KILAT
Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hashtag #2019gantipresiden

Arini[1]; Luh Kesuma Wardhani[2]; Dimas Octaviano[3]

[1, 2, 3]Program Studi Teknik Informatika, Fakultas Sains dan Teknologi
UIN Syarif Hidayatullah Jakarta
[1]arini@uinjkt.ac.id
[2]luhkesuma@uinjkt.ac.id
[3]dimasocta@mhs.uinjkt.ac.id

ABSTRACT

Towards an election year (elections) in 2019 to come, many mass campaign conducted through social media networks one of them on twitter. One online campaign is very popular among the people of the current campaign with the hashtag #2019GantiPresiden. In studies sentiment analysis required hashtag 2019GantiPresiden classifier and the selection of robust functionality that mendaptkan high accuracy values. One of the classifier and feature selection algorithms are Naive Bayes classifier (NBC) with Tri-Gram feature selection Character & Term-Frequency which previous research has resulted in a fairly high accuracy. The purpose of this study was to determine the implementation of Algorithm Naive Bayes classifier (NBC) with each selection and compare features and get accurate results from Algorithm Naive Bayes classifier (NBC) with both the selection of the feature. The author uses the method of observation to collect data and do the simulation. By using the data of 1,000 tweets originating from hashtag # 2019GantiPresiden taken on 15 September 2018, the author divides into two categories: 950 tweets as training data and 50 tweets as test data where the labeling process using methods Lexicon Based sentiment. From this study showed Naïve Bayes classifier algorithm accuracy (NBC) with feature selection Character Tri-Gram by 76% and Term-Frequency by 74%,the result show that the feature selection Character Tri-Gram better than Term-Frequency.

Keywords: *Comparison, Sentiment Analysis, Twitter, Naïve Bayes classifier, Feature Selection, Tri-Gram Character, Term-Frequency, Lexicon Based, Accuracy*

ABSTRAK

Menuju tahun pemilu (pemilihan umum) pada 2019 mendatang, banyak kampanye yang dilakukan secara massal melalui jejaring media sosial salah satunya di twitter. Salah satu kampanye online yang sangat populer dikalangan masyarakat saat ini adalah kampanye dengan hashtag #2019GantiPresiden. Dalam penelitian sentimen analisis hashtag #2019GantiPresiden diperlukan classifier & seleksi fitur yang tangguh agar mendaptkan nilai akurasi yang tinggi. Salah satu classifier & seleksi fitur tersebut adalah Algoritma Naïve Bayes Classifier (NBC) dengan seleksi fitur Tri-Gram Character & Term-Frequency dimana pada penelitian sebelumnya menghasilkan akurasi yang cukup tinggi. Tujuan dari penelitian ini adalah mengetahui implementasi dari Algoritma Naïve Bayes Classifier (NBC) dengan masing-masing seleksi fitur serta mendapatkan dan membandingkan hasil akurasi dari Algoritma Naïve Bayes Classifier (NBC) dengan kedua seleksi fitur tersebut. Penulis menggunakan metode observasi untuk mengumpulkan data dan melakukan simulasi. Dengan menggunakan 1.000 data tweet yang bersumber dari hashtag #2019GantiPresiden yang diambil pada 15 September 2018, penulis membagi menjadi dua kategori yaitu 950 tweet sebagai data latih dan 50 tweet sebagai data uji dimana proses labelling sentimen menggunakan metode Lexicon Based. Dari penelitian ini diperoleh hasil akurasi Algoritma Naïve Bayes Classifier (NBC) dengan seleksi fitur Tri-Gram Character sebesar 76% &

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

Term-Frequency sebesar 74%,, hal ini menunjukkan bahwa menggunakan seleksi fitur Tri-Gram Character lebih baik dari Term-Frequency.

Kata kunci: *Perbandingan, Analisis Sentimen, Twitter, Naïve Bayes Classifier, Seleksi Fitur, Tri-Gram Character, Term-Frequency, Lexicon Based, Akurasi*

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

1. PENDAHULUAN

Dalam kehidupan bermasyarakat, mengungkapan sentimen/pendapat kepada orang lain telah menjadi suatu aktivitas setiap harinya. Menurut KBBI, sentimen adalah pendapat atau pandangan yang didasarkan pada perasaan yang berlebih-lebihan terhadap sesuatu (bertentangan dengan pertimbangan pikiran). Seiring juga dengan berkembangnya kemajuan teknologi, pada saat ini masyarakat pun dapat mengungkapkan sentimen/pendapat/pandangannya ke publik melalui media sosial yang sedang trend digunakan seperti Facebook, Youtube, Twitter, dan Instagram. Media sosial sendiri merupakan sebuah media online yang para pengunanya bisa melakukan berbagai macam hal dengan mudah seperti berpartisipasi, berbagi, jejaring sosial, wiki, forum, dunia virtual, dan lain sebagainya [1].

