Ekspansi Fitur dengan *FastText* pada Klasisikadi Topik dengan Metode *Naïve Bayes-Support Vector Machine* (NBSVM) di Twitter

1st Kintari Nurul Utami
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia
kintarinrl@student.telkomuniversity.ac.id

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

Abstrak

Banyak informasi yang tersebar di berbagai jejaring sosial *online*, termasuk Twitter. Di Twitter, Banyak orang yang berbagi informasi di sekitarnya. Akan tetapi, Twitter hanya dapat mengirim *tweet* sebanyak 280 karakter. Oleh karena itu, banyak pengguna yang melakukan pemotongan/penyingkatan kata dan juga penggunaan variasi kata pada *tweet* agar pada setiap *tweet* mencakup banyak informasi. Penggunaan variasi kata seperti *emoticon*, bahasa gaul, dan singkatan pada *tweet* membuat ketidakcocokan kosa kata dan kalimat yang disampaikan sehingga sulit untuk dimengerti. Dalam hal ini, penulis melakukan ekspansi fitur untuk kasifikasi topik di Twitter agar dapat mengatasi permasalahan tersebut. Metode yang digunakan untuk ekspansi fitur adalah *FastText* dan metode yang digunakan dalam klasifikasi adalah *Naïve Bayes-Support Vector Machine* (NBSVM). Hasil dari penelitian ini menunjukkan bahwa sistem klasifikasi topik dengan ekspansi fitur *FastText* menggunakan metode NBSVM memiliki akurasi sebesar 82.01%.

Kata kunci : *klasifikasi, NBSVM, fasttext, ekspansi fitur*

Abstract

A lot of information is scattered in various online social networks, including twitter. On twitter, many people share information around them. However, Twitter can only send tweets of up to 280 characters. Therefore, many users cut/shorten words and also use variations of words in tweets so that one tweet includes a lot of information. The use of word variations such as emoticons, slang, and abbreviations in tweets makes the mismatch of words and sentences that are conveyed so difficult to find difficult. In this case, the author expands the feature to classify topics on Twitter in order to overcome these problems. The method used for expansion is FastText and the method used in classification is Nave Bayes-Support Vector Machine (NBSVM). The results of this study indicate that the topic classification system with FastText feature expansion using the NBSVM method has an accuracy of 82.01%.

Keywords: classification, NBSVM, fasttext, feature expansion

I. PENDAHULUAN

Twitter telah menjadi salah satu portal informasi terbesar yang menyediakan *platform* dengan kemudahan dan terpercaya bagi pengguna untuk berbagi apa pun yang terjadi di sekitar mereka dengan teman dan pengikut lainnya[1]. Dengan menggunakan Twitter, setiap individu dapat mengetahui berbagai informasi dan berpotensi menjadi sumber informasi bagi banyak orang[2]. Twitter merupakan layanan *microblogging* yang memiliki batasan penyampaian kalimat sebesar 280 karakter dalam satu kali *tweet*. Oleh karena itu, seringkali pengguna Twitter melakukan pemotongan/penyingkatan kata dalam setiap kali membuat *tweet*. Hal ini berakibat pada ketidakcocokan kosakata serta *tweet* yang disampaikan juga sulit untuk dipahami dan diklasifikasikan.

Kesulitan dalam mengklasifikasikan *tweet* ini dapat diatasi dengan penggunaan ekspansi fitur[3]. Pada penelitian[4], terbukti bahwa ekspansi fitur dapat meningkatkan akurasi dari proses klasifikasi. Terdapat berbagai variasi metode untuk ekspansi fitur, Salah satunya yaitu *FastText* yang mampu memberikan *word embedding* untuk kata yang terdapat kesalahan eja, kata langka dan juga kata-kata yang tidak ada dalam *Corpus* data latih karena *FastText* merepresentasikan setiap kata sebagai karakter n-gram[6].

Pada Penelitian ini, penulis menggunakan algoritma *Naïve Bayes-Support Vector Machine* (NBSVM). NBSVM banyak digunakan sebagai dasar untuk metode lain dalam klasifikasi teks[18]. Namun, kinerjanya sangat bervariasi tergantung pada varian, fitur, dan kumpulan data yang digunakan. NBSVM memiliki kinerja yang baik pada teks berukuran kecil serta memiliki kinerja yang baik juga pada dokumen yang lebih panjang untuk kasus klasifikasi topik dan hasilnya seringkali lebih baik daripada publikasi yang sebelumnya. Metode NBSVM juga dinilai sebagai metode canggih yang memiliki tujuan untuk menangani sekumpulan fitur[18].

