Penggunaan Metode GloVe untuk Ekspansi Fitur pada Analisis Sentimen Twitter dengan *Naïve Bayes* dan *Support Vector Machine*

1st Made Dwi Dharma Sreya
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
madedwidharmas@student.telkomuniversity.ac.id

2nd Erwin Budi Setiawan
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
erwinbudisetiawan@telkomuniversity.ac.id

Abstrak

Analisis Sentimen merupakan cara untuk mengetahui opini publik terhadap suatu informasi. Berbagai metode dapat digunakan untuk memperoleh akurasi serta meningkatkan akurasi informasi yang didapat dari Analisis Sentimen. Untuk mendapatkan akurasi yang maksimal ini dilakukan dengan 3 tahapan yaitu pencarian baseline, penggunaan *hyperparameter* dan TFIDF, dan penggunaan Korpus. Penggunaan korpus yang tidak efektif karena ketidaksesuaian kosakata dapat menurunkan akurasi analisis sentimen. Oleh karena itu digunakan Metode GloVe dalam ekspansi fitur pada Korpus untuk mengatasi ketidasesuaian kosakata pada data yang kita miliki. Selain menggunakan Metode GloVe sebagai cara untuk mengekspannsi fitur, *Support Vector Machine* serta *Naive Bayes* juga digunakan sebagai Metode Klasifikasi. Hasil yang didapat dari penelitian yaitu perbandingan akurasi sebelum dan sesudah melakukan ketiga tahapan tersebut. Peningkatan yang diperoleh adalah sebesar 44% dan 54% dengan akurasi yang sebelumnya sebesar 0.5394 dan 0.5406 meningkat menjadi 0.7786 dan 0.8323 untuk *Naïve Bayes* dan *Support Vector Machine*.

Kata kunci : analisis sentiment, feature expansion, GloVe. SVM, NB

Abstract

Sentiment analysis is a way to find out public opinion on an information. Various methods can be used to obtain accuracy and improve the accuracy of the information obtained from Sentiment Analysis. To get maximum accuracy, this is done in 3 stages, namely the baseline search, the use of hyperparameters and TFIDF, and the use of Corpus. The use of ineffective corpus due to inappropriate vocabulary can reduce the accuracy of sentiment analysis. Therefore, the GloVe method is used in the feature expansion in Corpus to overcome the incompatibility of vocabulary in the data we have. In addition to using the GloVe Method as a way to expand features, Support Vector Machine and Naive Bayes are also used as Classification Methods. The results obtained from the study are comparisons of accuracy before and after doing the three stages. The improvements obtained were 44% and 54% with the previous accuracy of 0.5394 and 0.5406 increasing to 0.7786 and 0.8323 for Naïve Bayes and Support Vector Machines.

Keywords: sentiment analysis, feature expansion, GloVe, SVM, NB

I. PENDAHULUAN

Salah satu media sosial yang ramai digunakan di Indonesia adalah Twitter. Sekalipun banyak kemunculan media-media sosial lainnya, twitter dinilai stabil dan memiliki peran yang signifkan sebagai media interaksi dan komunikasi di masyarakat. Berdasarkan data yang dirilis pihak Twitter Indonesia pada tahun 2016, 77% pengguna twitter di Indonesia merupakan pengguna twitter aktif[1]. Ini artinya penyebaran informasi melalui media twitter sangat mudah terjadi. Tentu saja penyebaran informasi yang sangat banyak berarti munculnya respon-respon dari masyarakat terhadap informasi yang diterima. Baik itu respon positif atau negatif.

Tweet hasil respon dari masyarakat ini merupakan hal menarik untuk dianalisis. Hal tersebut karena tweet yang disampaikan dapat mengandung informasi tentang pandangan, pemikiran, budaya, dan juga kebiasaan yang dimiliki oleh masyarakat pada suatu rentangan tertentu [2].Dengan mempelajari informasi dan pesan yang diunggah melalui media twitter, pemahaman terhadap kondisi masyarakat dapat dipelajari secara mendalam.