Berdasarkan data dari PT. Bakrie Telecom, media sosail Twitter saat ini memiliki 19,5 juta pengguna di Indonesia dari total pengguna global sebanyak 500 juta. Twitter sendiri menjadi salah satu jejaring sosial paling besar di dunia sehingga mampu meraup keuntungan mencapai USD 145 juta [2]. Kebanyakan pengguna Twitter di Indonesia adalah konsumen, yaitu yang tidak memiliki blog atau tidak pernah meng-upload video di Youtube namun sering update status di Twitter dan Facebook. Twitter dapat menjadi sumber data pendapat dan sentimen masyarakat yang mana data tersebut dapat digunakan secara efisien untuk pemasaran maupun studi sosial [3]. Twitter telah menjadi salah satu jejaring sosial media yang sering digunakan sebagai alat komunikasi, media untuk promosi, bahkan kampanye politik terlebih menjelang pemilihan umum (pemilu) pada tahun 2019.

Pada saat ini, sebagian besar kampanye politik yang dilakukan para politisi atau tokoh publik menggunakan/memanfaatkan media sosial agar dapat meningkatkan popularitas mereka dengan cepat sehingga mendapatkan banyak simpatisan khususnya dari masyarakat dunia maya (netizen) dan pada akhirnya kembali ke tujuan awal dari para politisi/tokoh publik tersebut berkampanye, yaitu agar bisa memenangkan pemilihan umum (pemilu). Pemilihan umum (pemilu) sendiri merupakan salah satu proses untuk memperjuangkan kepentingan politik dalam bentuk proses seleksi terhadap lahirnya wakil rakyat dan pemimpin dalam rangka perwujudan demokrasi, karena pemilihan umum merupakan suatu rangkaian kegiatan politik untuk menampung kepentingan rakyat, yang kemudian dirumuskan dalam berbagai bentuk kebijakan. Melalui Twitter, para politisi pun kerap memanfaatkan sentiment/pandangan masyarakat yang sedang berkembang untuk dapat masuk dan memperkenalkan program-program mereka agar terpilih pada pemilihan umum (pemilu) di tahun 2019 [4].

Berbicara mengenai banyaknya sentimen masyarakat di dunia maya (netizen) terutama pada twitter, ada satu tagar kampanye *online* yang menjadi trending *topic* dan masih banyak dibicarakan (*trend*) oleh kalangan masyarakat dunia maya (*netizen*) yaitu (hashtag) #2019GantiPresiden.

Tagar (*hashtag*) #2019GantiPresiden merupakan sebuah kampanye terbuka di media sosial Twitter dari pihak oposisi kepada pihak pemerintah yang bertujuan untuk mengganti presiden yang saat ini menjabat (Jokowi) dengan calon presiden lain pada pemilu 2019. Tagar (*hashtag*) #2019GantiPresiden menjadi sangat ramai di twitter karena banyak akun yang asli, palsu (*fake*), maupun akun bayaran (*buzzer*) yang ikut mengirimkan *tweet* menggunakan tagar (*hashtag*) #2019GantiPresiden setiap jam, menit, bahkan detik. Terlepas dari pro serta kontra yang terdapat pada tagar (*hashtag*) #2019GantiPresiden, analisis sentimen terhadap tagar (*hashtag*) tersebut dapat diteliti untuk mengetahui seberapa besar presentase sentimen positif, negatif dan netral dari *tweet* dengan tagar (*hashtag*) #2019GantiPresiden. Analisis sentimen (sentiment analysis) atau biasa disebut juga *opinion mining* bertujuan untuk menganalisis, memahami, mengolah, dan mengestrak

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

data tekstual yang berupa opini terhadap entitas seperti organisasi dan topik tertentu agar mendapatkan suatu informasi [6].

Dalam melakukan analisis sentimen, diperlukan suatu algoritma *classifier* & seleksi fitur yang mumpuni agar didapatkan hasil akurasi yang maksimal. Salah satu algoritma *classifier* yang paling sering dipakai dalam melakukan analisis sentimen adalah Algoritma *Naïve Bayes Classifier* (NBC), dimana algoritma *Naïve Bayes* memprediksi peluang terjadinya kejadian di masa depan berdasarkan data yang ada sebelumnya sehingga memiliki hasil akurasi yang cukup tinggi [7].