Berdasarkan uraian di atas, Penelitian ini dilakukan untuk klasifikasi kelas topik dengan Ekspansi Fitur *FastText* menggunakan metode NBSVM dan mengukur tingkat akurasi dari sistem yang telah dibuat. Adapun Batasan masalah dari penelitian ini yaitu dataset yang digunakan hanya dataset yang berbahasa Indonesia yang didapatkan melalui *crawling* pada media sosial Twitter.

Adapun Tujuan dari penelitian ini yaitu untuk mengimplementasikan optimalisasi model yang diterapkan melalui pengujian baseline dengan menggunakan ekstraksi fitur TF-IDF pada metode NBSVM dalam mengklasifikasikan topik di Twitter dan juga mengetahui pengaruh penerapan ekspansi fitur *FastText* pada metode NBSVM dalam mengklasifikasikan topik di Twitter.

Bagian pertama ini adalah pendahuluan. Selanjutnya terdiri dari bagian Studi Terkait, Sistem yang Dibangun, Evaluasi, dan Kesimpulan. Pada bagian Studi Terkait berisikan studi atau riset yang telah dilakukan sebelumnya. Bagian Sistem yang Dibangun menjelaskan bagaimana pembangunan sistem secara umum dengan menggunakan

skema dan penjelasan mengenai metode yang akan digunakan dalam membangun sistem. Evaluasi membuat hasil pengujian dan analisis dari sistem. Kesimpulan berisikan kesimpulan dan saran untuk penelitian selanjutnya.

II. KAJIAN TEORI

Pada penelitian[5], membahas Ekspansi Fitur dengan *word embbeding* untuk klasifikasi topik twitter menggunakan tiga metode yaitu *Naïve Bayes*(NB), *Support Vector Machine*(SVM), dan *Logistic Regression* dengan ekspansi fitur *World2vec*. Penelitian ini bertujuan untuk mengurangi ketidakcocokan kosakata dengan "*embedding word*" untuk klasifikasi topik *tweet*. Berdasarkan hasil skenario pengujian yang telah dilakukan menunjukkan bahwa penggunaan ekspansi fitur pada metode ini mampu melakukan proses klasifikasi. Penggunaan ekspansi fitur dapat meningkatkan performa secara konsisten saat menggunakan pengklasifikasian *Logistic Regression*, Berbeda dengan Pengklasifikasian NB yang memberikan hasil yang beragam.

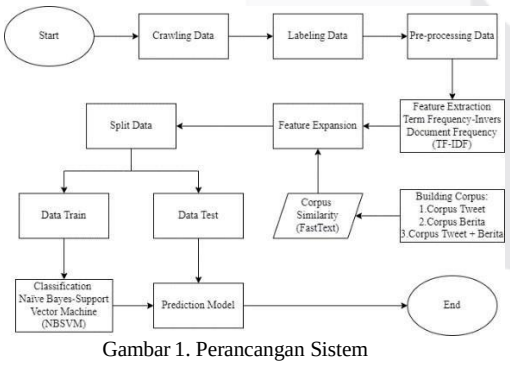
Terdapat juga pada[7], yang mengatakan bahwa jika ketidakmunculan kata pada data latih dapat diatasi dengan *FastText* karena memiliki kemampuan memberikan representasi kata tersebut dan juga mampu menangani *out of vocabulary*. Ketidakmunculan kata tersebut akan didapatkan *embedding* vektornya dengan cara memecah kata tersebut menjadi n-gram yang berupa kumpulan urutan suku kata. Pada hal tersebut, dari hasil eksperimen dapat dibuktikan bahwa kinerja *FastText* lebih baik dari *word2vec* dan *GloVe*.

NB sangat baik dalam pengklasifikasian teks. Kombinasi dari NB dan SVM menghasilkan tingkat akurasi yang lebih baik dan kinerja yang lebih kuat[8].

Oleh karena itu pada penelitian ini akan dibuktian apakah *algoritma Naïve Bayes-Support Vector Machine* (NBSVM) dengan ekspansi fitur *FastText* efektif pada klasifikasi topik Twitter.

III. METODE

Sistem topik klasifikasi yang akan dibangun pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Perancangan Sistem

A. *Crawling* Data

Pada penelitian ini, data yang digunakan berupa *tweet* dari Twitter. Pengumpulan data dilakukan dengan *Crawling* data. *Crawling* data merupakan pengumpulan data yang dilakukan dengan bantuan API(*Application Program Interface*) yang telah disediakan oleh *developer* twitter. Untuk mendapatkan API twitter harus melalui proses autentifikasi. Data yang didapatkan akan disimpan dalam file dengan format *Comma Separated Values*(CSV). Tabel 1 menggambarkan *tweet* yang berhasil diperoleh.