Analisis Sentimen atau dikenal dengan Penggalian Opini merupakan sistem untuk merangkum semua opini masyarakat dan mengelompokannya menjadi hal yang berguna secara

otomatis. Dengan dikelompokannya opini tersebut, kita dapat mengetahui bagaimana pandangan masyarakat terhadap suatu produk atau informasi. Penggunaan analisis sentimen ini sangat besar pengaruhnya hingga 20-30 perusahaan menggunakan sistem ini di Amerika [3].

Pada penelitian analisis sentimen yang dilakukan oleh E. Setiawan, dkk. dengan dengan menggabungkan basis fitur dengan ekspansi fitur telah terbukti bahwa nilai akurasi yang diperoleh dari SVM, Logit, dan NB adalah meningkat. Nilai akurasi tertinggi diperoleh ketika menggunakan Logit sebesar 98.81%[4]. Kemudian pada penelitian ekspansi fitur dengan menggunakan Word2Vec oleh E. Setiawan, dkk ditemukan bahwa menggunakan ekspansi fitur pada SVM dapat mengurangi performa dari sistem. Sementara itu menggunakan ekspansi fitur pada Logit dapat mengingkatkan performa secara konsisten dan performa campuran didapat ketika dilakukan pada NB[5]

Berdasarkan penelitian diatas, maka motivasi pada penelitian ini adalah melakukan penelitian dengan sistem yang menggunakan metode yang berbeda yaitu GloVe untuk mengekspansi fitur dari Analisis Sentimen Twitter pada model SVM dan Naïve Bayes

Masalah yang dibahas dalam Tugas Akhir ini adalah bagaimana pengaruh dan tingkat performansi sistem setelah diterapkan teknik ekspansi fitur dengan metode GloVe pada algoritma Support Vector Machine (SVM) dan Naïve Bayes (NB).

Batasan penelitian dalam Tugas Akhir ini, yaitu data yang digunakan adalah data sentimen Bahasa Indonesia sebanyak 16.597 *tweet* yang bertopik kebijakan publik di Indonesia, proses pelabelan sentimen dilakukan secara manual menjadi dua kategori, yaitu positif dan negatif, nilai matriks performansi yang digunakan adalah nilai akurasi, serta *word embedding* yang digunakan adalah GloVe

Tujuan yang ingin dicapai dari penelitian ini adalah mengimplementasi, mengukur nilai performansi pada nilai akurasi, serta menganalisis hasil sistem klasifikasi sentimen yang telah dibangun menggunakan teknik ekspansi fitur dengan metode GloVe pada data sentimen Bahasa Indonesia dalam *tweet* yang telah dikumpulkan

Tugas Akhir ini disusun dengan struktur yang pertaamn membahas teori/studi/literatur yang mendukung atau berkaitan erat dengan penelitian ini. Kemudian membahas teori terkait penelitian dan pemodelan sistem yang dibangun. Selanjutnya menjelaskan hasil, analisis, dan evaluasi model penelitian.Terakhir membahas kesimpulan dan saran untuk penelitian selanjutnya.

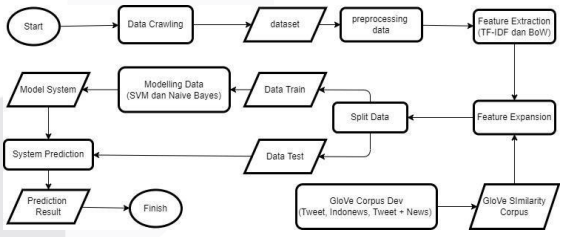
II. KAJIAN TEORI

Penelitian mengenai sentimen analisis bukanlah hal yang baru. Apalagi penelitian mengenai ekspansi fitur pada sentimen analisis. Berbagai variasi mengenai ekspansi fitur pun banyak dilakukan. Misalkan penelitian mengenai ekspansi fitur menggunakan Word Embedding. Penlitian telah dilakukan oleh E. Setiawan, dkk. sebelumnya. Pada penelitiannya, word embedding yang menggunakan Word2Vec ini berhasil diimplementasi pada ekspansi fitur de- ngan dataset dari IndoNews dan Google News. Implementasi ekspansi fitur pada dataset Google News, meningkatkan performa pada Logistic Regression, campuran pada Naive Bayes dan menurunkan performa pada Support Vector Machine. Hasil performa menggunakan dataset Google News lebih baik diban- ding menggunakan dataset Indo News[4]