Hasil akurasi yang cukup tinggi dari Algoritma *Naïve Bayes Classifier* (NBC) dapat lebih ditingkatkan lagi dengan menggunakan seleksi fitur (*feature selection*). Dalam beberapa penelitian, seleksi fitur N-Gram yang mana cara kerjanya memotong sub-urutan n karakter dari dokumen yang diberikan merupakan seleksi fitur yang banyak digunakan dengan Algoritma *Naïve Bayes Classifier* (NBC) untuk meningkatkan hasil akurasi. Tidak hanya N-Gram, beberapa peneliti juga menggunakan seleksi fitur *Term Frequency* (TF) yang juga dapat meningkatkan hasil akurasi dari *Algoritma Naïve Bayes Classifier* (NBC) [7]. Dalam penelitian lainnya yaitu [8], [9],[10] dan [11], sehingga beberapa hal yang menjadi fokus pada penelitian ini adalah sebagai berikut:

1. Analisis sentimen adalah data dari *Twitter* dengan tagar (*hashtag*) #2019GantiPresiden.
2. Menggunakan Algoritma *Naïve Bayes* dalam klasifikasi sentimen, dengan alasan dalam penelitian sebelumnya, *Naïve Bayes* memiliki tingkat akurasi yang besar, yaitu rata-rata 90%.
3. Menggunakan dua seleksi fitur, yaitu N-Gram *Character* dimana n = 3 atau disebut juga dengan Tri-Gram *Character* dan *Term frequency*.
4. Menggunakan metode *Lexicon Based* untuk mengklasifikasi sentimen pada data latihnya
5. Mengukur tingkat akurasi seleksi fitur antara Algoritma *Naïve Bayes* dengan Tri-Gram *Character* dan *Term Frequency*.

Text mining merupakan suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi [12]. *Text mining* dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian atau pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar. Dalam memberikan solusi, *text mining* mengadopsi dan mengembangkan banyak teknik dari bidang lain, seperti *Data mining, Information Retrieval*, Statistik dan Matematik, *Machine Learning*, *Linguistic*, *Natural Languange Processing* (NLP), dan *Visualization*. Tujuan dari *Text Mining* adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen, tetapi tujuan utama text mining adalah mendukung proses knowledge discovery pada koleksi dokumen yang besar. Adapun tugas khusus dari *text mining* antara lain yaitu pengkategorisasian teks (*text categorization*) dan pengelompokkan teks (*text clustering*) [12].

Analisis Sentimen (*opinion mining*) merupakan suatu bidang yang luas dari pengolahan bahasa alami, komputasi linguistik dan *text* mining dimana memiliki tujuan dalam menganalisa pendapat, sentimen, evaluasi, sikap, penilaian dan emosi seseorang apakah pembicara atau penulis berkenaan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu [6].

Algoritma *Naïve Bayes Classifier* (NBC) memberi nilai target kepada data baru menggunakan nilai V_(map), yaitu nilai kemungkinan tertinggi dari seluruh anggota himpunan set domain V. Setiap data tweet direpresentasikan dengan pasangan atribut “x_1,x_2,x_3…..x_n “ dimana x_1 adalah kata pertama, x_2 adalah kata kedua, x_3 adalah kata ketiga dan seterusnya. Sedangkan V adalah himpunan kategori sentimen. Pada saat klasifikasi, algoritma akan mencari

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

probabilitas tertinggi dari semua kategori yang diujikan (V_map), dimana persamaanya adalah sebagai berikut [13].

$$V_{map} = \underset{v_j \in V}{armax} \frac{P(x_1, x_2, x_3 \dots x_n | V_j) P(v_j)}{P(x_1, x_2, x_3 \dots x_n)}$$

(1)

Untuk P(x_1,x_2,x_3…x_n) nilainya konstan untuk semua kategori (V_j), sehingga persamaan dapat ditulis sebagai berikut :

$$V_{map} = \underset{v_j \in V}{armax} P(x_1, x_2, x_3 \dots x_n | V_j) P(V_j)$$

(2)

Persamaan diatas dapat disederhanakan menjadi sebagai berikut :

$$V_{map} = \underset{v_j \in V}{armax} \prod_{i=1}^{n} P(x_i | V_j) P(V_j)$$

(3)

Keterangan:

V_j = Kategori komentar j = 1, 2, 3,….n. Dimana dalam penulisan ini j_1 kategori komentar sentimen positif, j_2 = kategori komentar sentimen negatif dan j_3 = kategori komentar sentimen netral

