Dataset Indonesia untuk Analisis Sentimen

Article *in* Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) · November 2019

DOI: 10.22146/jnteti.v8i4.533

5 authors, including:

Ridi Ferdiana
Universitas Gadjah Mada
119 PUBLICATIONS **211** CITATIONS

SEE PROFILE

Wiliam Fajar
Universitas Gadjah Mada
3 PUBLICATIONS **0** CITATIONS

SEE PROFILE

Dataset Indonesia untuk Analisis Sentimen

Ridi Ferdiana[1], Fahim Jatmiko[2], Desi Dwi Purwanti[1], Artmita Sekar Tri Ayu[1], Wiliam Fajar Dicka[1]

***Abstract*—This paper present a text dataset which can be used in the field of text analysis, especially sentiment analysis. This dataset covers the primary data which consists of 10,806 lines of Indonesian text data originated from Twitter social media, which categorized into three categories that are positive, negative, and neutral; and the raw data which consists of 454,559 lines of unprocessed data. Other than that, on the labeled data, the data is cleaned by removing many kind of noises in the data, such as symbols or urls. In this paper, the presented dataset is tested using a sentiment analysis model to make sure that this dataset is suitable to be used in the field of text analysis. The testing is done by measuring the model accuracy which is trained using this dataset and then comparing it to other model which is trained using already published dataset. After testing the data using various algorithm, such as SVM, KNN, and SGD, the accuracy result between our data and the comparison data are more or less equal with around 4% to 12% differences in accuracy, and prove that the dataset presented in this paper is feasible to be used in sentiment analysis. Dataset can be downloaded from link at conclusion section.**

***Intisari*—Makalah ini menyajikan sebuah *dataset* teks berbahasa Indonesia untuk digunakan di bidang analisis teks, terutama analisis sentimen. *Dataset* ini mencakup data utama, yaitu 10.806 baris data berbahasa Indonesia yang diambil dari media sosial Twitter, yang telah dikategorikan ke dalam tiga label, yaitu positif, negatif, dan netral, beserta 454.559 baris data yang masih bersifat mentah. Selain itu, pada data yang sudah dilabeli, data sudah mengalami proses pembersihan dari elemen-elemen pengganggu di dalam data, misalnya simbol atau tautan halaman web. Dalam makalah ini, data yang disajikan sudah diuji terlebih dahulu menggunakan sebuah model sentimen analisis sederhana untuk memastikan bahwa data ini sudah sesuai untuk digunakan dalam sebuah pemodelan analisis teks secara umum. Pengujian ini dilakukan dengan melihat hasil nilai ketepatan sebuah model analisis sentimen yang menggunakan *dataset* ini pada proses pelatihan dan membandingkannya dengan model analisis yang menggunakan *dataset* lain pada proses pelatihan datanya. Setelah dilakukan pengujian menggunakan model analisis sentimen sederhana yang menggunakan algoritme SVM, KNN, dan SGD, terlihat bahwa nilai ketepatan dari data utama dan data pembanding seimbang pada masing-masing algortime, dengan perbedaan nilai ketepatan berkisar pada angka 4% sampai 12%, dan membuktikan bahwa data yang disajikan sudah layak untuk digunakan dalam pemodelan analisis sentimen. *Dataset* dapat diunduh pada tautan di bagian kesimpulan.**

***Kata Kunci—Dataset*, Analisis Teks, Analisis Sentimen, *Natural Language Processing*.**

[1] *Departemen Teknik Elektro dan Teknologi Informasi Universitas Gadjah Mada, Jln. Grafika No.2 Kampus UGM, Yogyakarta 55597 (email: ridi@ugm.ac.id)*

[2] *Microsoft Innovation Center, Jln. Grafika No.2 Kampus UGM, Yogyakarta 55597 (email: fahimj@micresearch.net)*

I. PENDAHULUAN

Faktor terpenting dalam sebuah pemodelan *machine learning* adalah data yang digunakan untuk proses pelatihan model tersebut, terlebih lagi pada pemodelan *machine learning* yang berfokus pada teks, misalnya analisis sentimen. Pada sebuah analisis sentimen, proses pelatihan yang dilakukan cenderung lebih sulit jika dibandingkan dengan bidang *machine learning* lainnya. Hal ini dikarenakan data yang digunakan pada proses pelatihan analisis sentimen merupakan data yang bersifat subjektif, seperti opini, yang tidak memiliki nilai konkret. Ditambah lagi, data ini bersumber dari manusia dan setiap manusia memiliki cara atau selera yang berbeda-beda dalam mengungkapkan opininya [1]. Oleh sebab itu, tanpa adanya data-data yang mendukung, membuat sebuah model analisis sentimen yang akurat sangatlah sulit. Ketersediaan data-data yang berkualitas pun masih terbilang minim, terutama data berbahasa non-Inggris. Hal inilah yang mendorong disajikannya sebuah *dataset* berbahasa Indonesia untuk dapat digunakan dalam proses sentimen analisis yang disebut *Indonesian-General-Sentiment-Analysis-Dataset*.

