

ANALISIS SENTIMEN CALON GUBERNUR JAWA TIMUR 2018 DENGAN METODE NAIVE BAYES CLASSIFIER

Ghulam Asrofi Buntoro

Teknik Informatika,
Fakultas Teknik, Universitas Muhammadiyah Ponorogo, Jl. Budi Utomo No. 10, Ponorogo, 63471,
Indonesia

E-mail: ghulamasrofibuntoro@gmail.com

Abstract

The East Java Governor Election 2018 is also felt in the virtual world especially Twitter. All people freely argue about their respective governor candidates, memorandum raises many opinions, not only positive or neutral also negative opinions. Media growth is so rapid, revealing a lot of online media from the news media to social media. Today's social media is not only used of friendship, but also for other activities. Promos of trading or buying and selling, until political party promos or campaigns of candidates for regents, governors, legislative candidates until presidential candidates. The research objective is to conduct a method of Sentiments Analysis for Governor candidates East Java 2018 in twitter with optimal and maximum optimization. While the benefits are to help the community conduct research on opinions on twitter which contains positive, neutral or negative sentiments. Sentiments Analysis for Governor candidates East Java 2018 in twitter using non-conventional processes that save costs, time and effort. The results of Khofifah's dataset are 77% accuracy, 79.2% precision, 77% recall, 98.6% TP rate and 22.2% TN rate. For the results of GusIpul dataset, accuracy is 76%, precision 74.4%, recall 76%, the TP rate is 93.8% and the TN rate is 52.9%.

Keywords: *sentiment analysis, governor, east java, naïve bayes classifier*

Abstrak

Pemilihan Gubernur Jawa Timur yang dilangsungkan pada tahun 2018 keramaiannya juga dirasakan di dunia maya khususnya di Twitter. Semua orang bebas berpendapat atau beropini tentang calon Gubernurnya masing-masing, sehingga memunculkan banyak opini, tidak hanya opini yang positif atau netral tapi juga negatif. Media sosial sekarang ini tidak hanya dimanfaatkan sebagai sarana pertemanan atau mencari teman, tapi juga digunakan untuk kegiatan lain. Promo dagangan atau jual beli, sampai promo partai politik atau kampanye calon-calon Bupati, Gubernur, Calon Legislatif sampai Calon Presiden. Tujuan penelitian adalah Untuk membangun suatu metode analisis sentimen calon Gubernur Jawa Timur 2018 pada media sosial Twitter dengan akurasi paling optimal dan maksimal. Sedangkan manfaatnya untuk membantu masyarakat melakukan riset atas opini masyarakat di twitter yang mengandung sentimen positif, netral atau negatif. Analisis sentimen calon Gubernur Jawa Timur 2018 pada media sosial twitter ini melalui proses nonkonvensional yang hemat biaya, waktu dan tenaga. Hasil dataset Khofifah akurasi 77% %, nilai presisi 79,2%, nilai *recall* 77%, nilai *TP rate* 98,6% dan nilai *TN rate* 22,2%. Untuk hasil dataset GusIpul akurasi 76%, nilai presisi 74,4%, nilai *recall* 76%, nilai *TP rate* 93,8% dan nilai *TN rate* 52,9%.

Kata Kunci: *analisis sentimen, gubernur, jawa timur, naïve bayes classifier*

1. PENDAHULUAN

Pemilihan Gubernur Jawa Timur tahun 2018 keramaiannya tidak hanya dirasakan di dunia nyata, di dunia maya khususnya media sosial Twitter orang-orang mulai membicarakan calon Gubernur jagoannya. Tahapan Pemilihan Gubernur Jawa Timur tahun 2018 sudah di umumkan Komisi Pemilihan Umum (KPU) Jawa Timur [1]. Sejak masih tahap pendaftaran sampai ditetapkannya siapa calon Gubernur Jawa Timur 2018 yang bakal maju di PILKADA Jatim 2018, nama-nama calonnya sudah mulai banyak

diperbincangkan. Dunia maya yang begitu bebas dan sulit dikontrol, membuat semua orang bebas berpendapat atau beropini tentang calon Gubernur jagoannya masing-masing, memunculkan banyak opini-opini masyarakat, tidak hanya opini yang positif atau netral tapi juga yang negatif.