$P(x_i|V_j)$ = Probabilitas x_i pada kategori V_j

$P(V_j)$ = Probabilitas dari (V_j)

Untuk $P(V_j)$ dan $P(x_i|V_j)$ dihitung pada saat pelatihan dimana persamaanya adalah sebagai berikut:

$$P(V_j) = \frac{|docs_j|}{|contoh|}$$

(4)

$$P(x_i|V_j) = \frac{n_k + 1}{n + |kosakata|}$$

(5)

Keterangan:

$|docs_j|$ = Jumlah dokumen setiap kategori j

$|contoh|$ = Jumlah dokumen dari semua kategori

n_k = Jumlah frekuensi kemunculan setiap kata

n = Jumlah frekuensi kemunculan kata dari setiap kategori

$|kosakata|$ = Jumlah semua kata dari semua kategori

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi, *recall*, *precision*, dan *error rate*. Dimana, *precision* mengevaluasi kemampuan sistem untuk

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

menemukan peringkat yang paling relevan, dan didefinisikan sebagai presentase dokumen yang di *retrieve* dan benar-benar relevan terhadap *query*. *Recall* mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai presentase dokumen yang relevan terhadap *query*. *Accuracy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan error rate merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus [14].

Tabel 1. Multiclass Confusion Matrix

		PREDIKSI		
		POSITIF	NEGATIF	NETRAL
AKTUAL	POSITIF	TPos	FPosNeg	FPosNet
	NEGATIF	FNegPos	TNeg	FNegNet
	NETRAL	FNetpos	FNetNeg	TNet

Dengan menggunakan tabel *multiclass confusion matrix* 3x3, maka untuk menghitung tingkat akurasi digunakan rumus :

$$Akurasi = \frac{TPos+TNeg+TNet}{TPos+FPosNeg+FPosNet+FNegPos+TNeg+FNegNet+FNetPos+FNetNeg+TNet} \quad (6)$$

2. METODE PENELITIAN

2.1. Metode Pengumpulan Data

Metode pengumpulan data menggunakan studi lapangan dengan metode observasi dengan mengamati dan mengambil data dari API Twitter tentang *tweet netizen* terhadap *hashtag* #2019GantiPresiden, periode 15 September 2018, dengan fitur *developer* Twitter di akses pada website https://developer.twitter.com/. Data sebanyak 1.000 *tweet* kemudian di simpan ke dalam *database* MySQL. Total data tersebut dibagi menjadi dua dengan rincian 950 data latih dan 50 data uji.

2.2 Metode Simulasi

Langkah-langkah metode simulasi dalam penelitiann ini yaitu:

a. Problem Formulation, pada tahap ini melakukan identifikasi masalah dari hasil penelitian.

b. Conceptual Model, tahap ini peneliti mengidentifikasikan hasil *crawling* yaitu *tweet-tweet* yang bersumber dari *hashtag* #2019GantiPresiden, kemudian dibagi menjadi data latih dan data uji serta dilakukan *preprosesing*. Klasifikasi atau *labelling* sentimen data latih pada penelitian ini menggunakan metode *Lexicon Based.* Data latih diolah dan diproses dengan Algoritma *Naïve Bayes Classifier* dengan dua seleksi fitur, yaitu *Tri-Gram Character* dan *Term Frequency*. Berikutnya akan mencari tingkat akurasi antara Algoritma *Naïve Bayes Classifier* dengan seleksi fitur *Tri-Gram Character* dan *Term Frequency*.

c. Collection of Input/Output Data, pada tahap ini data yang telah diamabil dari *tweet-tweet* API Twitter sebelumnya dijadikan *input* data dalam aplikasi berbasis PHP. Data *tweet* yang telah tersimpan dalam database *MySQL* ini diolah menggunakan Algoritma *Naïve Bayes Classifier* (NBC) dengan seleksi fitur Tri-Gram *Character* dan *Term Frequency*

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

sehingga menghasilkan *output* berupa tingkat akurasi algoritma dengan masing-masing seleksi fitur.

d. Modelling Phase, pada ini penulis membangun representasi yang rinci berdasarkan pemodelan konsep dan masukan/keluaran data yang dikumpulkan. Penulis menggunakan diagram UML (*Unified Modelling Language*) untuk tahap pemodelannya yaitu *use case diagram, activity diagram*, dan *deployment diagram*. Penulis juga melakukan pemodelan berupa perhitungan manual terhadap Algoritma *Naïve Bayes Classifier* (NBC) dengan seleksi fitur Tri-Gram *Character* dan *Term Frequency*.