Selain ekspansi fitur menggunakan word embedding, penelitian mengenai Ekspansi Fitur pada Sentimen Analisis Twitter juga dilakukan oleh E. Setiawan, dkk. Penelitian kali ini befokus pada perbandingan performa menggu- nakan ekspansi fitur dengan TF-IDF dan ekspansi fitur dengan tweet-based fitur. Sama seperti penelitian sebelumnya, model klasifikasi yang digunakan juga menggunakan Logit, NB dan SVM. Pada penelitian ini diketahui bahwa ekspansi fitur terbukti meningkatkan akurasi sentiment analisis dengan peningkatan tertinggi terjadi pada model Logit dengan akurasi 98.81%[5]

III. METODE

Sistem analisis sentimen dibangun sesuai Gambar 1. Sistem dibagi menjadi 5 bagian seperti Data Crawling, Preprocessing Data, Representasi data, Ekspansi Fitur, dan Memodelkan Analisis Sentimen. Masing-masing bagian akan mewakili beberapa proses dari Gambar 1. Berikut merupakan penjelasan dari setiap tahap.



GAMBAR 1 Sistem Analisis Sentimen dengan Menggunakan Metode Glove

A. Crawling dan Pelabelan Data

Crawling data dilakukan dengan menggunakan Application Program Interface (API) Twitter yang sudah disediakan oleh Twitter [6]. Penulis menggunakan kata kunci yang dirasa memiliki hubungan dengan peristiwa yang sedang viral dan data dikumpulkan selama bulan Oktober. Kata kunci yang digunakan seperti: *#omnibuslaw*, *#covid19* dan *#vaksin*. Data yang didapat kemudian dilabeli dengan 1 sebagai tweet positif dan -1 sebagai tweet negatif.

Data yang telah dilabeli kemudian dilakukan validasi label dengan bantuan 3 orang lain untuk memvoting, suara terbanyaklah yang diambil sebagai label yang sesuai. Tabel 1 meerupakan contoh data tweet yang dilabeli.

TABEL 1 Pelabelan Data

Tweet	Label
“banyak manfaat dari undang undang cipta kerja yang baru disahkan ini lah gaes, mantap”	1
“klu gagal uu omnibuslaw kepung istana amp gedung dprri tumbang rezim khianat revolusi harga mati khianat rakyat revolusi”	-1

B. Preprocessing Data

Data yang didapat dari Twitter mengandung banyak informasi yang susah untuk diolah pada sistem. Maka dari itu data perlu diolah terlebih dahulu menjadi data yang sesuai sehingga dapat digunakan secara efektif. Pada tahap ini data diproses melalui beberapa tahap yang akan dijelaskan sebagai berikut

a. Case Folding

Pada tahap ini, tiap huruf pada tweet akan diubah menjadi huruf kecil semua. Hal ini bertujuan untuk menghindari masalah karena adanya perbedaan besar kecil dari huruf

b. Tokenization

Tahap Tokenization merupakan tahap yang memecah tiap kata pada tweet yang dipisah oleh spasi. Bertujuan dalam mendapatkan kata yang sering muncul pada suatu topik. Misalkan terdapat tweet seperti ”uu cipta kerja tingkatkan e-commerce jadi bisnis potensial”. Tweet ini akan dipecah sehingga menjadi ”uu”, ”cipta”, ”kerja”, ”tingkatkan”, ”e-commerce”, ”jadi”, ”bisnis”, dan ”potensial”.

c. Stopword Removal

Stopword Removal atau menghapus stopword, merupakan tahap meng-hilangkan kata-kata yang tidak perlu pada kamus. Kata-kata yang tidak perlu seperti ”dan”, ”jadi”, ”atau”, ”untuk” dan kata-kata yang tidak ada hubungannya dengan topik yang akan kita proses akan dihapus dari kamus.

d. Stemming

Tahap terakhir yaitu Stemming. Pada tahap ini tiap kata akan diubah menjadi kata dasarnya dengan menghilangkan imbuhan yang ada.