Perkembangan dunia informasi yang begitu cepat, memunculkan banyak media online, mulai informasi berita sampai media sosial atau pertemanan, media sosial mulai dari Facebook, Twitter, Path, Instagram, Google+ dan masih banyak lagi. Pada tahun 2015 Indonesia menjadi nomer dua pengguna aktif Media sosial Twitter

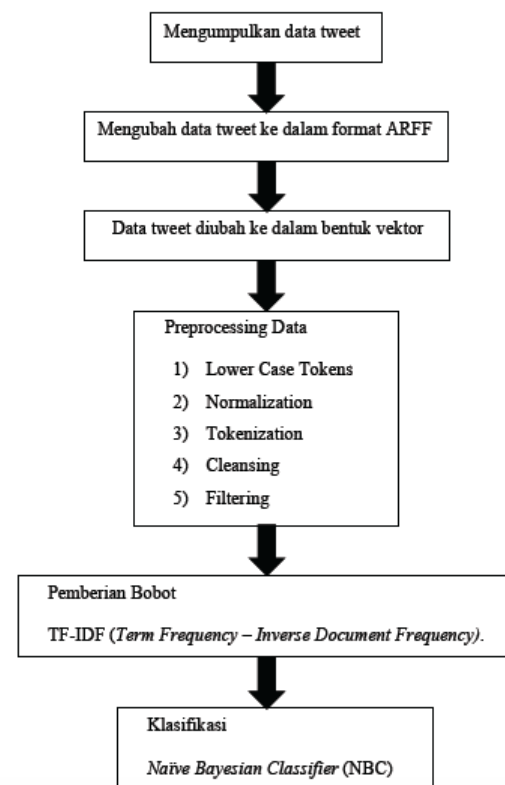
dari Jumlah Total Pengguna Twitter yang Aktif seluruh dunia sampai saat ini 330 juta, Jumlah Tweet yang dikirim per hari untuk seluruh dunia sekitar 500 juta dan Jumlah Pengguna Harian yang Aktif di seluruh dunia sekitar 100 juta [2].

Kemeriahan Pilkada Jawa Timur 2018 sudah dirasakan di media sosial khususnya Twitter, media sosial khususnya twitter sekarang ini menjadi tempat yang sangat penting untuk calon dan tim suksesnya melakukan kampanye. Tim sukses sebuah pasangan calon Gubernur atau kepala daerah sekarang ini misalnya, mereka sampai menghalalkan segala cara dalam mengkampanyekan calonnya, terbukti pada setiap masa kampanye banyak *Black Campaign* terutama di media social terhadap seorang calon. Jaman sekarang kampanye atau pencitraan tidak hanya dilakukan di dunia nyata tapi juga merambah di dunia maya. Media sosial khususnya *Twitter* sekarang ini menjadi salah satu tempat kampanye yang efektif dan efisien.

Analisis sentimen masih bagian dari penelitian *opinion mining* yaitu proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini [3]. Pada penelitian ini Analisis sentimen dilakukan untuk melihat dan mengambil informasi sebuah opini seseorang dalam Bahasa Indonesia di Twitter yang ditujukan kepada calon Gubernur Jawa Timur 2018, apakah opini itu masuk kategori opini positif, netral atau negatif. Metode pembobotan menggunakan *Lexicon Based Features* dan untuk menguji akurasi analisis sentimen dalam penelitian ini menggunakan dua metode yaitu Metode *Naïve Bayes Classifier* (NBC), karena metode tersebut banyak digunakan untuk analisis sentimen dengan hasil akurasi yang cukup bagus [4].

2. Metode Penelitian

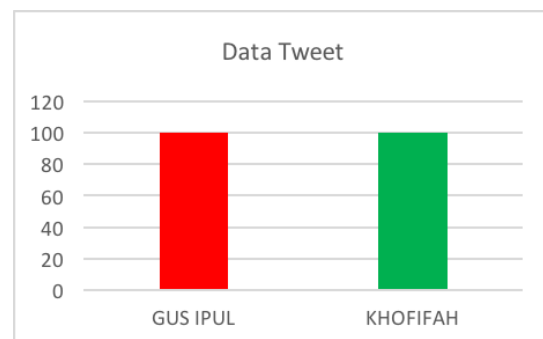
Langkah-langkah penelitian sesuai dengan alur penelitian adalah sebagai berikut:



Gambar 1. Alur Penelitian

A. Mengumpulkan data *tweet*

Data *tweet* diambil dengan metode *Crawling* dari media sosial *Twitter*. Data yang diambil hanya *tweet* dalam bahasa Indonesia, yaitu 100 *tweet* dengan kata kunci masing-masing kandidat. Data diambil secara acak baik dari user biasa ataupun media *online* di *Twitter*.