a. Simulation Phase, pada tahap ini melakukan simulasi perbandingan Algoritma *Naïve Bayes Classifier* (NBC) dengan seleksi fitur *Tri-Gram Character* dan *Term Frequency* menggunakan aplikasi berbasis PHP, simulasi dilakukan dengan mengambil data mentah yang telah tersimpan di *database MySQL* untuk dianalisa dan dikelompokkan sentimennya menggunakan metode *Lexicon Based* lalu diberikan masing-masing seleksi fitur sehingga menghasilkan tingkat akurasi dari masing-masing algoritma.

b. Conclusion (*Verification, Validation, and Experimentation*), pada tahap ini validasi dan verifikasi bertujuan untuk meyakinkan hasil dari perbandingan Algoritma *Naïve Bayes Classifier* (NBC) dengan seleksi fitur Tri-Gram *Character* dan *Term Frequency*, sedangkan pada eksperimen bertujuan untuk mengevaluasi hasil simulasi pada aplikasi.

c. Output Analysis Phase, pada tahap ini melakukan analisis terhadap *output* berdasarkan pengujian yang telah dilakukan, yaitu memfilter data uji berdasarkan *labelling* data latih dengan masing-masing seleksi fitur. Hasil yang di dapatkan dianalisa serta dibandingkan dengan pemodelan sebelumnya, yaitu hasil Algoritma *Naïve Bayes Classifier* (NBC) dengan seleksi fitur *Tri-Gram Character* dan *Term Frequency*, serta menghitung tingkat akurasi dari masing-masing algoritma.

2.3. Perangkat Penelitian

Dalam melakukan penelitian ini, penulis menggunakan perangkat keras (*hardware*) dan persngkat lunak (*software*) sebagai berikut:

1. Perangkat Keras (*hardware*)

Tabel 2. Perangkat Keras (*hardware*)

No.	Perangkat Keras	Spesifikasi
1.	Device	Laptop Asus A43S
2.	Processor	Core i5 @2.00GHz
3.	Monitor	14 inch
4.	VGA	NVIDIA GEFORCE 610M 2.00 GB
5.	Memori (RAM)	8.00 GB
6.	Keyboard dan Mouse	Standard
7.	Speaker	Standard

2. Perangkat Lunak (*software*)

Tabel 3. Perangkat Lunak (*software*)

No.	Perangkat Lunak	Spesifikasi
1.	Sistem Operasi	Windows 10 Profesional 64-bit
2.	Tools	XAMPP
3.	Bahasa Pemrograman	PHP

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

3. HASIL DAN PEMBAHASAN

3.1. Hasil Klasifikasi Sentimen

Berikut pada tabel 4 adalah hasil klasifikasi sentiment dari 50 *tweet* data uji yang diambil dari *Twitter* dengan tagar (*hashtag*) #2019GantiPresiden menggunakan Algoritma *Naïve Bayes* dan seleksi fitur *Term-Frequency* & Tri-Gram *Character* serta metode *Lexicon Based*.

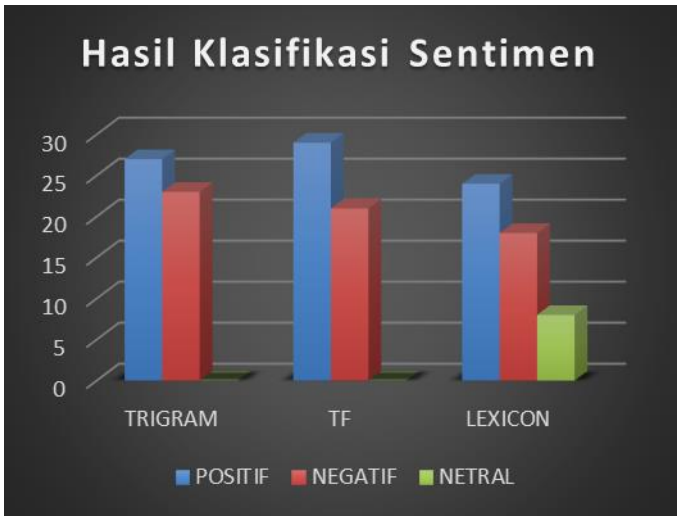
Tabel 4. Hasil Klasifikasi Sentimen Data Uji