Tabel 1. Gambaran Hasil Proses Crawling

No	Tweet
1	Kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta, terbebas dari pandemi dan perekonomian nasional kembali pulih. #EkonomiTumbuhIndonesiaTangguh #TransportasiMaju #EkonomiBangkit #PakeMaskerHargaMati #BersatuLawanCovid #LawanCovid19
2	Jumlah peserta yang mendaftar dan berpartisipasi di Piala Presiden Esports (PPE) 2021 menembus angka 107.389 atlet dari seluruh Indonesia.
3	Peringkat maskapai kebanggaan kita, Garuda Indonesia, semakin terpuruk. Posisinya semakin menjauhi 10 besar karena berada di peringkat ke-15. https://t.co/C1SAKZKniG

Total keseluruhan data yang diperoleh pada tahap *crawling* data dari Twitter sebanyak 36.396 yang nantinya digunakan untuk data latih dan data uji. Terdapat juga data berita yang diperoleh sebanyak 142.544 dari beberapa media yaitu republika, cnnindonesia, sindonews, Kompas, detik, dan liputan 6. Data yang terkumpul dari media tersebut akan digunakan untuk pembuatan *Corpus Similarity*. Persebaran data tersebut dapat dilihat pada tabel 2.

Tabel 2 Persebaran data Berita

Sumber	Jumlah
republika.com	53812
cnnindonesia.com	29349
sindonews.com	22401
kompas.com	15055
detik.com	7974
liputan6.com	251
Total	142544

B. *Labeling* Data

Karena menggunakan algoritma *supervised learning*, maka diperlukan label agar memudahkan untuk proses klasifikasinya. Data yang sudah terkumpul pada proses *crawling* akan diberi label yang didapat dari penelitian yang telah dilakukan sebelumnya yaitu sebanyak 12 label. *Labeling* data dilakukan dengan prinsip *majority votes* oleh 4 orang. Tabel 3 merupakan contoh hasil data *crawling* yang telah diberi label.

Tabel 3 Contoh pemberian label pada Tweet

No	Tweet	Label
1	Kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta, terbebas dari pandemi dan perekonomian nasional kembali pulih. #EkonomiTumbuhIndonesiaTangguh #TransportasiMaju #EkonomiBangkit #PakeMaskerHargaMati #BersatuLawanCovid #LawanCovid19	Kesehatan
2	Jumlah peserta yang mendaftar dan berpartisipasi di Piala Presiden Esports (PPE) 2021 menembus angka 107.389 atlet dari seluruh Indonesia.	Olahraga
3	Peringkat maskapai kebanggaan kita, Garuda Indonesia, semakin terpuruk. Posisinya semakin menjauhi 10 besar karena berada di peringkat ke-15. https://t.co/C1SAKZKniG	Perhubungan

Pelabelan data terbagi menjadi 12 label dengan 2-3 *keyword* yang digunakan pada setiap label. Persebaran label memiliki persentase yang cukup seimbang antara 5.79% hingga 12.54%. Tabel 4 merupakan persebaran data yang telah dilakukan pelabelan.

Tabel 4 Persebaran data Topik Tweet

No	Topik	Jumlah	Persentase
1	Pendidikan	4565	12.54%
2	Umum	4397	12.08%
3	Teknologi	3531	9.70%
4	Olahraga	3054	8.39%
5	Agama	3023	8,31%
6	Entertainment	2941	8.08%
7	Kesehatan	2923	8.03%
8	Perhubungan	2585	7.10%
9	Politik	2552	7.01%
10	Hukum	2525	6.94%
11	Bisnis	2191	6.02%
12	Iklan	2109	5.79%
	Total	36396	

C. *Pre-processing* Data

Pre-Processing merupakan tahap pertama dalam pemodelan sistem klasifikasi, yang bertujuan untuk mempersiapkan data sehingga siap diolah dan meningkatkan performa klasifikasi pada topik twitter itu sendiri [9]. Berikut merupakan Langkah *Pre-processing* data, antara lain:

1. Data *Cleaning* melakukan penghapusan punctuation atau tanda baca dan karena data yang diambil dari twitter maka perlu juga dilakukan penghapusan pada *mention*, *hastag* dan *retweet* dengan menggunakan *Regular Expression*[9].
2. *Case Folding* adalah mengubah kata menjadi huruf kecil semua dan menghilangkan huruf selain a sampai huruf z dengan menggunakan *Regular Expression*[9].
3. *Tokenization* dilakukan untuk memotong kalimat menjadi levelnya per kata pada suatu dokumen yang nantinya digunakan sebagai fitur pada saat proses klasifikasi dengan menggunakan *library* NLTK[9].
4. *Stopword Removal* dilakukan untuk menghilangkan kata-kata yang tidak penting seperti kata sandang dan kata hubung seperti: “ya”, “dan”, “ini”, “itu”, “yang”, “atau”, dan lain-lain[5][9]. Pada tahapan ini, digunakan *library* NLTK.
5. *Stemming* untuk menjadikan suatu kata sebagai kata dasar[9]. Pada tahapan ini dilakukan dengan menggunakan *library* Sastrawi.