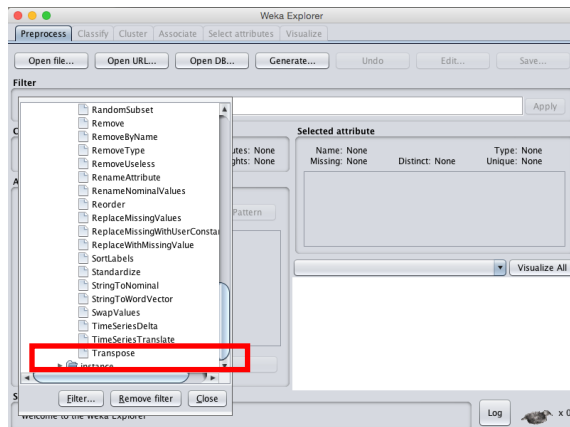
Gambar 2. Data *tweet*

B. Mengubah *data tweet* ke dalam format *ARFF*

Data *tweet* yang dikumpulkan yang berbentuk teks, kemudian dijadikan file *ARFF* (Attribute Relation File Format)[12]. Untuk pembuatan file *ARFF* dilakukan dengan cara manual.

C. *Data tweet* diubah ke dalam bentuk vektor

Data *tweet* yang sudah berbentuk ARFF, kemudian diubah ke dalam bentuk vector[14]. Cara mengubah data *tweet* menjadi vector dengan cara memilih *StringToWordVector* pada tool WEKA.



Gambar 3. mengubah data *tweet* menjadi vector

Hasilnya dari pengubahan file ARFF menjadi bentuk vector dapat dilihat pada Gambar 3.

184: masyarakat	185: masyarakatny	186: may	187: melanjutkan	188: memimpin	189: memperhatikan	190: mendukung
Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
2.092108	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0

Gambar 4. Contoh vector dari data *tweet*

Pada Gambar 4. bagian yang ada dalam kotak merah merupakan kata-kata yang ada dalam data *tweet*. Untuk tiap baris mewakili tiap-tiap data *tweet*. Pada baris ke-1 dalam kotak berwarna biru dapat dilihat bahwa kata “masyarakat” memiliki nilai 2.092108 dan kata “mendukung” memiliki nilai 2.33402. Sedangkan yang lainnya bernilai 0,0 itu artinya ada kata “masyarakat” dan “mendukung” terdapat pada data *tweet* ke-1.

D. Preprocessing Data

Melakukan *preprocessing* data *tweet*. *Preprocessing* meliputi lower case tokens, normalization, tokenization *cleansing*, dan *filtering*. Semua tahapan preprocessing data dilakukan menggunakan tools WEKA 3.8.1. Tahapan-tahapan preprosesing yang dilakukan sebagai berikut:

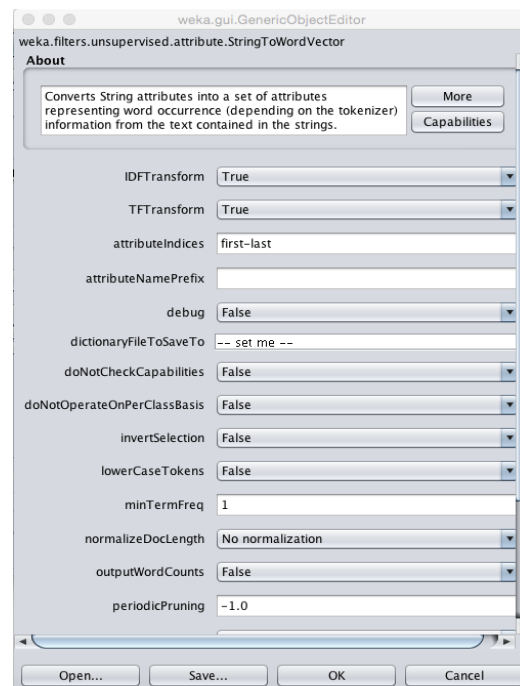
- 1) *Lower Case Tokens* berfungsi untuk membuat data *tweet* menjadi huruf kecil semua, misalnya dari huruf capital menjadi huruf kecil biasa.
- 2) *Normalization* dilakukan untuk menormalkan kata-kata yang tidak baku, misalnya kata yang disingkat, kata gaul atau alay ke kata yang bakunya di Bahasa Indonesia.
- 3) *Tokenization* dilakukan untuk memecah *tweet* menjadi beberapa kata atau kumpulan