Twitter merupakan salah satu media sosial yang paling populer untuk digunakan sebagai sumber data pada analisis teks, seperti ditunjukkan pada Tabel I. Hal ini karena tulisan-tulisan pada media sosial Twitter, atau dapat disebut juga dengan *tweet*, memiliki struktur yang sangat cocok untuk digunakan pada analisis. Tidak heran jika *dataset-dataset* yang sudah dipublikasi dari penelitian-penelitian lain sering menggunakan Twitter sebagai sumber datanya, sebagai contoh ASTD, yang merupakan *dataset* berbahasa Arab [2]. Dengan alasan ini, dipilih media sosial Twitter sebagai sumber data untuk ditampilkan pada makalah ini.

II. KONTEN UTAMA

A. Analisis Sentimen Berbahasa Indonesia

Pada konteks analisis sentimen berbahasa Indonesia, sebenarnya sudah ada *dataset-dataset* yang dipublikasikan secara umum, misalnya *dataset* dari penelitian *tweet* bahasa Indonesia nonformal yang berisi 4.000 *tweet* dengan label polaritas utama, yaitu positif, negatif, dan netral [3], atau *Indonesian-Emotion-Twitter-Dataset*, yang berisi 4.403 *tweet* dengan polaritas emosi yang lebih spesifik (*love, joy, anger, sadness, fear*) [4].

Namun, *dataset* berbahasa Indonesia yang lebih banyak ditemukan adalah data yang berasal dari penelitian-penelitian analisis sentimen yang memiliki cakupan yang lebih spesifik dan kurang cocok untuk digunakan pada analisis sentimen yang jangkauannya lebih luas, misalnya analisis sentimen terhadap operator seluler Indonesia yang berjumlah 1.000 *tweet* yang berisi komentar mengenai operator seluler dan dibagi menjadi dua polaritas, yaitu positif dan negatif [5]. Beberapa contoh lainnya yaitu analisis sentimen yang berfokus pada

TABEL I
PERBANDINGAN MEDIA SOSIAL UNTUK ANALISIS SENTIMEN

Pembanding	Twitter	Facebook	Instagram
Jenis Data	Hampir seluruhnya teks, sangat sedikit foto	Gabungan antara teks, foto, video, *games*, dll.	Hampir seluruhnya nonteks, teks hanya untuk *caption*
Isi Data	Sebagian besar berisi opini (*tweet)*	Bermacam-macan (opini, berita, cerita)	Teks hanya untuk *caption* gambar
Panjang Data	Maksimal 280 karakter	Maksimal 63.206 karakter	Maksimal 2.200 karakter

TABEL II
BEBERAPA *DATASET* UNTUK ANALISIS SENTIMEN BERBAHASA INDONESIA

Dataset	Sumber	Ukuran
Informal Indonesian Tweets Dataset	Twitter	4.000
Indonesian-Emotion-Twitter-Dataset	Twitter	4.403
Indonesian Mobile Operator Dataset	Twitter	1.000
Dataset Cerita Pendek	cerpenmu.com	121
Automatic Indonesian Translations Movie Reviews	imdb.com	1.400

pengategorian cerita pendek yang menggunakan beberapa cerita pendek sebagai *dataset* [6].

Selain itu, *dataset* lain yang sering digunakan pada analisis sentimen bahasa Indonesia adalah *dataset* terjemahan yang bersumber dari bahasa Inggris, seperti pada penelitian analisis sentimen untuk ulasan film [7]. Terlebih lagi, ada beberapa penelitian analisis sentimen yang tidak memublikasikan data yang digunakan untuk pelatihan modelnya karena beberapa alasan, misalnya data-data yang digunakan merupakan data sensitif atau bersumber dari perusahaan komersial. Masih kurangnya publikasi *dataset* analisis sentimen berbahasa Indonesia menjadi salah satu faktor pendorong dari publikasi *Indonesian-General-Sentiment-Analisis* ini. Beberapa *dataset* analisis sentimen bahasa Indonesia yang ditemukan diperlihatkan pada Tabel II.