HASIL KLASIFIKASI SENTIMEN			
Data ke-	Seleksi Fitur Tri-Gram	Seleksi Fitur *Term Frequency*	Lexicon (actual)
1	NEGATIF	POSITIF	NEGATIF
2	NEGATIF	NEGATIF	NEGATIF
3	NEGATIF	NEGATIF	NEGATIF
4	NEGATIF	NEGATIF	POSITIF
5	NEGATIF	NEGATIF	NEGATIF
6	NEGATIF	NEGATIF	NEGATIF
7	POSITIF	POSITIF	POSITIF
8	POSITIF	POSITIF	POSITIF
9	NEGATIF	POSITIF	NEGATIF
10	POSITIF	POSITIF	POSITIF
11	NEGATIF	NEGATIF	NETRAL
12	POSITIF	POSITIF	POSITIF
13	POSITIF	POSITIF	POSITIF
14	NEGATIF	NEGATIF	NEGATIF
15	POSITIF	POSITIF	POSITIF
16	NEGATIF	NEGATIF	NEGATIF
17	POSITIF	POSITIF	POSITIF
18	NEGATIF	POSITIF	NETRAL
19	NEGATIF	NEGATIF	NETRAL
20	NEGATIF	NEGATIF	NEGATIF
21	POSITIF	POSITIF	POSITIF
22	POSITIF	POSITIF	POSITIF
23	NEGATIF	NEGATIF	NEGATIF
24	POSITIF	NEGATIF	NEGATIF
25	POSITIF	POSITIF	POSITIF
26	NEGATIF	POSITIF	NEGATIF
27	NEGATIF	POSITIF	POSITIF
28	POSITIF	POSITIF	POSITIF
29	NEGATIF	NEGATIF	NETRAL
30	POSITIF	POSITIF	POSITIF
31	POSITIF	NEGATIF	NETRAL
32	NEGATIF	NEGATIF	NEGATIF
33	POSITIF	POSITIF	POSITIF
34	NEGATIF	NEGATIF	NEGATIF
35	POSITIF	POSITIF	POSITIF
36	POSITIF	POSITIF	POSITIF
37	POSITIF	NEGATIF	NETRAL
38	NEGATIF	NEGATIF	NEGATIF
39	POSITIF	POSITIF	POSITIF

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

40	NEGATIF	NEGATIF	NEGATIF
41	POSITIF	POSITIF	POSITIF
42	POSITIF	POSITIF	POSITIF
43	NEGATIF	NEGATIF	NETRAL
44	NEGATIF	NEGATIF	NEGATIF
45	POSITIF	POSITIF	POSITIF
46	POSITIF	POSITIF	POSITIF
47	POSITIF	POSITIF	POSITIF
48	POSITIF	POSITIF	POSITIF
49	POSITIF	POSITIF	POSITIF
50	POSITIF	POSITIF	NEGATIF

Data dari tabel 4 diatas, jika disajikan dalam bentuk grafik adalah seperti pada gambar 1.



Gambar 1. Hasil Klasifikasi Sentimen

Berikut ini adalah penjelasan hasil analisis klasifikasi sentimen berdasarkan tabel 4 dan grafik pada gambar 1 diatas:

1. Pada hasil klasifikasi negative menggunakan Algoritma *Naïve Bayes* dan seleksi fitur Tri-Gram *Character,* didapatkan 27 data uji bersentimen positif dan 23 data uji bersentimen negative.
2. Pada hasil klasifikasi negative menggunakan Algoritma *Naïve Bayes* dan seleksi fitur *Term-Frequency*, didapatkan 29 data uji bersentimen positif dan 21 data uji bersentimen negative.
3. Pada hasil klasifikasi egative menggunakan metode *Lexicon Based*, didapatkan 24 data uji bersentimen positif, 18 data uji bersentimen egative dan 8 data uji bersentimen netral.
4. Berdasarkan hasil klasifikasi menggunakan metode *Lexicon Based* serta Algoritma *Naïve Bayes* dengan seleksi fitur *Term-Frequency* dan Tri-Gram *Character*, jumlah egative positif lebih dominan dibandingkan dengan egative egative maupun netral.

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

5. Jumlah negative positif yang lebih banyak pada klasifikasi menggunakan metode *Lexicon Based* dikarenakan banyaknya *tweet* yang mengandung kata pada kamus *Lexicon* positif negative g kata pada kamus *Lexicon* negative.
6. Salah satu contoh *tweet* yang lebih banyak mengandung kata pada kamus *Lexicon* positif adalah "rezim anjlok, #2019GantiPresiden solusi masa depan Indonesia yang lebih baik" dimana kata rezim ada di kamus *Lexicon* negatif sedangkan kata solusi dan baik ada di kamus *Lexicon* positif sehingga *tweet* tersebut ditandai bersentimen positif