C. Term Frequency – Inverse Document Frequency (TF-IDF)

Dalam merepresentasikan data, tiap kata pada tweet akan diberi nilai Boolean. Nilai boolean ini mencerminkan ada atau tidaknya kata tersebut pada suatu tweet. Tiap tweet diasumsikan memiliki sejumlah kata yang dapat merepresentasikan tweet tersebut. Mencari kata representasi itu perlu menghitung nilai bobot dari kata-kata yang ada.

Term Frequency – Inverse Document Frequency atau disingkat TF-IDF merupakan pembobotan dalam menggambarkan data dalam suatu model vector ruang[7]. Nilai bobot yang berhasil dihitung ini akan digunakan dalam mendapatkan fitur pada suatu topik. Representasi dari suatu data bergantung dari jumlah fitur yang didapat pada seluruh topik. Perhitungan TF-IDF dilakukan dengan rumus:

$$Wij = t_{fij} \times Idf_j, Idf_j = \log(\frac{N}{df}) \quad (1)$$

D. Global Vector (GloVe)

Global Vector atau GloVe merupakan model yang dapat menyimpan statistik kemunculan kata secara global yang nantinya model ini akan digunakan untuk merepresentasikan kata atau makna. Model Glove ini dapat tercipta karena dari pengamatan yang dilakukan diketahui bahwa rasio dari kemunculan kata-kata memiliki potensi untuk ditarik suatu kesimpulan.Hasil yang diperoleh dari GloVe merupakan besar kedekatan suatu kosakata terhadap kosakata lain. Nilai inilah yang akan digunakan untuk mengekspansi fitur dari sentimen analisis. Tabel 2 merupakan kedekatan suatu kata dari korpus yang telah dibuat menggunakan GloVe

TABLE 2 Korpus GloVe

Contoh Kata	Top Similarity	Korpus Tweet	Korpus IndoNews	Korpus IndoNews + Tweet
Kampanye	1	('zalim', 0.9133729740400165)	('pilkada', 0.6396626233549042)	('pilkada', 0.6457906848274794)

E. Ekspansi Fitur

Setelah mendapatkan korpus dari GloVe, fitur-fitur yang kita miliki dapat diperluas pengertiannya dengan ekspansi fitur. Ekspansi fitur dilakukan dengan mengganti vektor-vektor yang bernilai 0 pada data dengan *similar word* pada korpus GloVe.

Algoritma ekspansi fitur yang menggunakan input berupa list dari data vektor yang akan dicari kedekatannya dengan korpus GloVe. Algoritma 1 merupakan contoh algoritma untuk ekspansi fitur yang menggunakan top 1 similarity dalam mencari bobotnya.

Algoritma 1 *ekspansiFitur(teksVektor, korpusGloVe)* → matriks teks Vektor
// *implementasi ekspansi fitur pada setiap teks pada data*

```
Input  : teksVektor merupakan list vektor dari dataset
         korpusGlove merupakan list vektor dari glove
1      : for teks ∈ teksVektor do
2      :     temp ← teks.copy()
3      :     i ← korpusGlove.get_similar(temp)
4      :     if (temp.weight = 0) and (i.weight ≠ 0)
5      :         teks.weight = i.weight
6      :     endif
7      : endfor
```

Contoh penggunaan algoritma 1, misal terdapat tweet “Waktu kampanye mengemis2 giliran jadi wakil kalian Bengis” dan bobot vektor untuk kata “Kampanye” adalah 0. Maka akan dicari kedekatan dari kata “kampanye” pada korpus GloVe yang mana dapat dilihat pada Tabel 2. Setelah itu, didapatkan bahwa kata “kampanye” memiliki kedekatan dengan kata “zalim” maka bobot “kampanye” akan diisi dengan bobot “zalim” .