Jika melihat pada lingkup internasional, analisis sentimen adalah bidang yang perkembangannya sangat bergantung pada negara masing-masing. Di saat bidang klasifikasi kuantitatif dapat melakukan kolaborasi data dan pemodelan secara internasional, bahasa menjadi sebuah penghalang terjadinya kolaborasi internasional di bidang analisis sentimen. Masalahnya pun tidak sebatas data saja, tetapi model yang dikembangkan juga perlu memperhatikan aturan bahasa yang tentu berbeda-beda pada masing-masing negara. Oleh karena itu, bidang analisis sentimen di Indonesia dituntut untuk berkembang lebih pesat dari bidang klasifikasi lainnya agar dapat bersaing di lingkup internasional. *Dataset* ini sendiri menjadi salah satu bentuk kontribusi di bidang analisis sentimen bahasa Indonesia.

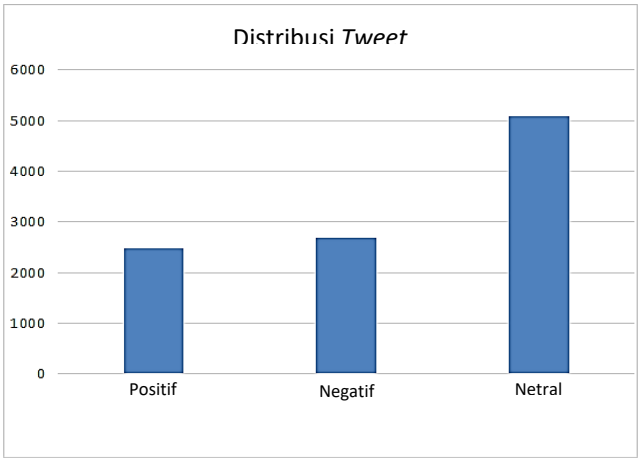
TABEL III
INDONESIAN-GENERAL-SENTIMENT-ANALYSIS-DATASET

Sentimen	*Tweet*
-1	lagu bosan apa yang aku save ni huhuhuhuhuhuhuhuhuhuuuuuuuuuuuuuu
-1	kita lanjutkan saja diam ini hingga kau dan aku mengerti tidak semua kebersamaan harus melibatkan hati
1	doa rezeki tak putus inna haa zaa larizquna maa lahu min na fadesungguhnya ini ialah pemberian kami kepada kamu
1	makasih loh ntar kita bagi hasil aku 99 9 sisanya buat kamu
0	yg selama ini nunggun video dari aku nih retweet yg mau full
0	barusan liat tulisan di belakang truk rela injek kopling demi kamu bisa shopping
-1	ku kira aku introvert ternyata karna nga punya uang aja jadi males ke mana mana
1	ya aku akan menjadi satu satunya bukan nomor satu tetapi satu satunya
0	aku nang kampus untuk ngopi ketemu wong2 podo kabeh takok e lapo mengubah sosyeti

Memajukan bidang analisis sentimen bahasa Indonesia memang memiliki berbagai tantangan. Selain kebutuhan data yang belum terpenuhi, ruang lingkup di dalam bidang analisis sentimen sendiri sangatlah luas. Sumber data yang berbeda-beda, seperti berita, ulasan, dan pesan singkat, dapat dikelompokkan menjadi pemodelan analisis sentimen yang berbeda. Menyediakan sebuah data yang dapat digunakan untuk semua jenis analisis sentimen tentu tidak memungkinkan. Namun, untuk pemodelan analisis sentimen yang spesifik seperti contoh tersebut, *dataset* ini dapat digunakan secara parsial. Dengan menyediakan *dataset* yang mencakup lingkup yang luas, seorang peneliti dapat memanfaatkan bagian dari *dataset* yang relevan dengan penelitiannya. Hal inilah yang menjadi salah satu sasaran dalam penyediaan *dataset* ini.

B. Properti Data

Dataset ini diambil dari media sosial Twitter menggunakan API Twitter dan *library tweepy* dari Python dengan jenjang waktu antara bulan September sampai Desember 2018. Metode pengambilan data adalah dengan mengambil *tweet-tweet* berbahasa Indonesia secara acak menggunakan kata kunci berupa kata-kata yang sering digunakan dalam percakapan sehari-hari. Dalam jenjang waktu tersebut, telah diambil *tweet* sebanyak 454.559 *tweet. Tweet-tweet* yang telah diambil kemudian diseleksi terlebih dahulu untuk memilih *tweet* yang cocok untuk digunakan dalam pelatihan model analisis sentimen. Dari proses seleksi tersebut, didapatkan jumlah akhir *tweet* pada *Indonesian-General-Sentiment-Analysis-Dataset* ini, yaitu sebanyak 10.806 *tweet*. Beberapa contoh *tweet* pada *dataset* ini diperlihatkan pada Tabel III.