kata yang berdiri sendiri. Penelitian ini menggunakan 3 metode tokenization, yaitu *unigram*, *bigram*, dan *n-gram* dengan nilai n minimum = 1 dan n maksimum = 3. Proses tokenization menggunakan menu yang ada dalam WEKA. Pada *tokenizer* pilih *choose* kemudian pilih metode tokenization yang akan digunakan.

- 4) *Cleansing*, yaitu proses menghapus simbol-simbol yang kurang penting dalam data *tweet* yang bisa mengganggu proses klasifikasi nantinya. Proses ini dilakukan dengan menggunakan menu *delimiters* pada WEKA.
- 5) *Filtering* dilakukan untuk menghapus kata-kata yang kurang penting atau kurang berpengaruh terhadap proses klasifikasi nantinya. Proses ini dilakukan dengan menggunakan *stopword list*. *Stopword list* yang digunakan dalam penelitian ini adalah *stopword list* WEKA dan *stopword list* Bahasa Indonesia yang dibuat oleh Tala[13].

E. Pemberian Bobot

Pada tahap selanjutnya adalah memberikan bobot pada tiap-tiap kata (*term*). Pembobotan dilakukan untuk mendapatkan nilai dari kata yang hasil diekstrak. Metode yang digunakan untuk pemberian bobot dalam penelitian ini adalah TF-IDF (*Term Frequency – Inverse Document Frequency*). Karena metode ini paling cocok jika dikombinasi dengan metode klasifikasi *Naive Bayes Classifier* (NBC).



Gambar 5. TF-IDF (*Term Frequency – Inverse Document Frequency*)



F. Klasifikasi

Pada penelitian ini, tahap klasifikasi data menggunakan bantuan tools WEKA 3.8.1. Metode klasifikasi yang digunakan dalam penelitian ini adalah *Naïve Bayes Classifier (NBC)*. *Naïve Bayes Classifier (NBC)* adalah metode *classifier* berdasarkan probabilitas dan *Teorema Bayes* dengan asumsi bahwa setiap variabel X bersifat bebas (*independence*). Dengan kata lain, *Naïve Bayesian Classifier (NBC)* mengasumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan beradaan atribut (variabel) yang lain. Berikut adalah Rumusnya.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{1}$$

Pada proses klasifikasi data diuji dengan menggunakan metode *10 fold cross validation*[15]. Jadi dataset akan dibagi menjadi dua, yaitu 10 bagian dengan 9/10 bagian digunakan untuk proses *training* dan 1/10 bagian digunakan untuk proses *testing*. *Iterasi* berlangsung 10 kali dengan variasi data *training* dan *testing* menggunakan kombinasi 10 bagian data.

Pengujian	Dataset									
1	■									
2		■								
3			■							
4				■						
5					■					
6						■				
7							■			
8								■		
9									■	
10										■

Gambar 6. Ilustrasi 10 fold cross validation

G. Evaluasi Hasil

Melakukan evaluasi performa *TP rate*, *FP rate*, *Precision*, *Recall* dan *F-measure* dari eksperimen yang telah dilakukan. Evaluasi dilakukan dengan menggunakan *Confusion Matrix* yaitu *true positive rate (TP rate)*, *true negative rate (TN rate)*, *false positive rate (FP rate)* dan *false negative rate (FN rate)* sebagai indikator. *TP rate* adalah persentase dari kelas positif yang berhasil diklasifikasi sebagai kelas positif, sedangkan *TN rate* adalah persentase dari kelas negatif yang berhasil diklasifikasi sebagai kelas negatif. *FP rate* adalah kelas negatif yang

diklasifikasi sebagai kelas positif. *FN rate* adalah kelas positif yang diklasifikasi sebagai kelas negatif[16].