Tahapan pada *pre-processing* dapat dilihat pada Gambar table 2.

No	Tahapan	Sebelum	Sesudah
1	Data *Cleaning*	Kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta, terbebas dari pandemi dan perekonomian nasional kembali pulih. #EkonomiTumbuhIndonesiaTangguh #TransportasiMaju #EkonomiBangkit #PakeMaskerHargaMati #BersatuLawanCovid #LawanCovid19	Kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta terbebas dari pandemi dan perekonomian nasional kembali pulih
2	*Case Folding*	Kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta terbebas dari pandemi dan perekonomian nasional kembali pulih	kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta terbebas dari pandemi dan perekonomian nasional kembali pulih
3	*Tokenization*	kegiatan ini semoga mempercepat target vaksinasi nasional agar kekebalan komunal segera tercipta terbebas dari pandemi dan perekonomian nasional kembali pulih	[‘kegiatan’, ‘ini’, ‘semoga’, ‘mempercepat’, ‘target’, ‘vaksinasi’, ‘nasional’, ‘agar’, ‘kekebalan’, ‘komunal’, ‘segera’, ‘tercipta’, ‘terbebas’, ‘dari’, ‘pandemi’, ‘dan’, ‘perekonomian’, ‘nasional’, ‘kembali’, ‘pulih’]
4	*Stopword Removal*	[‘kegiatan’, ‘ini’, ‘semoga’, ‘mempercepat’, ‘target’, ‘vaksinasi’, ‘nasional’, ‘agar’, ‘kekebalan’, ‘komunal’, ‘segera’, ‘tercipta’, ‘terbebas’, ‘dari’, ‘pandemi’, ‘dan’, ‘perekonomian’, ‘nasional’, ‘kembali’, ‘pulih’]	[‘kegiatan’, ‘semoga’, ‘mempercepat’, ‘target’, ‘vaksinasi’, ‘nasional’, ‘kekebalan’, ‘komunal’, ‘tercipta’, ‘terbebas’, ‘pandemi’, ‘perekonomian’, ‘nasional’, ‘pulih’]
5	*Stemming*	[‘kegiatan’, ‘semoga’, ‘mempercepat’, ‘target’, ‘vaksinasi’, ‘nasional’, ‘kekebalan’, ‘komunal’, ‘tercipta’, ‘terbebas’, ‘pandemi’, ‘perekonomian’, ‘nasional’, ‘pulih’]	[‘giat’, ‘moga’, ‘cepat’, ‘target’, ‘vaksinasi’, ‘nasional’, ‘kebal’, ‘komunal’, ‘cipta’, ‘bebas’, ‘pandemi’, ‘ekonomi’, ‘nasional’, ‘pulih’]

Gambar 2. Contoh Input dan Output Tahapan Pre-processing

D. *Feature Extraction Term Frequency – Invers Document Frequency*

Ekstraksi fitur merupakan hal yang penting dilakukan dalam klasifikasi karena mempengaruhi perfomansi klasifikasi. Ekstraksi fitur didasarkan pada model ruang vektor yang dilakukan dengan menghitung bobot kata-kata dan kemudian membentuk vektor fitur[10]. Pada penelitian ini ekstraksi fitur yang digunakan adalah *Term Frequency-Invers Document Frequency* (TF-IDF). TF-IDF merupakan kombinasi antara *Term Frequency* (TF*)* dan *Invers Document Frequency* (IDF).

TF menyatakan berapa kata yang muncul dalam sebuah dokumen Sedangkan IDF memberi bobot pada kata dalam sebuah dokumen, untuk kata kata yang lebih sering muncul akan diberikan bobot lebih rendah daripada kata-kata yang jarang muncul[11]. TF-IDF merupakan perkalian antara TF dengan IDF.

E. *Feature Expansion FastText*

Ekspansi fitur digunakan untuk mengatasi kesulitan dalam mengklasifikasikan *tweet* dengan cara menambahkan kata lain yang terkait dengan fitur[1][5]. Pada penelitian ini ekspansi fitur yang digunakan adalah *FastText*. *FastText* didasarkan pada model skipgram yang akan mengubah teks ke dalam bentuk vektor dengan ekstraksi fitur dan mewakili setiap kata sebagai karakter n-gram. Fitur ini meningkatkan pembelajaran pada Bahasa yang sangat terpengaruh, seperti kata-kata cinta, dicintai dan mencinta semua memiliki representasi vektor yang serupa, meskipun cenderung muncul dalam konteks yang berbeda[12].