Gbr. 1 Distribusi label pada *dataset*.

Tweet yang sudah diseleksi ini kemudian diberi label secara manual menggunakan tiga jenis polaritas, yaitu polaritas positif (1) untuk melambangkan *tweet* yang memiliki sentimen yang bersifat positif, misalnya persetujuan dan kebahagiaan; polaritas negatif (-1) untuk melambangkan *tweet* yang memiliki sentimen negatif, seperti penolakan atau amarah dan kekecewaan; dan yang terakhir polaritas netral (0) untuk *tweet* yang tidak menunjukkan sentimen positif ataupun negatif. Distribusi label positif, negatif, dan netral pada *Indonesian-General-Sentiment-Analysis-Dataset* ini diperlihatkan pada Gbr. 1.

Dari segi *pre-processing*, pada *dataset* ini telah dilakukan pembersihan elemen-elemen pengganggu pada data, misalnya simbol-simbol, tanda baca, angka, tautan halaman web, *hashtag*, dan *mention*. Pembersihan data ini dilakukan guna meningkatkan kualitas data saat digunakan untuk pelatihan model analisis sentimen. Selain itu, juga disediakan beberapa versi data dengan penerapan teknik *pre-processing* tambahan, yaitu *stemming*, yang mengubah kata menjadi bentuk dasarnya dan penghilangan kata sambung. Berikut ini adalah contoh penerapan *pre-processing* pada *Indonesian-General-Sentiment-Analysis-Dataset*.

1) Pembersihan Elemen Pengganggu: Dalam konteks data teks untuk analisis sentimen, *noise* atau gangguan dalam data berbentuk simbol-simbol, misalnya titik, koma, tanda kurung, dan lain-lain, serta angka. Dengan menghilangkan *noise* tersebut, proses klasifikasi akan menjadi lebih mudah dan fitur-fitur yang dihasilkan dari data-data ini akan menjadi lebih konsisten.

~~"RT @atrenal:~~ kita lanjutkan saja diam ini**~~,~~** hingga kau dan aku mengerti**~~,~~** tidak semua kebersamaan**~~,~~** harus melibatkan hati**~~."~~**

⇩

kita lanjutkan saja diam ini hingga kau dan aku mengerti tidak semua kebersamaan harus melibatkan hati

2) Stemming: Stemming adalah proses mengubah kata-kata menjadi bentuk dasar kata tersebut, misalnya kata “berjalan” menjadi “jalan”. *Stemming* bekerja berdasarkan *database* atau kamus untuk memastikan hasil yang didapatkan akurat dan sesuai dengan tata bahasa yang digunakan.

"RT @atrenal: kita **lanjutkan** saja diam ini, hingga kau dan aku **mengerti**. tidak semua **kebersamaan**, harus **melibatkan** hati."

⇩

“RT @atrenal: kita **lanjut** saja diam ini, hingga kau dan aku **erti**. tidak semua **sama**, harus **libat** hati.”

3) Penghilangan Kata Sambung: Stopwords atau kata sambung adalah kata-kata yang sangat sering digunakan dalam bahasa. Contoh *stopwords* dalam konteks bahasa Indonesia antara lain “aku”, “kamu”, “dia”, dan “kita”. *Stopwords* dapat dihilangkan karena terkadang kata-kata tersebut terdapat pada hampir semua data yang ada. Dengan demikian, menghilangkan *stopwords* tidak akan berpengaruh pada tahap pelatihan model.

"RT @atrenal: **~~kita~~** lanjutkan saja diam **~~ini~~**, hingga **~~kau dan aku~~** mengerti. tidak semua kebersamaan, harus melibatkan hati."

⇩

"RT @atrenal: lanjutkan saja diam, hingga mengerti. tidak semua kebersamaan, harus melibatkan hati."

III. HASIL DAN PEMBAHASAN

Indonesian-General-Sentiment-Analysis-Dataset ini akan digunakan untuk pelatihan sebuah model analisis sentimen sederhana dan kemudian dibandingkan akurasinya terhadap *dataset* pembanding untuk memastikan bahwa *Indonesian-General-Sentiment-Analysis-Dataset* ini dapat menghasilkan nilai akurasi yang setara jika dibandingkan dengan *dataset* lainnya. *Dataset* yang digunakan sebagai data pembanding adalah *dataset SemEval-2018*. *Dataset* ini dipilih karena *dataset SemEval-2018* adalah *dataset* yang sudah teruji dan digunakan pada konferensi *Semantic Evaluation* yang berjenjang internasional [8].