F. Naïve Bayes

Naive Bayes atau disingkat NB merupakan model klasifikasi yang menggunakan Teorema Bayes sebagai dasar dalam perancangannya. NB menggunakan teori peluang dalam mengklasifikasikan model. Karena simpelnya model ini dalam mengklasifikasikan data, model mudah dalam diimplementasi. Model ini juga bekerja dengan baik dalam memprediksi model dengan banyak kelas [6]. Hal inilah alasan kenapa model Naive Bayes digunakan.

Teorema Bayes:

$$P(Ci|X) = \frac{P(X|Ci) \times P(Ci)}{P(X)} \quad (2)$$

Keterangan
$P(Ci|X)$: peluang kemunculan Ci dengan kondisi X
X : kejadian X
Ci : kelas yang tersedia (C1 C2, C3, … Ci)
$P(Ci)$: peluang kemunculan Ci
$P(X)$: peluang kemunculan X
$P(X|Ci)$: peluang kemunculan X dengan kondisi Ci

G. Support Vector Machine

Support Vector Machine atau disingkat SVM merupakan model yang dikenal karena keakuratannya dalam mengklasifikasikan data [9]. Model ini sering digunakan untuk menyelesaikan masalah yang berkaitan dengan pembagian kelas pada bidang hiper atau data yang memiliki banyak fitur. Proses pembagian kelas dilakukan dengan membuat garis untuk memisahkan data, garis akan terus dibuat sehingga data dapat terbagi menjadi beberapa kelas. Metode SVM ini dapat diimplementasikan untuk membagi data sesuai dengan label sehingga data terpisah menjadi 2 kelas, yaitu kelas sentimen positif dan sentimen negatif

IV. HASIL DAN PEMBAHASAN

Hasil yang diperoleh dari pengujian sistem akan dijelaskan pada bagian ini.

A. Data

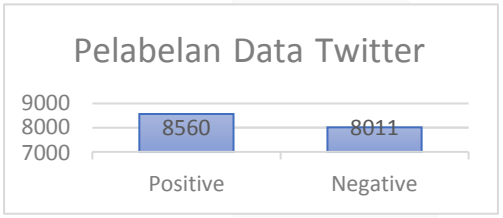
Data yg telah diperoleh dari *crawling data* sebanyak 16.597 tweet ini merupakan data yang akan digunakan untuk pada. Data diperoleh dengan menggunakan hashtag *#omnibuslaw*, *#covid19* dan

#vaksin. Tabel 3 menunjukan persebaran data dari hashtag yang digunakan.

TABEL 3 Persebaran Data Twitter

Hashtag	Jumlah
#omnibuslaw	4861
#covid19	5916
#vaksin	5820
Total	**16.597**

Data hasil crawling kemudian dilakukan pelabelan. Sebelum dilakukan pelabelan data dicek apabila ada data yang duplikat. Data yang duplikat sebanyak 26 data dihapus sehingga didapat data yang siap dilabeli sebanyak 16.571. Munculnya data yang duplikat terjadi akibat crawling data yang dilakukan oleh orang yang berbeda, sehingga ketika data dikumpulkan ada kemungkinan terjadinya duplikasi data. Setelah data yang duplikat dihapus, data dilabeli dan dilakukan validasi secara manual. Dari pelabelan diperoleh data tweet dengan sentimen positif sebanyak 8560 tweet dan sentimen negatif sebanyak 8011. Gambar 2 menunjukan hasil pelabelan dari data tweet.