Table 1
CONFUSION MATRIX

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positif	c	d

3. HASIL DAN PEMBAHASAN

Hasil berupa data penelitian yang telah diolah dan dituangkan dalam bentuk tabel, grafik, foto atau gambar. Pembahasan berisi hasil analisis dan hasil penelitian yang dikaitkan dengan struktur pengetahuan yang telah mapan (tinjauan pustaka yang diacu oleh penulis), dan memunculkan teori-teori baru atau modifikasi terhadap teori – teori yang telah ada.

Dataset pada penelitian ini menggunakan format ARFF yang dikumpulkan dari Twitter dengan metode *Crawling* dari media sosial Twitter. Data yang diambil hanya *tweet* dalam bahasa Indonesia, yaitu *tweet* dengan kata kunci GusIpul untuk Calon Gubernur Jawa Timur 2018 Syaifullah Yusuf dan Khofifah untuk Calon Gubernur Jawa Timur 2018 Khofifah Indar Parawansa. Data diambil secara acak baik dari user biasa ataupun media *online* di *Twitter*. Dataset yang digunakan sebanyak 200 Tweet, data dibagi secara seimbang (*balanced*) setiap kelasnya, karena dengan data yang tidak seimbang (*imbalanced*), klasifikasi yang dibangun memiliki kecenderungan untuk mengabaikan *minority class* [11]. Data dibagi menjadi GusIpul 100 Tweet, dan Khofifah 100 Tweet. Pemberian label dilakukan dengan metode *Lexicon Based Features* dan bantuan ahli Bahasa Indonesia.

Hasil Analisis Sentimen Calon Gubernur Jawa Timur 2018 menggunakan metode *Lexicon Based Features* dengan tiga class atribut yaitu positif, netral dan negatif.

Tabel 4.

Hasil Analisis Sentimen metode *Lexicon Based Features*

Sentimen	GusIpul	Khofifah
Positif	65	72
Netral	18	19
Negatif	17	9

Untuk mengetahui akurasinya, Analisis Sentimen calon Gubernur DKI Jakarta 2017 dengan metode *Lexicon Based Features*



diklasifikasi menggunakan metode *Naïve Bayes Classifier (NBC)* dengan perangkat lunak WEKA versi 3.8.1. WEKA menggunakan tipe dokumen *Atribut-Relation File Format (ARFF)* sebagai masukan untuk melakukan klasifikasi data.

Hasil dari proses klasifikasi kemudian diuji dengan menggunakan metode *10-fold cross validation*, data dibagi mejadi 10 bagian dengan 9/10 bagian digunakan untuk proses *training* dan 1/10 bagian digunakan untuk proses *testing*. Iterasi berlangsung 10 kali dengan variasi data *training* dan *testing* menggunakan kombinasi 10 bagian data.

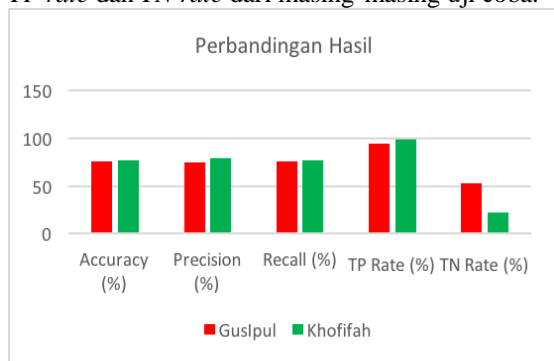
Perbandingan hasil dari metode klasifikasi *Naïve Bayes Classifier (NBC)* dengan dataset Calon Gubernur Jawa Timur 2018 GusIpul dan Khofifah.

Tabel 5.
Perbandingan Hasil Klasifikasi

Calon Gubernur	Accur acy (%)	Preci sion (%)	Reca ll (%)	TP Rate (%)	TN Rate (%)
GusIpul	76	74,4	76	93,8	52,9
Khofifah	77	79,2	77	98,6	22,2

*) Nilai *Precision* dan *Recall* meruapakan nilai rata-rata dari nilai kelas positif dan kelas negatif.