A. Tahap Pengumpulan Data

Indonesian-General-Sentiment-Analysis-Dataset merupakan sebuah *dataset* yang disusun dengan sumber *tweet* atau kicauan dari media sosial Twitter. Pertama-tama, diambil *tweet* secara acak menggunakan API Twitter. Karena batasan dari API Twitter, *tweet* harus diambil berdasarkan sebuah kata kunci. Untuk memastikan penggunaan kata kunci tidak memengaruhi isi *tweet* yang diambil, kata kunci yang dipilih merupakan kata sambung baku bahasa Indonesia, contohnya “adalah”, “yaitu”, “juga”, dan “seperti”. Proses pengambilan data ini dilakukan dalam bentuk *streaming data* secara kontinu dalam periode waktu tertentu (selama API Twitter masih memperbolehkan) dengan menggunakan sebuah kode bahasa pemrograman Python dengan dibantu oleh *library tweepy*. Pengambilan data berlangsung dari bulan September sampai Desember 2018, dengan data atau *tweet* yang didapat sebanyak 454.559 baris *tweet*.

Kemudian, data atau *tweet* ini masih diseleksi lagi kelayakannya untuk digunakan sebagai *dataset* analisis sentimen bahasa Indonesia. Kriteria kelayakan yang dipilih adalah berbahasa Indonesia (baik bahasa Indonesia baku maupun bahasa daerah) dan memiliki penulisan yang baku (tidak ada huruf yang diubah menjadi angka atau simbol). Proses penyaringan ini dilakukan secara manual dengan cara pembacaan sepintas sampai titik tertentu, baru selanjutnya data dibaca secara lebih dalam lagi pada proses pelabelan data.

Pada sisi proses pelabelan data, diputuskan untuk mengategorikan data menjadi tiga kategori, yaitu positif, negatif, dan netral. Kategori positif digunakan untuk emosi seperti senang, ceria, rileks, santai, dan emosi yang secara garis besar berhubungan dengan kebajikan. Sedangkan kategori negatif digunakan untuk untuk emosi seperti sedih, marah, tertekan, takut, dan emosi-emosi lain yang mengakibatkan penderitaan. Lalu, *tweet-tweet* yang memenuhi kriteria penyaringan tetapi tidak dapat dikategorikan ke dalam dua kategori ini diberi label sebagai kategori netral. Proses pengategorian ini dilakukan secara manual dan subjektif. Setelah itu, didapatkan data akhir berjumlah 10.806 baris *tweet*.

B. Tahap Pre-processing

Proses perbandingan data dilakukan dengan membandingkan akurasi model analisis sentimen yang dihasilkan saat menggunakan masing-masing *dataset* pada tahap pelatihan model. Selain itu, pada data yang diuji juga diberikan beberapa parameter tertentu dari sisi *pre-processing* dan sisi ekstraksi fitur. Pada sisi *pre-processing*, pada masing-masing data, yaitu data *Indonesian-General-Sentiment-Analysis-Dataset* dan data pembanding, yaitu *dataset SemEval-2018* akan diterapkan dua macam teknik *pre-processing*: yang pertama adalah penerapan pembersihan elemen-elemen pengganggu (*noise*) dan yang kedua adalah gabungan penerapan pembersihan *noise*, *stemming*, dan penghilangan *stopwords* (kata sambung).

Selanjutnya, tahap ekstraksi fitur dilakukan dengan menerapkan teknik perubahan kata menjadi vektor (*word2vec*) dan *Term Frequency – Inverse Document Frequency* (TF-IDF). TF-IDF adalah proses ekstraksi fitur berbasis *word2vec* yang menerapkan pemberian nilai pada fitur-fitur berdasarkan frekuensi kemunculan fitur tersebut pada sebuah data (dalam kasus ini *tweet*) dan dibagi dengan frekuensi kemunculan fitur tersebut pada *dataset* secara keseluruhan. TF-IDF digunakan agar fitur-fitur yang sering muncul pada *dataset* berupa kata-kata yang sering digunakan dalam percakapan tidak mengurangi pengaruh dari fitur lainnya dalam pemodelan analisis sentimen.

Secara teori, kedua data yang diuji mengalami peningkatan, baik dari sisi *pre-processing*, dengan menambahkan *stemming* dan penghilangan *stopwords*, maupun dari sisi ekstraksi fitur, dengan mengubah teknik ekstraksi fitur dari *word2vec* standar menjadi TF-IDF [9].