GAMBAR 2 Pelabelan Data Twitter

Selanjutnya akan digunakan data *IndoNews* untuk pembuatan kamus, data ini diambil dari beberapa media *IndoNews* seperti Kompas, Tempo, Detik, Dan Yang Lainnya Dengan Topik Yang Berbeda Beda Seperti Agama, Bisnis, Budaya, Ekonomi, Entertainment, Hankam, Hukum, Iklan, Jurnalistik, Kesehatan, Keuangan, Motivasi, Olahraga, Pemerintahan, Pendidikan, Perhubungan, Politik, Sosial, Teknologi, Dan Umum, ada sebanyak 142.551, dan untuk pengambilannya dari tanggal 01 Mei 2016 hingga tanggal 01 Maret 2017 berikut adalah data *IndoNews* yang digunakan untuk pembuatan kamus katanya bisa dilihat pada data tabel 4.

TABEL 4 Persebaran Data *IndoNews*

Nama Redaksi	Jumlah
CNN Indonesia	29350
Detik	7975
Kompas	15056
Liputan6	252
Republika	53813
SIndoNews	22402
Tempo	13703
Total	**142.551**

B. Preprocessing Data

Pada tahap ini, data akan diolah sebelum digunakan untuk pengujian. Data-data tersebut akan dihapus atau diganti bagian-bagian yang tidak diperlukan, seperti simbol, angka, dan singkatan-singkatan yang ada. Hal ini dilakukan untuk meningkatkan kualitas data saat sebelum masuk tahap pelatihan model [10]. Tabel 5 merupakan langkah-langkah saat melakukan preprocessing data.

TABLE 5 Preprocessing Data

Step	Input	Output
Case Folding	UU Cipta Kerja tingkatkan E-Commerce jadi bisnis potensial	uu cipta kerja tingkatkan e-commerce jadi bisnis potensial
Tokenization	uu cipta kerja tingkatkan e-commerce jadi bisnis potensial	”uu”, ”cipta”, ”kerja”, ”tingkatkan”, ”e-commerce”, ”jadi”, ”bisnis”, ”potensial”.
Stopword Removal	”uu”, ”cipta”, ”kerja”, ”tingkatkan”, ”e-commerce”, ”jadi”, ”bisnis”, ”potensial”.	”uu”, ”cipta”, ”kerja”, ”tingkatkan”, ”e-commerce”, ”bisnis”, ”potensial”.
Stemming	”uu”, ”cipta”, ”kerja”, ”tingkatkan”, ”e-commerce”, ”bisnis”, ”potensial”.	”uu”, ”cipta”, ”kerja”, ”tingkat”, ”e-commerce”, ”bisnis”, ”potensial”.

C. Pembuatan Kamus Kata dengan GloVe (Corpus)

Pembuatan kamus kata dilakukan dengan menggunakan word embedding dengan metode GloVe. Kamus kata akan terbentuk dengan mengumpulkan kata dengan kata lain yang memiliki

kemiripan yang sama. Kemiripan dari tiap kata akan dihitung menggunakan metode GloVe. Berikut merupakan hasil similarity (kemiripan) pada corpus.

a. Korpus Data Tweet
Korpus data tweet merupakan korpus yang didapat dengan menggunakan data tweet dengan kosakata 15.952 kata. Tabel 6 merupakan hasil *similarity* dari salah satu kata yg digunakan dari korpus.

TABEL 6 Korpus Data Tweet

Kata	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
uu	jerat	ite	cipta ker	cilaka	poin
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	cipk er	prod uk	miner ba	perekono mian	ngang gur

b. Korpus Data Berita (indonews)
Korpus data tweet merupakan korpus yang didapat dengan menggunakan data dari Indonews dengan kosakata sejumlah 225.932 kata. Tabel 7 merupakan hasil *similarity* dari salah satu kata yg digunakan dari korpus.

TABEL 7 Corpus Data Berita

Kata	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
kampanye	pilka da	anie s	sosialis asi	kandi dat	pasl on
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	pons en	giat	debat	sandia ga	ahok

c. Korpus Data Berita dan Tweet
Korpus ini merupakan korpus yang menggabungkan data dari Indonews dengan data tweet sehingga menghasilkan kosakata sejumlah 230.042 kata. Tabel 8 merupakan hasil *similarity* dari salah satu kata yg digunakan dari korpus.