Nilai vektor kata direpresentasikan dengan menjumlahkan nilai pada tiap n-gram. Dengan *FastText*, sangat memungkinkan untuk beberapa n-gram yang membentuk kata yang tidak ditemui dalam *corpus* muncul pada n-gram yang berada di dalam *corpus*[13]. Sehingga dapat direpresentasikan dengan baik. *Output* dari *FastText* adalah berupa *word similarity*. Tabel 5 merupakan kata-kata yang memiliki kemiripan dengan kata “Ajar” dan peringkat 1 hingga 10 menunjukkan tingkat kemiripan dengan kata “Ajar”.

Tabel 5 Kata-kata yang serupa dengan Ajar

Kata	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5

Ajar	matakuliah	siswa	matematika	murid	kurikuler
	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
	matematik	nonkurikuler	intrakurikuler	matematikawan	kuliah

F. Naïve Bayes-Support Vector Machine

Naïve Bayes-Support Vector Machine (NBSVM) banyak digunakan sebagai dasar untuk metode lain dalam klasifikasi teks[18]. Namun, kinerjanya sangat bervariasi tergantung pada varian, fitur, dan kumpulan data yang digunakan. Metode ini menggunakan fitur *Naive Bayes* (NB) untuk menimbang representasi bag-of-n-gram yang jarang. N-gram menangkap urutan kata dalam konteks pendek dan fitur NB memberikan bobot lebih pada kata-kata penting tersebut.

NBSVM memiliki kinerja yang baik pada teks berukuran kecil serta memiliki kinerja yang baik juga pada dokumen yang lebih panjang untuk kasus klasifikasi topik dan hasilnya seringkali lebih baik daripada publikasi yang sebelumnya. Metode NBSVM juga dinilai sebagai metode canggih yang memiliki tujuan untuk menangani sekumpulan fitur[18].

G. Performansi Sistem

Mengukur Perfomansi Sistem penting ketika mengevaluasi kinerja sistem yang telah dibangun. Untuk mengukur Perfomansi sistem digunakan akurasi, presisi, *recall*, dan *F1 Measure*. Langkah-langkah ini didefinisikan oleh *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Pada penelitian ini, digunakan *confusion matrix* yaitu matriks yang menunjukan berapa banyak sampel yang berada di kelas yang benar dan berapa banyak sampel yang berada di kelas yang salah. Nilai TP dan TN mengacu pada jumlah sampel yang kelasnya diperkirakan benar. FP dan FN adalah perkiraan jumlah kelas yang salah. Tabel 6 merupakan tabel *confusion matrix*.

Tabel 6. Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	*True Positive* (TP)	*False Positive* (FP)
	Negative	*False Negative* (FN)	*True Negative* (TN)

Akurasi merupakan hubungan antara nilai yang diperoleh dengan nilai yang sebenarnya. Berikut persamaan menghitung akurasi.

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

Presisi merupakan perbandingan antara jumlah sampel yang diprediksi berada di kelas yang benar dengan jumlah sampel diprediksi oleh sistem klasifikasi[6]. Berikut persamaan menghitung presisi.

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

Recall merupakan rasio antara jumlah sampel yang diprediksi benar dengan jumlah sampe yang seharusnya diprediksi[6]. Berikut Persamaan menghitung recall.

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

F1-Measure adalah statistic pengukuran untuk menganalisi kinerja klasifikasi. Nilai *F1-Measure* berkisar antara 0 dan 1, dimana 1 menunjukkan yang berkinerja terbaik[6]. Berikut Persamaan menghitung *F1-Measure*.

$$F1 - Measure = 2\,x\,\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (4)$$

IV. HASIL DAN PEMBAHASAN

Hasil Pengujian dan Analisis Hasil Pengujian.

A. Skenario dan Hasil Pengujian

Untuk mencapai tujuan dari penelitian ini terdapat dua skenario, yang pertama adalah baseline dengan menggunakan ekstraksi fitur *Term Frequency-Invers Document Frequency* (TF-IDF) pada metode *Naïve Bayes-Support Vector Machine* (NBSVM) dalam mengklasifikasikan topik di twitter dan skenario kedua yaitu penerapan ekspansi fitur *FastText* pada metode NBSVM dalam mengklasifikasikan topik di twitter.

4.1.1 Pengujian pertama

Pengujian pertama yaitu menggunakan baseline dengan pembobotan TF-IDF dengan parameter *max feature*=1000 pada metode NBSVM. Dalam pengujian ini rasio yang digunakan adalah 70:30, 80:20, dan 90:10 yang nantinya akan dibandingkan untuk perbandingan data latih dan data ujinya. Pengujian ini dilakukan sebanyak 3 kali pada setiap rasio lalu dihitung nilai rata-rata dari akurasi dan *F1-Measure* yang akan diambil dari pengujian tersebut. Tabel 6 merupakan hasil dari pengujian pertama.