Dari segi *pre-processing*, akurasi pada model analisis sentimen dapat meningkat, karena teknik-teknik *pre-processing* yang diterapkan pada *dataset* dapat mengurangi jumlah fitur pada data. Dengan mengurangi jumlah fitur, algoritme yang digunakan untuk pelatihan model analisis sentimen akan semakin mudah menemukan pola-pola pada data yang berhubungan dengan label pada data itu sendiri. Sedangkan dari sisi ekstraksi fitur, penerapan TF-IDF dapat meningkatkan akurasi model analisis sentimen, karena dengan menggunakan TF-IDF, semua fitur akan diberi bobot yang berbeda satu sama lain. Dengan bobot yang diberikan ini, algoritme pemodelan analisis sentimen dapat menemukan fitur yang paling berpengaruh pada data yang digunakan.

Jika akurasi yang dihasilkan pada model analisis sentimen mengalami peningkatan dan memiliki nilai yang seimbang antara *dataset* utama (*Indonesian-General-Sentiment-Analysis-Dataset*) dengan *dataset pembanding* (*dataset SemEval-2018*), maka dapat diambil kesimpulan bahwa *Indonesian-General-Sentiment-Analysis-Dataset* ini layak untuk digunakan pada pemodelan analisis sentimen.

C. Algoritme Analisis Sentimen

Pada percobaan ini, algoritme analisis sentimen yang digunakan adalah *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), dan *Stochastic Gradient Descent* (SGD). Model algoritme ini dijalankan menggunakan *library scikit-learn* pada bahasa pemrograman Python.

Metode klasifikasi SVM bekerja dengan cara meletakkan fitur-fitur yang diperoleh dari data latihan pada sebuah bidang vektor. Fitur-fitur ini nantinya akan dikelompokkan antara satu dengan yang lain berdasarkan label yang ada pada data pelatihan ini. Dengan demikian, data-data baru dapat diukur nilainya berdasarkan letak data tersebut dibandingkan dengan data-data latihan sebelumnya.

Metode klasifikasi KNN bekerja dengan prinsip yang sama dengan metode klasifikasi SVM, yaitu dengan pertama-tama meletakkan semua fitur pada sebuah bidang vektor. Perbedaannya adalah metode klasifikasi KNN mengukur nilai data baru berdasarkan jarak antara data baru tersebut dengan data latihan yang sudah ditempatkan terlebih dahulu.

Sedangkan metode klasifikasi SGD adalah metode klasifikasi yang menerapkan fungsi *gradient descent* pada sebuah model *standard linear* untuk membuat model tersebut menjadi lebih optimal. SGD melakukan iterasi untuk menemukan titik terendah dalam sebuah gradien (dalam kasus ini, titik terendah merupakan akurasi tertinggi yang dihasilkan oleh model).

Algoritme pemodelan SVM dan KNN adalah dua algoritme yang paling sering digunakan untuk model analisis sentimen pada penelitian-penelitian yang sudah dipublikasi. Hal ini karena (terutama pada SVM) model algoritme ini dapat menganalisis data yang memiliki jumlah fitur yang banyak (data teks) dengan optimal [10]. Dalam melakukan proses pelatihan model, pertama-tama, masing-masing *dataset* dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian. Pada sebuah proses pelatihan model *machine learning*, pembagian data menjadi data pelatihan dan pengujian biasanya dilakukan dengan menggunakan rasio 8:2 atau 80% untuk data pelatihan dan 20% untuk data pengujian.

Selanjutnya, model analisis sentimen yang sudah dilatih diukur akurasinya menggunakan data pengujian yang telah disiapkan tadi. Selain menggunakan akurasi, perbandingan kedua model analisis sentimen ini juga dilakukan dengan melihat nilai-F pada masing-masing model. Akurasi dan nilai-F dapat diukur menggunakan (1) sampai (4).

TABEL IV
TABEL *CONFUSION MATRIX*

	Predicted YES	***Predicted NO***
Actual YES	TP	FN
Actual NO	FP	TN

TABEL V
PERBANDINGAN AKURASI ANTARA *DATASET* UTAMA DENGAN *DATASET* PEMBANDING

Algoritme	**TF-IDF**	***Dataset* utama**		***Dataset* pembanding**	
		Data 1	**Data 2**	**Data 1**	**Data 2**
SVM	no	0,596	0,598	0,522	0,586
	yes	0,614	0,629	0,531	0,607
KNN	no	0,487	0,508	0,375	0,396
	yes	0,523	0,527	0,531	0,543
SGD	no	0,578	0,592	0,531	0,561
	yes	0,621	0,627	0,544	0,573