TABEL 8 Corpus Data Berita dan Tweet

Kata	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
tugas	emban	aparat	laksana	polisi	plt
	Rank 6	**Rank 7**	**Rank 8**	**Rank 9**	**Rank 10**
	mesti	awas	tanggung	piket	satu

D. Evaluasi dan Hasil Pengujian

Skenario yang dilakukan terdiri dari 3 skenario yang setiap skenarionya menggunakan *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB) sebagai model klasifikasinya. Skenario pertama membahas pembuatan baseline, skenario kedua membahas penggunaan *hyperparameter* dan TFIDF untuk meningkatkan akurasi, dan skenario terakhir membahas penggunaan korpus Tweet, korpus Indonews, dan korpus Tweet+Indonews yang merupakan hasil ekspansi fitur menggunakan GloVe.

a. Skenario Baseline

Skenario awal dalam pengujian yaitu membuat baseline dari masing-masing model dengan berbagai macam perbandingan data tes dan data latih. Pengujian dilakukan untuk mendapatkan perbanndingan yang menghasilkan akurasi yang maksimal. Penggunaan ratio 20:80 memperoleh data baseline tertinggi dengan akurasi 54.06% dan 53.94% untuk SVM dan NB. Tabel 9 menunjukan hasil akurasi dari baseline dengan beragam perbandingan data tes dan data latih

TABEL 9 Data Baseline

Ratio	SVM	NB
10:90	52.83	53.08
20:80	**54.06** ***(+1.22)***	**53.94** ***(+0.86)***
30:70	53.07 *(-0.98)*	52.94 *(-1.00)*
40:60	53.64 *(+0.57)*	53.69 *(+0.75)*
50:50	52.45 *(-1.19)*	52.55 *(-1.14)*

b. Skenario *Hyperparameter* dan TFIDF

Skenario selanjutnya adalah penggunaan *hyperparameter* dan TFID. Setelah didapat perbandingan dengan akurasi yang maksimal, akurasi

baseline kemudian ditingkatkan. Akurasi ditingkatkan dengan menggunakan *hyperparameter* pada baseline saja dan *hyperparameter* pada baseline dengan TFIDF. Tabel 10 menunjukan peningkatan akurasi setelah dilakukan optimasi dengan menggunakan *hyperparameter* dan TFIDF.

TABEL 10 Data *Hyperparameter* dan TFIDF

Classifier	**SVM**	**NB**
Baseline	54.06	53.94
Baseline (*Hyperparameter*)	80.54 *(+26.49)*	77.59 *(+23.65)*
Baseline + *TF-IDF*	**81.54 *(+1.00)***	***77.77 (+0.18)***

c. Skenario GloVe

Skenario selanjutnya adalah mencoba meningkatkan akurasi dengan menggunakan korpus Tweet, korpus Indonews, dan korpus Indonews+Tweet. Korpus-korpus ini diperoleh dari hasil ekspansi fitur-fitur yang ada pada dataset Tweet dan Indonews menggunakan metode GloVe. Selain dengan penggunaan korpus, akurasi juga di cek perubahannya dengan mengatur jumlah top similaritynya. Tabel 11 menunjukan hasil akurasi yang diperoleh dengan mengatur korpus yang digunakan dan top similaritynya

TABEL 11 Data dengan Korpus

	SVM			**NB**		
	Korpus *Tweet*	**Korpus *IndoNews***	**Korpus *IndoNews+Tweet***	**Korpus *Tweet***	**Korpus *IndoNews***	**Korpus *IndoNews+Tweet***
Top 1	81.75 *(+0.25)*	81.39 *(-0.18)*	81.24 *(-0.37)*	77.44 *(-0.42)*	77.35 *(-0.54)*	76.92 *(-1.09)*
Top 5	81.03 *(-0.63)*	81.27 *(-0.33)*	**83.23 *(+2.02)***	77.13 *(-0.82)*	77.83 *(+0.07)*	**77.86 *(+0.11)***
Top 10	81.90 *(+0.44)*	80.54 *(-1.23)*	82.20 *(+0.81)*	77.56 *(-0.27)*	76.05 *(-2.26)*	77.47 *(-0.38)*