Tabel 7 Hasil Perfomansi Baseline NBSVM + TF-IDF

Rasio	Akurasi (%)	*F1-Measure*(%)
90 : 10	**80,81**	**82.06**
80 : 20	80,51	81,50
70 : 30	79,99	81,00

Setelah didapatkan rasio dengan perfomansi terbaik dalam akurasi dan *F1-Measure* adalah menggunakan rasio 90:10 untuk perbandingan data latih dan data uji, selanjutnya dilakukan percobaan untuk pembobotan TF-IDF dengan parameter *max feature* yang merupakan maksimal fitur kata yang digunakan dalam proses TF-IDF yang diambil berdasarkan frekuensi kemunculan kata. Parameter *max feature* yang digunakan dalam penelitian ini menggunakan *max feature* 1000, 5000 dan 10000 dengan rasio perbandingan data latih dan data ujinya 90:10. Karena dalam pengujian rasio pada di atas sudah dilakukan pembobotan TF-IDF dengan parameter *max feature* 1000, maka akan dilakukan pengujian terhadap 5000 dan 10000. Tabel 8 merupakan hasil dari pengujian pembobotan TF-IDF dengan parameter *max feature* dengan rasio 90:10 untuk perbandingan data latih dan data uji.

Tabel 8. Perbandingan Hasil Perfomansi NBSVM + TF-IDF

Max Feature	Akurasi (%)	*F1-Measure*(%)
1000	**80,81**	**82.06**

5000	78,25	79,45
10000	77,88	79,30

Dari tabel 8 dapat disimpulkan bahwa TF-IDF dengan parameter *max feature* sebanyak 1000 dan rasio 90:10 untuk perbandingan data latih dan data uji memiliki perfomansi terbaik untuk akurasi sebesar 80,81% dan *F1-Measure* sebesar 82,06%. Jadi, untuk pengujian selanjutnya akan menggunakan rasio dengan perbandingan 90:10 untuk data latih dan data ujinya serta pembobotan TF-IDF dengan parameter *max feature* sebanyak 1000.

4.1.2 Pengujian kedua

Pengujian kedua yaitu penerapan ekspansi fitur dengan *FastText* setelah dilakukan uji terhadap baseline dengan pembobotan TF-IDF pada metode NBSVM. Pada pengujian ini dilakukan dengan menggunakan 3 Corpus *FastText*, yaitu *Corpus* data *tweet*, *Corpus* data berita dan *Corpus* data *tweet* + berita yang dilakukan sama dengan pengujian sebelumnya yaitu dengan perbandingan rasio 90:10 untuk data latih dan data ujinya serta pembobotan TF-IDF dengan parameter *max feature* sebanyak 1000. Pada tiap *corpus FastText* dilakukan *Top Similarity* yaitu dengan mengambil beberapa kata-kata yang memiliki nilai *similarity* tertinggi dari corpus *FastText* berdasarkan kata target untuk proses ekspansi fitur. *Top Similarity* yang dilakukan dalam pengujian ini adalah *Top Similarity* 1, *Top Similarity* 5 dan *Top Similarity* 10. Pengujian ini dilakukan sebanyak 3 kali pada setiap *corpus FastText* lalu dihitung nilai rata-rata dari akurasi dan *F1-Measure* yang akan diambil dari pengujian tersebut. Tabel 9 merupakan hasil dari pengujian pertama.

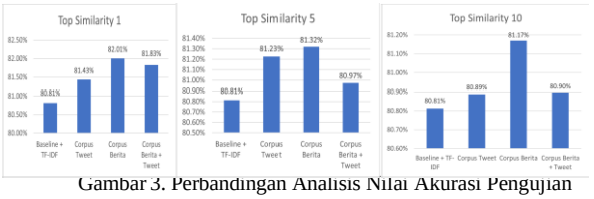
Tabel 9. Perbandingan Hasil Perfomansi Baseline + TF-IDF dengan Ekspansi Fitur pada NBSVM

Top Similarity	Akurasi (%)				*F1-Measure (%)*			
	Baseline + TF-IDF	*Corpus* Tweet	*Corpus* Berita	*Corpus* Tweet + Berita	Baseline + TF-IDF	*Corpus* Tweet	*Corpus* Berita	*Corpus* Tweet + Berita
1	80,81	81,43 (+0,62)	**82,01 (+1,20)**	81,83 (+1,02)	82.06	82,43 (+0,37)	**82,72 (+0,66)**	82,80 (+0,74)
5	80,81	81,23 (+0,42)	81,32 (+0,51)	81,22 (+0,41)	82.06	82,20 (+0,14)	82,33 (+0,27)	82,33 % (+0,27)
10	80,81	80,89 (+0,08)	81,17 (+0,36)	80,90 (+0,09)	82.06	81,84 (-0,22)	82,18 (-0,12)	82,12 (+0,06)