TABEL VI
PERBANDINGAN NILAI-F ANTARA DATASET UTAMA DENGAN DATASET PEMBANDING

Algoritme	**TF-IDF**	***Dataset* utama**		***Dataset* pembanding**	
		Data 1	**Data 2**	**Data 1**	**Data 2**
SVM	no	0,594	0,596	0,521	0,584
	yes	0,605	0,613	0,551	0,597
KNN	no	0,465	0,477	0,343	0,394
	yes	0,519	0,520	0,520	0,536
SGD	no	0,565	0,587	0,525	0,555
	yes	0,606	0,617	0,537	0,566

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

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

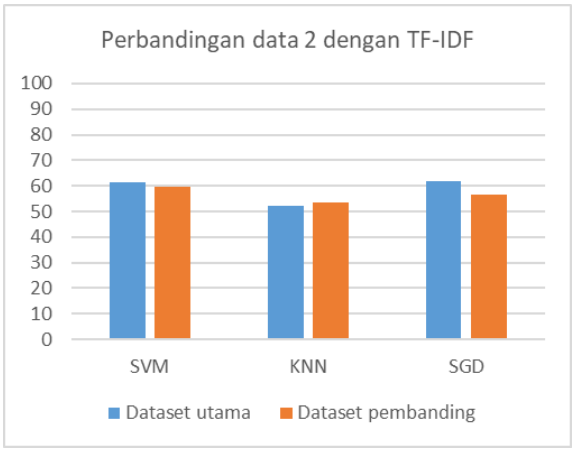
$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Nilai - F = \frac{2 \times presisi \times recall}{presisi+recall}. \quad (4)$$

Sedangkan nilai *True Positive* (TP), *True Negative* (TN), *False Positve* (FP), dan *False Negative* (FN) sendiri dapat diketahui dengan merujuk pada Tabel IV.

Hasil akurasi dan nilai-F dari model analisis sentimennya ditunjukkan pada Tabel V dan Tabel VI. Pada tabel-tabel tersebut, *dataset* utama mengacu pada *dataset* makalah ini, yaitu *Indonesian-General-Sentiment-Analysis-Dataset*, dan *dataset* pembanding mengacu pada *dataset SemEval-2018*. Sedangkan data 1 mengacu pada penerapan pembersihan elemen-elemen penganggu (*noise*) pada data dan data 2 mengacu pada penerapan gabungan antara pembersihan *noise*, *stemming*, dan penghilangan *stopwords*. Gbr. 2 memperlihatkan perbandingan *dataset* utama dan pembanding untuk parameter data 2 dengan TF-IDF

Berdasarkan Tabel V dan Tabel VI, dapat dilihat bahwa rata-rata akurasi dan nilai-F yang dihasilkan dari kedua *dataset*



Gbr. 2 Perbandingan antara *dataset* utama dan *dataset* pembanding untuk paremeter data 2 dengan TF-IDF.

tersebut kurang lebih seimbang satu sama lain. Pada ketiga algoritme, *dataset* utama memiliki nilai akurasi yang sedikit lebih tinggi dibandingkan *dataset* pembanding, dengan perbedaan akurasi antara 4% sampai 12%.

Selanjutnya, jika dilihat pada penerapan teknik *pre-processing* dan TF-IDF, secara konsisten akurasi dan nilai-F telah meningkat pada semua algoritme, yaitu sebesar 0,5% sampai 5% pada bagian penerapan *pre-processing* dan sebesar 5% sampai 15% pada bagian penerapan TF-IDF. Peningkatan akurasi dan nilai-F ini dialami baik pada *dataset* utama maupun *dataset* pembanding. Hal ini konsisten dengan peningkatan akurasi yang terjadi pada penelitian-penelitian analisis sentimen yang sudah dilakukan dan dipublikasi.

Dengan melihat bahwa *dataset* utama dan *dataset* pembanding memiliki akurasi dan nilai-F yang relatif setara, serta melihat peningkatan yang konsisten saat diterapkan teknik *pre-processing* dan TF-IDF, dapat disimpulkan bahwa *Indonesian-General-Sentiment-Analysis-Dataset* ini sudah layak untuk digunakan dalam pelatihan pemodelan analisis sentimen secara umum.

IV. KESIMPULAN

Pada makalah ini, dipublikasikan sebuah *dataset* berbahasa Indonesia, yaitu *Indonesian-General-Sentiment-Analysis-Dataset*, yang merupakan *dataset* berupa *tweet-tweet* dari media sosial Twitter untuk analisis sentimen. Data yang diberikan ini berjumlah 10.806 *tweet* dan telah diberi label dengan tiga polaritas sentimen, yaitu positif, negatif, dan netral. Selain itu, juga dipublikasikan hasil pengukuran berupa perbandingan data *Indonesian-General-Sentiment-Analysis-Dataset* dengan *dataset SemEval-2018* yang dapat digunakan sebagai patokan dasar untuk penelitian-penelitian analisis sentimen di waktu yang akan datang. Beberapa hal yang dapat diambil dari *dataset* ini, serta percobaan yang telah dilakukan adalah sebagai berikut.

