (9.21 %)	Bantu (10.00 %)	Masyarakat (23.08 %)
19	Lampung	Harga (9.42 %)	Harga (5.18 %)	Dukung (9.74 %)
20	Maluku	Naik (7.41 %)	Harga (15.00 %)	Harga (10.59 %)
21	Maluku Utara	Harga (8.47 %)	Utara (8.33 %)	Utara (13.24 %)
22	Nusa Tenggara Barat	Harga (10.17 %)	Harga (5.61 %)	Tenggara (5.39 %)
23	Nusa Tenggara Timur	Nusa (12.00 %)	Harga (5.99 %)	Dukung (9.11 %)
24	Papua	Turun (18.18 %)	Sa (7.53 %)	Masyarkat (5.84 %)
25	Papua Barat	Oktober (16.67 %)	Harga (9.93 %)	Perintah (7.59 %)
26	Riau	Harga (6.63 %)	Harga (5.40 %)	Harga (14.10 %)
27	Sulawesi Barat	Harga (11.70 %)	Harga (7.38 %)	Utara (8.56 %)
28	Sulawesi Selatan	Harga (10.22 %)	Harga (5.47 %)	Subsidi (5.00 %)
29	Sulawesi Tengah	Harga (14.29 %)	Harga (11.90 %)	Dukung (9.43 %)
30	Sulawesi Tenggara	Harga (10.28 %)	Bantu (5.84 %)	Subsidi (5.27 %)
31	Sulawesi Utara	Harga (10.34 %)	Harga (8.77 %)	Harga (15.53 %)
32	Sumatera Barat	Harga (11.59 %)	Harga (9.09 %)	Jambi (10.93 %)
33	Sumatera Selatan	Harga (8.96 %)	Bantu (4.08 %)	Dukung (8.23 %)
34	Sumatera Utara	Turun (15.38 %)	Harga (4.90 %)	Dukung (10.67 %)

Kata "harga" adalah kata yang paling sering digunakan pada kategori negatif, netral, dan positif. Pada kategori negatif, kata "harga" merupakan kata yang paling sering digunakan di 25 provinsi dari total 34 provinsi di Indonesia. Pada kategori netral, kata "harga" juga mendominasi kata yang paling banyak digunakan di 25 provinsi dari total 34 provinsi di Indonesia. Pada kategori positif, kata yang paling sering digunakan cukup beragam. Kata "harga" merupakan kata yang paling sering digunakan

di 16 provinsi, kata "dukung" merupakan kata yang sering digunakan di 5 provinsi, kata "subsidi" paling sering digunakan di 2 provinsi yaitu Sulawesi Utara dan Sulawesi Selatan.

Terdapat perbedaan frekuensi kata yang paling sering digunakan untuk setiap provinsi di Indonesia. Daerah yang cukup berbeda jika dibandingkan dengan daerah lain adalah provinsi DKI Jakarta, Kalimantan Timur, Kepulauan Riau, Maluku Utara, Nusa Tenggara Timur, Papua, Papua Barat, Sulawesi Tenggara, Sumatera Selatan, dan Sumatera Utara. Di daerah-daerah tersebut, kata yang paling sering digunakan pada setiap kategori cukup beragam dan tidak ada kata yang mendominasi untuk semua kategori.



Gambar 4. Jumlah Sentimen Per Label Per Provinsi

Rata-rata frekuensi sentimen negatif mendominasi setiap provinsi di Indonesia, hal ini menunjukan bahwa penolakan kenaikan BBM terjadi di berbagai provinsi di Indonesia. Ketidak seimbangan label pada analisis kasifikasi menimbulkan masalah serius (Gambar 4). Apabila terdapat data baru masuk dan diprediksi berdasarkan model terbentuk, data tersebut akan diarahkan ke label yang paling dominan. Metode ADASYN membangkitkan data sintetik untuk memnyeimbangkan frekuensi data. Hasil ADASYN akan diklasifkasikan menggunakan metode SVM. Berikut hasil klasifikasi menggunaman metode SVM *multiclass*.

Tabel 3. Akurasi Metode SVM *Multiclass* Per Provinsi

Provinsi	**Akurasi**	**Provinsi**	**Akurasi**
Aceh	0.70	Kepulauan Riau	0.90
Bali	0.90	Lampung	0.90
Bangka Belitung	0.90	Maluku	0.85
Banten	0.90	Maluku Utara	0.85
Bengkulu	0.80	NTB	0.90
DI Yogyakarta	0.95	NTT	0.85
DKI Jakarta	0.95	Papua	0.95
Gorontalo	0.85	Papua Barat	0.95
Jawa Barat	0.95	Riau	0.90
Jambi	0.90	Sulawesi Barat	0.90
Jawa Tengah	0.90	Sulawesi Selatan	0.85
Jawa Timur	0.90	Sulawesi Tengah	0.90
Kalimantan Barat	0.85	Sulawesi Tenggara	0.90
Kalimantan Selatan	0.90	Sulawesi Utara	0.75
Kalimantan Tengah	0.90	Sumatera Barat	0.90
Kalimantan Timur	0.80	Sumatera Selatan	0.85
Kalimantan Utara	0.85	Sumatera Utara	0.90

Berdasarkan Tabel 3 akurasi metode SVM *multiclass* terendah adalah provinsi Aceh sebesar 70%. Sedangkan Akurasi tertinggi terdapat pada provinsi DI Yogyakarta, DKI Jakarta, Jawa Barat, Papua, Papua Barat sebesar 95%. Rata-rata akurasi seluruh provinsi di Indonesia sebesar 87.94%. Hal ini

menunjukan bahwa metode ADASYN SVM *multiclass* mampu memberikan akurasi yang baik untuk pengklasifikasian data tidak seimbang (*imbalance data*).