Tabel 5. berisi informasi mengenai nilai akurasi, presisi, *recall*, *TP rate* dan *TN rate* dari masing-masing uji coba yang telah dilakukan. Bagian kolom berisi informasi mengenai Calon Gubernur Jawa Timur 2018. Sedang bagian baris berisi nilai akurasi, presisi, *recall*, *TP rate* dan *TN rate* dari masing-masing uji coba yang telah dilakukan. Dari proses data *preprocessing* menghasilkan sejumlah token yang kemudian digunakan sebagai input sebuah proses klasifikasi. Proses klasifikasi dilakukan menggunakan metode *Naïve Bayes Classifier (NBC)*. Dari proses klasifikasi diperoleh nilai akurasi, presisi, *recall*, *TP rate* dan *TN rate* dari masing-masing uji coba.



Gambar 3. Grafik tingkat akurasi

Dari Gambar 3. dapat dilihat hasil akurasi dari nilai Analisis Sentimen dengan metode *Lexicon Based Features* yang diklasifikasi dengan

metode *Naïve Bayes Classifier (NBC)*. Nilai akurasi dataset Khofifah mencapai 77%, nilai presisi 79,2%, nilai *recall* 77%, nilai *TP rate* 98,6% dan nilai *TN rate* 22,2%. Untuk dataset GusIpul akurasinya mencapai 76%, nilai presisi 74,4%, nilai *recall* 76%, nilai *TP rate* 93,8% dan nilai *TN rate* 52,9%. Dataset Khofifah mendapatkan akurasi paling tinggi karena dari 72 data positif, 71 data berhasil diklasifikasi metode *Naïve Bayes Classifier (NBC)* dengan benar sesuai sentimennya yaitu sentimen positif. Sedangkan untuk dataset GusIpul, dari 65 data positif, 61 data berhasil diklasifikasi metode *Naïve Bayes Classifier (NBC)* dengan benar sesuai sentimennya yaitu sentimen positif. Metode ini cenderung *Naïve Bayes Classifier (NBC)* lebih stabil karena berbasis probabilitas kemunculan kata dalam sebuah kalimat. Nilai akurasi adalah salah satu parameter penilaian dari metode yang telah digunakan, nilai akurasi didapat dari jumlah banyaknya data yang berhasil diklasifikasi dengan benar sesuai kelas sentimennya dari seluruh jumlah data yang diklasifikasi. Nilai akurasi yang tinggi didapat ketika banyak data yang berhasil diklasifikasi dengan benar sesuai kelas sentimennya.

Dari Gambar 3. dapat diketahui juga nilai *Precision* dan *Recall*. Nilai *Precision* mengikuti nilai akurasi, nilai akurasi semakin tinggi maka akan diikuti nilai *Precision* yang tinggi juga, begitu sebaliknya. Nilai *Precision* adalah jumlah data positif yang benar diklasifikasi sebagai data positif dibagi total data yang diklasifikasi sebagai data positif. Sedangkan Nilai *recall* adalah jumlah data positif yang benar diklasifikasi sebagai data positif dibagi jumlah data positif sebenarnya.

Dari Gambar 3. kita juga dapat mengetahui nilai *TP Rate* dan *TN Rate*. *TP Rate* adalah nilai data positif yang diklasifikasi dengan benar sesuai kelas sentimennya, yaitu positif. Nilai *TN Rate* adalah nilai data sentimen yang diklasifikasi dengan benar sesuai kelas sentimennya, yaitu negatif.

Dari penelitian yang telah dilakukan, terbukti metode klasifikasi *Naïve Bayes Classifier (NBC)* dapat digunakan untuk mengklasifikasi sentimen-sentimen (positif, netral dan negatif) *tweet* Bahasa Indonesia terhadap Calon Gubernur Jawa Timur 2018. Selanjutnya, dataset Khofifah mendapat akurasi lebih tinggi dibandingkan dengan akurasi dataset GusIpul, dengan akurasi 77% dibanding 76%. Dalam dataset Khofifah, sentimen positifnya paling banyak yaitu 72 dan sentimen negatifnya hanya 9. Sedangkan didataset GusIpul sentimen positifnya 65 dan sentiment negatifnya 17. Jadi bisa disimpulkan, di media sosial Twitter Khofifah lebih banyak dicintai dibandingkan dengan GusIpul. Meskipun menghasilkan akurasi

yang cukup tinggi, model yang dibangun ini masih melakukan sedikit kesalahan klasifikasi untuk dataset yang pembagian sentimennya tidak seimbang. Karena dengan menggunakan dataset yang tidak seimbang akan menyebabkan data *minority class* yang salah diklasifikasi sebagai data *majority class* (Kohavi, 1998). Pada akhirnya menjadikan selisih nilai menjadi besar.