Dari tabel 9 dapat dilihat nilai akurasi pada pengujian ini terjadi peningkatan sebanyak 0.62% dibandingkan dengan akurasi dan *F1-Measure* pada pengujian sebelumnya begitu juga dengan *F1-Measure* yang memiliki peningkatan sebanyak 0,66% dari pengujian sebelumnya. Nilai akurasi tertinggi berada pada *Top similarity* 1 dengan menggunakan *corpus* berita sebesar 82,01% yang dimana pada pengujian sebelumnya 80,81% dan untuk *F1-Measure* pada *Top Similarity* yang sama yaitu 1 sebesar 82,72% yang pada pengujian sebelumnya hanya 82,06%.

B. Analisis Hasil Pengujian

Setelah dilakukannya pengujian, hasilnya akan divisualisasikan yang dapat dilihat pada gambar 3.



Gambar 3. Perbandingan Analisis Nilai Akurasi Pengujian

Gambar 3 merupakan visualisasi *bar chart* yang dikelompokkan berdasarkan variasi penggunaan fitur pada proses ekspansi fitur. Ketika menerapkan ekspansi fitur *FastText* pada sistem, nilai akurasi meningkat dibandingkan dengan yang tidak dilakukan ekspansi fitur dengan *Fasttext*.

Dapat dilihat pada baseline dengan penerapan TF-IDF mendapatkan akurasi sebesar 80,81%, ketika dilakukan penerapan ekspansi fitur untuk model *Top Similarity* 1 menggunakan *corpus* data berita terjadi peningkatan nilai akurasi sebesar 1,2% dari baseline dengan TF-IDF menjadi 82,01%. Lalu, saat menggunakan *corpus* data *tweet* terjadi peningkatan nilai akurasi sebesar 0,62% dari baseline dengan TF-IDF menjadi 81,83%. Dan saat menggunakan *corpus* data *tweet* + berita terjadi peningkatan nilai akurasi sebesar 1,2% dari baseline dengan TF-IDF menjadi 82,01%.

Untuk model *Top Similarity* 5 menggunakan *corpus* data berita terjadi peningkatan nilai akurasi sebesar 0,51% dari baseline dengan TF-IDF menjadi 81,32%. Lalu, saat menggunakan *corpus* data *tweet* terjadi peningkatan nilai akurasi sebesar 0,42% dari baseline dengan TF-IDF menjadi 81,23%. Dan saat menggunakan corpus data *tweet* + berita terjadi peningkatan nilai akurasi sebesar 0,41% dari baseline dengan TF-IDF menjadi 82,22%.

Untuk model *Top Similarity* 10 menggunakan *corpus* data berita terjadi peningkatan nilai akurasi sebesar 0,36% dari baseline dengan TF-IDF menjadi 81,17%. Lalu, saat menggunakan *corpus* data *tweet* terjadi peningkatan nilai akurasi sebesar 0,08% dari baseline dengan TF-IDF menjadi 80,89%. Dan saat menggunakan *corpus* data *tweet* + berita terjadi peningkatan nilai akurasi sebesar 0,09% dari baseline dengan TF-IDF menjadi 80,90%.

Dapat dilihat dari gambar 2, pada model *Top Similarity* 1, tepatnya pada saat menggunakan *corpus* data berita mendapatkan peningkatan akurasi paling terbaik dibandingkan dengan baseline yang menggunakan TF-IDF dengan kenaikkan akurasi sebesar 1,2% dari baseline dengan TF-IDF menjadi 82,01%.

Top Similarity 1 memiliki performansi terbesar karena pada model *Top Similarity* 1, proses ekspansi fitur akan mengganti fitur kata pada vektor TF-IDF dengan kata yang memiliki nilai *similarity* terbesar yang terdapat pada *corpus*. Berbeda halnya pada model *Top Similarity* 5 dan *Top Similarity* 10, kata yang menjadi pengganti dari fitur kata bisa saja memiliki nilai *similarity* rendah, oleh karena hal itu bisa saja membuat performansi dari sistem yang dibuat menurun.