Pada klasifikasi sentimen data uji menggunakan Algoritma *Naïve Bayes* dengan seleksi fitur *Term-Frequency* & Tri-Gram *Character*, sistem lebih condong menghasilkan sentimen positif dibandingkan dengan sentimen negatif maupun netral. Hal ini dipengaruhi oleh ketidakseimbangan data pada klasifikasi data latih menggunakan metode *Lexicon Based* dimana jumlah sentimen positif yang lebih besar dibandingkan sentimen negatif maupun netral

3.2. Hasil Tingkat Akurasi

Untuk mengetahui tingkat akurasi, peneliti menggunakan *multiclass confussion matrix* 3x3 dimana sistem akan mengelompokkan hasil klasifikasi sentimen masing-masing seleksi fitur (kelas prediksi) dan membandingkan dengan hasil klasfikasi *Lexicon Based* (kelas nyata) dan menghitung banyaknya nilai pada kategori **TPos, FPosNeg, FPosNet, FNegPos, TNeg, FNegNet, FNetPos, FNetNeg, TNet**. Berikut pada tabel 5 adalah hasil tingkat akurasi.

Tabel 5. Hasil *Multiclass Confusion Matrix Trigram*

LEXICON	***TRIGRAM***			
		Positif	Negatif	Netral
	Positif	22	2	0
	Negatif	2	16	0
	Netral	3	5	0

Tabel 6. Hasil *Multiclass Confussion Matrix Term Frequency*

LEXICON	***TERM FREQUENCY***			
		Positif	Negatif	Netral
	Positif	23	1	0
	Negatif	4	14	0
	Netral	2	6	0

Dengan menggunakan rumus $Akurasi$, maka didapatkan hasil akurasi sebagai berikut :

a. Pada tabel 5 *Multiclass Confussion Matrix* 3x3 kelas Tri-Gram *Character*, didapatkan nilai *match* sebanyak 22 *True* Positif, 16 *True* Negatif, dan 0 True Netral, lalu dibagi dengan total data uji sebanyak 50 menghasilkan akurasi sebesar 76%.
b. Pada tabel 6 *Multiclass Confussion Matrix* 3x3 kelas *Term Frequency*, didapatkan nilai *match* sebanyak 23 *True* Positif, 14 *True* Negatif, dan 0 *True* Netral, lalu dibagi dengan total data uji sebanyak 50 menghasilkan akurasi sebesar 74%.

Hasil akurasi Algoritma *Naïve Bayes* dengan seleksi fitur Tri-Gram *Character* lebih tinggi dibandingkan seleksi fitur *Term-Frequency* dikarenakan pemotongan *string* kata menjadi karakter memiliki tingkat ketelitian yang lebih baik dalam memeriksa/menganalisa suatu *tweet* terlebih jika terdapat kata-kata yang salah ketik (*typo*).

a. Misal jika terdapat kata JOKOWO pada data uji dan pada model data latih terdapat kata JOKOWI di setiap sentimen, dengan menggunakan seleksi fitur Tri-Gram *Character*,

maka kata JOKOWO akan dipecah menjadi JOK, OKO, KOW, OWO. Dimana, kata JOK, OKO, KOW pada JOKOWO memiliki kesamaan (*similaritas*) dengan kata JOK, OKO, KOW pada JOKOWI sehingga kata JOKOWO masih dianggap sebagai kata JOKOWI & nilai probabilitas tiap karakter tersebut bisa langsung diambil dari model data latih.

b. Lalu, pada seleksi fitur *Term-Frequency*, jika terdapat kata JOKOWO pada data uji dan pada model data latih terdapat kata JOKOWI di setiap sentimen, maka kata JOKOWO tersebut dinyatakan sebagai kata baru sehingga nilai probabilitas kata tersebut dihitung dengan frekuensi 0.

Hasil akurasi Algoritma *Naïve Bayes* dengan seleksi fitur Tri-Gram *Character* yang lebih tinggi dibanding seleksi fitur *Term-Frequency* diperkuat dalam penelitian [15], dimana nilai akurasi Tri-Gram sebesar 86% dan *Term-Frequency* sebesar 68,5%. Untuk kasus klasifikasi sentimen pada *twitter* dalam penelitian ini metode seleksi fitur Tri-Gram *Character* terbukti memiliki ketelitian yang lebih baik dan menghasilkan nilai akurasi yang lebih tinggi dibanding seleksi fitur *Term-Frequency.*