4. KESIMPULAN

Dari penelitian yang telah dilakukan, terbukti metode klasifikasi *Naïve Bayes Classifier (NBC)* dapat digunakan untuk mengklasifikasi sentimen-sentimen (positif, netral dan negatif) *tweet* Bahasa Indonesia terhadap Calon Gubernur Jawa Timur 2018. Selanjutnya, dataset Khofifah mendapat akurasi lebih tinggi dibandingkan dengan akurasi dataset GusIpul, Hasil dataset Khofifah akurasi 77% %, nilai presisi 79,2%, nilai *recall* 77%, nilai *TP rate* 98,6% dan nilai *TN rate* 22,2%. Untuk hasil dataset GusIpul akurasi 76%, nilai presisi 74,4%, nilai *recall* 76%, nilai *TP rate* 93,8% dan nilai *TN rate* 52,9%. Dalam dataset Khofifah, sentimen positifnya paling banyak yaitu 72 dan sentimen negatifnya hanya 9. Sedangkan didataset GusIpul sentimen positifnya 65 dan sentiment negatifnya 17. Jadi bisa disimpulkan, di media sosial Twitter Khofifah lebih banyak dicintai dibandingkan dengan GusIpul. Terbukti Analisis Sentimen dapat digunakan untuk mengetahui Sentimen masyarakat khususnya netizen Twitter terhadap Calon Gubernur Jawa Timur 2018, sehingga membantu masyarakat awam untuk mengetahui sentimen masyarakat lainya terhadap Calon Gubernur Jawa Timur 2018. Untuk penelitian selanjutnya perlu dicoba dikembangkan dengan menggunakan data yang lebih banyak dan Real Time. Perlu dikembangkan juga *stopword list* dan *stemmer* Bahasa Indonesia yang mampu meningkatkan akurasi dalam analisis Sentimen Bahasa Indonesia.

5. REFERENSI

- [1] "KEGIATAN TAHAPAN PILGUB JATIM 2018 MAKIN PADAT, KETUA KPU JATIM AJAK JAGA SOLIDITAS," *KPU PROVINSI JAWA TIMUR*, 08-Feb-2018.
- [2] S. Aslam, "Twitter by the Numbers (2018): Stats, Demographics & Fun Facts," 01-Jan-2018.
- [3] B. Liu, "Sentiment Analysis and Subjectivity.," *Handb. Nat. Lang. Process.*, vol. 2, pp. 627–666, 2010.
- [4] D. S. Pamungkas, et. al (2015). ANALISIS SENTIMENT PADA SOSIAL MEDIA TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER TERHADAP KATA KUNCI "KURIKULUM 2013" *Techno.COM*, Vol. 14, No. 4, November 2015: 299-314.
- [5] G. A. Buntoro, (2016). " Sentiment Analysis Candidates of Indonesian Presiden 2014 with Five Class Attribute" in *International Journal of Computer Applications* (0975 – 8887).
- [6] N. Adiyasa, "Analisis Sentimen Pada Opini Berbahasa Indonesia Menggunakan Pendekatan Lexicon-Based," *Catatan Kecil*, 2011. [Online]. Available: <http://adiyasan.wordpress.com/2013/02/08/sentiment-analysis-menggunakan-pendekatan-lexicon-based/>. [Accessed: 10-Mar-2018].
- [7] A. F. Hadi and M. Hasan, "TEXT MINING PADA MEDIA SOSIAL TWITTER STUDI KASUS: MASA TENANG PILKADA DKI 2017 PUTARAN 2." Seminar Nasional Matematika dan Aplikasinya, Universitas Airlangga 2017.
- [8] J. Ariawan, "Data Preprocessing." [Online]. Available: <https://www.google.com/search?q=apa+itu+noise+dalam+data+mining&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official> [Accessed: 10-Mar-2018].
- [9] G. A. Buntoro, (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, 2(1).
- [10] Ian H. Witten. (2013) *Data Mining with WEKA*. Department of Computer Science University of Waikato New Zealand.
- [11] Kohavi, & Provost. (1998) *Confusion Matrix* http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html