Indonesian-General-Sentiment-Analyisis-Dataset memiliki rasio sebesar 2:1:1 antara data yang memiliki label netral dibanding dengan label positif dan negatif. *Dataset* ini diambil dari media sosial Twitter secara acak, sehingga dapat digunakan pada pemodelan analisis sentimen yang topiknya bersifat

umum. Berdasarkan percobaaan yang dilakukan, *Indonesian-General-Sentiment-Analyisis-Dataset* menghasilkan akurasi dan nilai-F yang sepadan dengan *dataset* pembanding, yaitu *dataset SemEval-2018*, dengan perbedaan antara 4% sampai 12%. Kemudian, penerapan teknik *pre-processing* dan TF-IDF pada *Indonesian-General-Sentiment-Analysis-Dataset* telah meningkatkan akurasi model analisis sentimen sesuai dengan teori berdasarkan penelitian-penelitian analisis sentimen lainnya. Peningkatan akurasi yang didapatkan dengan menerapkan teknik *pre-processing* dan TF-IDF adalah sebesar 0,5% sampai 5% dari sisi *pre-processing* dan 5% sampai 15% dari sisi TF-IDF.

Namun, pada makalah ini, dan pada *Indonesian-General-Sentiment-Analysis-Dataset*, masih ada beberapa kekurangan, antara lain belum dimilikinya label polaritas emosi yang spesifik, seperti senang, sedih, amarah, takut, dan lain-lain. Lalu, *dataset* ini tidak dapat digunakan untuk pemodelan analisis sentimen yang spesifik, misalnya analisis sentimen terhadap suatu produk tertentu.

Di masa yang akan datang, diharapkan *Indonesian-General-Sentiment-Analyisis-Dataset* dapat terus ditingkatkan, antara lain dengan menambah jumlah *tweet* pada *dataset*; menambah jenis label polaritas pada *dataset*; meningkatkan kualitas *tweet* yang disajikan pada *dataset*; dan mengembangkan lagi metode pengambilan data dari Twitter. *Dataset* ini dapat diunduh pada tautan http://ugm.id/idsadataset.

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Microsoft Rinna yang telah memberikan bantuan finansial dan fasilitas untuk menjalankan penelitian ini.

REFERENSI

[1] G. Vinodhini dan R. M. Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey," *Int. J. of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 6, hal. 282-292, 2012.

[2] M. Nabil, M. Aly, dan A.F. Atiya, "ATSD: Arabic Sentiment Tweets Dataset," *Conf. on Empirical Methods in Natural Language Processings*, 2015, hal. 2515–2519.

[3] T.A. Lee, D. Moeljadi, Y. Miura, dan T. Ohkuma, "Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets," *Proc. 12th Workshop on Asian Language Resources*, 2016, hal. 123-131.

[4] M.S. Saputri, R. Mahendra, dan M. Adriani, "Emotion Classification on Indonesian Twitter Dataset," *Int. Conf. on Asian Language Processing*, 2018, hal. 90-95.

[5] H. Wijaya, A. Erwin, A. Soetomo, dan M. Galinium, "Twitter Sentiment Analysis and Insight for Indonesian Mobile Operators," *Information Systems Int. Conf.*, 2013, hal. 367-372.

[6] O. Somantri, "Text Mining Untuk Klasifikasi Kategori Cerita Pendek Menggunakan Naive-Bayes (NB)," *Jurnal Telematika*, Vol. 12, No. 1, hal. 7-12, 2017.

[7] Franky dan R. Manurung, "Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews," *Proc. of the Int. Conf. on Advanced Computational Intelligence and Its Applications 2008 (ICACIA 2008)*, 2008, hal. 1-6.

[8] S.M. Mohammad, M. Salameh, F. Bravo-Marquez, dan S. Kiritchenko, "SemEval-2018 Task 1: Affects in Tweets," *Proc. of the 12th Int. Workshop on Semantic Evaluation (SemEval-2018)*, 2018, hal. 1-17.

[9] E. Haddi, X. Liu, dan Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science*, Vol. 17, hal. 26-32, 2013.

[10] R.H. Mohammad dan A. Ahmad, "Sentiment Analysis on Twitter Data using KNN and SVM," *Int. J. of Advanced Computer Science and Applications*, Vol. 8, No. 6, hal. 19-25, 2017.