E. Analisis Hasil Pengujian

Dari hasil pengujian yang telah dilakukan, peningkatan akurasi terbesar terjadi saat melakukan hyperparameter dengan peningkatan sebesar 48.9% untuk SVM dan 43.8% untuk NB. Lalu peningkatan akurasi terjadi lagi saat data ditambahkan dengan TF-IDF dengan nilai sebesar 1,2% pada SVM dan 0.2% pada NB. Kemudian nilai akurasi mulai berubah-ubah ketika dilakukan fitur ekspansi dengan corpus yang berbeda-beda dengan nilai top similarity yang berbeda pula. Namun peningkatan akurasi terbesar terjadi pada Top 5 similarity dengan nilai 2.02% pada SVM dan 0.11% pada NB dengan Corpus Indonews+Tweet.

V. KESIMPULAN

Pada penelitian ini telah dilakukan penelitian untuk analisis sentimen menggunakan teknik Ekspansi Fitur dengan metode GloVe pada model *Support Vector Machine* (SVM) dan *Naïve Bayes* (NB). Metode GloVe berhasil diimplementasikan sehingga menghasilkan 3 korpus yang digunakan saat ekspansi fitur. Korpus yang digunakan yaitu korpus tweet, korpus Indonews, serta korpus gabungan keduanya. Dari korpus tersebut diperoleh peningkatan performa akurasi pada model SVM dan NB dari model baselinenya. Peningkatan performa terbaik diperoleh pada Top 5 similarity dengan menggunakan korpus Indonews+Tweet dengan akurasi 83.23% untuk SVM dan 77.86% untuk NB. Akurasi meningkat sebesar 48.9% dan 43.8% dari akurasi awal.

REFERENSI

[1] Herman, "Indonesia Masuk Lima Besar Pengguna Twitter," *https://www.beritasatu.com/*, 2017. http://www.beritasatu.com/digital-life/428591-indonesia-masuk- limabesar-pengguna-twitter.html,.

[2] P. Fornacciari, M. Mordonini, and M. Tomauiolo, "Social network and sentiment analysis on twitter: Towards a combined approach," *CEUR Workshop Proc.*, vol. 1489, no. January 2018, pp. 53–64, 2015.

[3] B. Liu, "Sentiment analysis and subjectivity," *Handb. Nat. Lang. Process. Second Ed.*, pp. 627–666, 2010.

[4] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion for sentiment

analysis in twitter," *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 2018-Octob, pp. 509–513, 2018, doi: 10.1109/EECSI.2018.8752851.

[5] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," *Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol.*, no. October, 2017, doi: 10.1109/TSSA.2016.7871085.

[6] Y. Wategulis Syaifudin and D. Puspitasari, "Twitter Data Mining for Sentiment Analysis on Peoples Feedback Against Government Public Policy," *MATTER Int. J. Sci. Technol.*, vol. 3, no. 1, pp. 110–122, 2017, doi: 10.20319/mijst.2017.31.110122.

[7] W. W, W. Priatna, and J. Hidayat, "A IMPLEMENTASI TERM FREQUENCY – INVERSE DOCUMENT FREQUENCY (TF-IDF) DAN VECTOR SPACE MODEL (VSM) UNTUK PENCARIAN BERITA BAHASA INDONESIA," *Pelita Teknol.*, vol. 14, no. 2, pp. 119–133, 2019, [Online]. Available: https://jurnal.pelitabangsa.ac.id/index.php/pelitatekno/article/view/237.

[8] C. D. M. Jeffrey Pennington, Richard Socher, "GloVe: Global Vectors for Word Representation." https://nlp.stanford.edu/projects/glove/.

[9] D. K. Srivastava and L. Bhambhu, "Data classification using support vector machine," *J. Theor. Appl. Inf. Technol.*, vol. 12, no. 1, pp. 1–7, 2010.

[10] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019, doi: 10.22146/jnteti.v8i4.533.