IV. KESIMPULAN

Sentimen negatif terhadap kenaikan BBM di seluruh provinsi di Indonesia mendominasi tweet. Hal ini dibuktikan dengan frekuensi sentimen negatif lebih banyak dibandingkan dengan netral dan positif. Provinsi dengan sentimen negatif terbanyak adalah DKI Jakarta. Penerapan metode ADASYN mengatasi ketidaksamaan frekuensi antara sentimen negatif, positif, dan netral. Hasil penyamaan ADASYN dilakukan klasifikasi dengan menggunakan metode SVM multiclass menghasilkan rata-rata akurasi 87.94%. Akurasi tertinggi adalah provinsi DKI Jakarta, DI Yogyakarta, Jawa Barat, Papua, dan Papua Barat dengan akurasi mencapai 95%.

UCAPAN TERIMA KASIH

REFERENSI

1. Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.
2. D. Anjas Ramadhan and E. Budi Setiawan SSi, "ANALISIS SENTIMEN PROGRAM ACARA DI SCTV PADA TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN SUPPORT VECTOR MACHINE."
3. A. Ridok and R. Latifah, "Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN," in *Konferensi Nasional Sistem & Informatika*, 2015, pp. 1–6. [Online]. Available: www.kompas.com
4. H. N. Irmanda and R. Astriratma, "Klasifikasi Jenis Pantun dengan Metode Support Vector Machines (SVM)," *RESTI*, vol. 1, no. 3, pp. 915–922, 2017.
5. H. Guo and W. Wang, "An active learning-based SVM multi-class classification model," *Pattern Recognit*, vol. 48, no. 5, pp. 1577–1597, 2015, doi: https://doi.org/10.1016/j.patcog.2014.12.009.
6. V. Blanco, A. Japón, and J. Puerto, "Optimal arrangements of hyperplanes for SVM-based multiclass classification," *Adv Data Anal Classif*, vol. 14, no. 1, pp. 175–199, 2020, doi: 10.1007/s11634-019-00367-6.
7. C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans Neural Netw*, vol. 13, no. 2, pp. 415–425, 2002, doi: 10.1109/72.991427.
8. A. Ridok and R. Latifah, "Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN," 2015. [Online]. Available: www.kompas.com
9. S. Chaudhari, A. R, V. G. Tekkur, P. G. L, and S. R. Karki, "Ingredient/Recipe Algorithm using Web Mining and Web Scraping for Smart Chef," *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2020.
10. T. Karthikeyan, K. Sekaran, D. Ranjith, V. Vinoth kumar, and J. M. Balajee, "Personalized content extraction and text classification using effective web scraping techniques," *International Journal of Web Portals*, vol. 11, no. 2, pp. 41–52, Jul. 2019, doi: 10.4018/IJWP.2019070103.
11. A. Kurniasih and L. P. Manik, "On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, pp. 927–934, 2022, doi: 10.14569/IJACSA.2022.01306109.
12. A. F. Hidayatullah, S. Cahyaningtyas, and A. M. Hakim, "Sentiment Analysis on Twitter using Neural Network: Indonesian Presidential Election 2019 Dataset," *IOP Conf Ser Mater Sci Eng*, vol. 1077, no. 1, p. 012001, Feb. 2021, doi: 10.1088/1757-899x/1077/1/012001.
13. L. P. Manik *et al.*, "Aspect-Based Sentiment Analysis on Candidate Character Traits in Indonesian Presidential Election," in *Proceeding - 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications, ICRAMET 2020*, Institute of

Electrical and Electronics Engineers Inc., Nov. 2020, pp. 224–228. doi: 10.1109/ICRAMET51080.2020.9298595.
14. M. Mayo, "A General Approach to Preprocessing Text Data," Dec. 01, 2017.
15. A. A. V. A. Jayaweera, Y. N. Senanayake, and P. S. Haddela, "Dynamic Stopword Removal for Sinhala Language," 2019.
16. R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "Data mining with imbalanced class distributions: concepts and methods," 2009.
17. H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," 2008.
18. F. Fauzi, "K-Nearset Neighbor (K-NN) dan Support Vector Machine (SVM) untuk Klasifikasi Indeks Pembangunan Manusia Provinsi Jawa Tengah Info Artikel," *Jurnal MIPA*, vol. 40, no. 2, 2017, [Online]. Available: http://journal.unnes.ac.id/nju/index.php/JM
19. M. Y. Darsyah, I. J. Suprayitno, F. Fuzi, B. W. Otok, and B. S. S. Ulama, "Smooth Support Vector Machine (SSVM) for classification of human development index," in *Journal of Physics: Conference Series*, 2019. doi: 10.1088/1742-6596/1217/1/012114.
20. Z. F. Hussain *et al.*, "A new model for iris data set classification based on linear support vector machine parameter's optimization," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 1079–1084, 2020, doi: 10.11591/ijece.v10i1.pp1079-1084.
21. D. Agustina, E. Putri, F. Fauzi, S. N. Alawiyah, and R. Wasono, "METODE SUPPORT VECTOR MACHINE (SVM) UNTUK KLASIFIKASI DATA EKSPRESI GEN MICROARRAY," in *EDUSAINTEK 4*, 2020, pp. 1–10.