4. KESIMPULAN

Dari klasifikasi sentimen dari 50 *tweet* data uji didapatkan jumlah sentimen positif lebih dominan dibandingkan sentimen negatif maupun netral dikarenakan dengan metode *Lexicon Based* terdapat lebih banyak *tweet* yang mengandung kata dalam kamus *Lexicon* positif dibanding kata dalam kamus *Lexicon* negatif. Sedangkan untuk seleksi fitur, sentimen positif lebih dominan dikarenakan ketidakseimbangan jumlah sentimen positif, negatif dan netral dalam klasifikasi data latih menggunakan metode *Lexicon Based* dimana sentimen positif lebih besar sehingga sistem lebih condong dalam mengklasifikasi ke sentimen positif. Nilai akurasi pada Algoritma *Naïve Bayes* dengan seleksi fitur Tri-Gram *Character* memiliki nilai akurasi yang lebih besar (76%) dibandingkan dengan seleksi fitur *Term-Frequency* (74%). Hal ini disebabkan karena pemenggalan kata menjadi karakter pada seleksi fitur Tri-Gram *Character* memiliki tingkat ketelitian yang lebih baik dalam memeriksa / menganalisa suatu *tweet*.

DAFTAR PUSTAKA

[1] Jagad.id. (2018). Pengertian Media Sosial: Sejarah, Jenis, Ciri Ciri dan Fungsi Tujuan. Retrieved December 18, 2018, https://jagad.id/pengertian-media-sosial-sejarah-jenis-ciri-ciri-dan-fungsi-tujuan/

[2] Kominfo. “Pengguna Internet di Indonesia 63 Juta Orang”. Artikel diakses pada tanggal 10 Agustus 2018. https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker

[3] Pak, A. & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Dalam Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10. Valletta: Malta)

[4] Farisa, F. C. (2018). Peserta Pemilu Diizinkan Kampanye Lewat Sosial Media, tapi Harus Hati-hati. Retrieved December 18, 2018, from https://nasional.kompas.com/read/2018/08/30/19462201/peserta-pemilu-diizinkan-kampanye-lewat-sosial-media-tapi-harus-hati-hati

Vol. 9, No. 1, April 2020, P-ISSN 2089-1245, E-ISSN 2655-4925
DOI: https://doi.org/10.33322/kilat.v9i1.878

[5] Purwanto. "Mengapa Gerakan 2019 Ganti Presiden Makin Populer?". Artikel diakses pada tanggal 10 Agustus 2018 dari http://www.siagaindonesia.com/181491/mengapa-gerakan-2019-ganti-presiden-makin-populer.html

[6] Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. (H. Graeme, Ed.) (1st ed.). Chicago: Morgan & Claypool Publisher. Retrieved from https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[7] Hakimi, F. D. D. (2018). *Sistem Analisis Sentimen Publik Tentang Opini Pemilihan Kepala Daerah Jawa Timur 2018 Pada Dokumen Twitter Menggunakan Naive Bayes Classifier*. Universitas Islam Negeri Sunan Ampel Surabaya.

[8] Rossi, A., Lestari, T., Perdana, R. S., & Fauzi, M. A. (2017). Analisis Sentimen Tentang Opini Pilkada Dki 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Näive Bayes dan Pembobotan Emoji. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(12), 1718–1724.

[9] Hidayatullah, A. F., & Sn, A. (2014). Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter. *Seminar Nasional Informatika 2014*, *2014* (August 2013), 0–8.

[10] Fath, M. K. Al. (2018). *Analisis Sentimen Komentar Kebijakan Full Day School Dari Facebook Page Kemendikbud Ri Menggunakan Algoritma Naïve Bayes Classifier*. Jurusan Teknik Informatika. Fakultas Sains dan Teknologi. Universitas Islam Negeri Syarif Hidayatullah

[11] Indrayuni, E., Wahyudi, M., Informasi, S., Selatan, J., Komputer, I., & Selatan, J. (2015). Penerapan Charachter N-Gram Untuk Sentiment Review Hotel Menggunakan Algoritma Naive Bayes. In *Konfrensi Nasional Ilmu Pengetahuan dan Teknologi (KNIT)* (pp. 88–93). Bekasi.

[12] Feldman, Ronen, dan Sanger, James. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.* New York: Cambridge University Press.

[13] Dewi, D. F., & Supriyanto, C. (2012). Analisis Sentimen Publik Seputar Tren Wisata Pada Twitter Menggunakan Naive Bayes Classifier dengan Penambahan Fitur N-Gram, 1–11.

[14] Ferdinandus, Subari. (2015). Sistem Information Retrieval Layanan Kesehatan Untuk Berobat dengan Metode Vector Space Model berbasis WebGis. Jurnal. Teknik Informatika. STIKI Malang.

[15] Tripathi, G., & N aganna, S. (2015). Feature Selection And Classification Approach For Sentiment Analysis, *2*(2).