V. KESIMPULAN

Telah dilakukan penelitian klasifikasi kelas topik di Twitter dengan dua skenario, yaitu baseline dengan menggunakan ekstraksi fitur *Term Frequency-Invers Document Frequency* (TF-IDF) pada metode *Naïve Bayes-Support Vector Machine* (NBSVM) dalam mengklasifikasikan topik di twitter dan skenario kedua yaitu penerapan ekspansi fitur *FastText* pada metode NBSVM dalam mengklasifikasikan topik di twitter. Ekspansi Fitur diterapkan dengan tiga corpus *FastText* yaitu *Corpus* data *tweet*, *Corpus* data berita dan *Corpus* data *tweet* + berita. Pada tiap corpus, dilakukan dengan tiga model top similarity yaitu *Top Similarity* 1, *Top Similarity* 5 dan *Top Similarity* 10.

Berdasarkan pengujian dan hasil analisis yang telah dilakukan pada penelitian ini, dapat disimpulkan bahwa penerapan ekspansi fitur dengan menggunakan *FastText* terbukti berpengaruh dapat meningkatkan perfomansi pada metode NBSVM untuk klasifikasi topik di twitter. Klasifikasi topik di twitter dengan menggunakan NBSVM dengan ekspansi fitur *FastTest* mendapatkan akurasi sebesar 82,01% dengan model *Top Similariy* 1 dengan *corpus* data berita sedangkan jika dibandingkan dengan yang tidak menggunakan ekspansi fitur hanya mendapatkan akurasi 80,81% yang berarti terjadi peningkatan akurasi sebesar 1,2%.

Adapun saran untuk studi selanjutnya perlu dilakukan perbandingan NBSVM dengan *Naïve bayes* (NB) dan *Support Vector Machine* (SVM) secara terpisah untuk mengetahui perbedaan perfomansi sistemnya.

REFERENSI

[1] W. Xie, F. Zhu, J. Jiang, E. Lim and K. Wang, "TopicSketch: Real-time Bursty Topic Detection from Twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2216-2229, 1 Aug 2016.

[2] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman and C. Espina, "Cl sifying Twitter Topic-Networks Using Social Network Analysis," *social media + society*, January 2017.

[3] F. F. Irfani, M. A. Fauzi and Y. A. Sari, "News Classification on Twitter Using Naive Bayes and Hypernym-Hyponym Based Feature Expansion," *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, pp. 317-321, 2018.

[4] M. A. Fauzi, R. F. Nur Firmansyah and T. Afirianto, "Improving Sentiment Analysis of Short Informal Indonesian Product Reviews using Synonym Based Feature Expansion," *Telkomnika*, vol. 16, no. 3, pp. 1345-1350, June 2018.

[5] E. B. Setiawan, D. H. Widyantoro and K. Surendro, "Feature tweet topic classification," *2016 10th International Confere Services and Applications (TSSA)*, p

[6] A. G. D'sa, I. Illina and D. Fohr, "BERT and fastText Embedd Speech," *2020 International Multi-Conference on: "Organi Technologies" (OCTA)*, pp. 1-5

[7] A. Nurdin, B. A. Seno Aji, A. Bustamin and Z. Abidin, "PE EMBEDDING WORD2VEC, GLOVE,," *Jurnal TEKNOKOM*

[8] A. N. Muhammad, S. Bukhori and P. Pandunata, "Sentiment YouTube Comments Using Naïve Bayes – Support Vector M *ICOMITEE 2019*, 2019.

[9] H. Tantyoko, A. and U. N. Wisesty, "Perbandingan Pembob menggunakan Decision Tree," *Univer*

[10] R. Dzisevič and D. Šešok, "Text Classification using Differen *Open Conference of Electrical, Electronic and Information*

[11] "Text Mining: Use of TF-IDF to Examine the Relevance of *Journal of Computer Applications*, vol. 181, no. 1,

[12] I. Santos, N. Nedjah and L. d. M. Mourelle, "Sentiment analys with fastText embeddings," *2017 IEEE Latin American Conf (LA-CCI)*, pp. 1-5, 2017

[13] F. Alfariqi, W. Maharani and J. H. Husen, "Klasifikasi Senti Pemilihan Kandidat Karyawandengan Menggunakan Convol Embeddings," *e-Proceeding of Engineering*,

[14] A. S. Nugroho, A. B. Witarto and D. Handoko, "Support Ve dalam Bioinformatika-," *IlmuKom*

[15] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessi Classifier," *2016 IEEE 40th Annual Computer Software and A*

pp. 618-619, 2016.

[16] M. Wongkar and A. Angdresey, "Sentiment Analysis Using Crawler : Twitter," *2019 Fourth International Conference on I*

1-5, 2019.

[17] V. S and S. K.C, "Prediction of Loan Risk using Naive Ba *International Conference on Advancements in Computing Tec*

pp. 110-113, 2018.

[18] S. Wang and C. D. Manning, "Baselines and Bigrams: Si Classification," *50th Annu. Meet. Assoc. Comput. Linguist. AC*

July 2012.