Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202294706
Vol. 9, No. 4, Oktober 2022, hlm. 1101-1108 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PEMBENTUKAN DAFTAR *STOPWORD GOFFMAN TRANSITION POINT* DENGAN PEMBOBOTAN *EMOJI* PADA ANALISIS SENTIMEN DI TWITTER

Rizky Maulana Iqbal*[1], Yuita Arum Sari[2], Edy Santoso[3]

[1,2,3]Universitas Brawijaya, Malang
Email: [1]rizkyiqbal1606@gmail.com, [2]yuita@ub.ac.id, [3]edy144@ub.ac.id
*Penulis Korespondensi

(Naskah masuk: 15 Februari 2021, diterima untuk diterbitkan: 24 Oktober 2022)

Abstrak

Analisis sentimen atau *opinion mining* merupakan proses mengekstrak data teks sehingga didapatkan informasi yang terkandung dalam suatu data. Dalam proses ekstraknya, terdapat tahapan *stopword removal* untuk menghapus kata-kata tidak penting dengan menggunakan *stopword*. *Stopword* telah banyak disediakan dalam *digital library* dengan berisikan kata-kata tidak penting, tetapi tidak semua kata-kata tersebut tidak penting dalam suatu data atau kasus tertentu. Penelitian ini berfokus pada perbandingan terhadap *stopword* Tala dengan pembentukan *stopword* dari data latih menggunakan metode *Goffman Transition Point* yang merupakan pengembangan dari metode *Zipf Law* dengan menggunakan metode klasifikasi *K-Nearest Neighbour* (KNN) serta menambahkan pembobotan *emoji* dalam proses pembobotannya. Hasil penelitian ini menunjukkan dengan pembentukan *stopword* menggunakan metode *Zipf Law* menunjukkan nilai akurasi sebesar 73,6% dan menggunakan pembobotan *emoji* dengan nilai K yang dipakai metode KNN K = 17 tetapi jika tidak menggunakan pembobotan *emoji* akurasinya menjadi 72.9%. Formula jarak yang digunakan adalah *Cosine distance*. Jika dengan menggunakan *stopword* Tala dengan parameter yang sama menghasilkan akurasi sebesar 73% dengan pembobotan *emoji* dan 71,9% tanpa pembobotan *emoji*. Berdasarkan hasil tersebut dapat disimpulkan bahwa pembentukan *stopword* dan pembobotan *emoji* dapat meningkatkan akurasi.

Kata kunci: *Analisis Sentimen, Stopword, Goffman Transition Point, Zipf Law, Pembobotan Emoji*

ESTABLISHMENT OF GOFFMAN TRANSITION POINT STOPWORD LIST WITH EMOJI WEIGHTING ON SENTIMENT ANALYSIS IN TWITTER

Abstract

Sentiment analysis or opinion mining is the process of extracting text data, so that the information contained in the data is obtained. In the extracting process, there are stopword removal steps to remove unnecessary words by using a stopword. Many stopwords have been provided in digital libraries containing unimportant words, but not all of these words are not important in a particular data or case. This study focuses on the comparison of the stopword tala with the formation of a stopword from training data using the Goffman Transition Point which is a development of the Zipf Law method using the K-Nearest Neighbor (KNN) classification method and adding emoji weighting in the weighting process. The results of this study indicate that the formation of a stopword using the zipf law method shows an accuracy value of 73.6% and using emoji weighting with the K value used by the KNN method with K = 17 but if you don't use emoji weighting the accuracy will be 72.9%. The distance formula used is the cosine distance. Using a stopword Tala with the same parameters produces an accuracy of 73% with emoji-weighted and 71.9% without emoji-weighted. Based on these results it can be concluded that the formation of stopwords and weighting of emojis can improve accuracy.

Keywords: *Sentiment Analysis, Stopword, Goffman Transition Point, Zipf Law, Emoji Weighting*

1. PENDAHULUAN

Dalam *preprocessing* data terdapat tahapan *stopword removal*. *Stopword removal* merupakan salah satu metode dengan melakukan pemrosesan awal untuk mereduksi *noise* dalam *tweet* (Fathan, 2016). Tujuan dari tahapan *stopword removal* diharapkan mampu mempercepat proses pembobotan setelah dilakukan *preprocessing*. Tahapan tersebut membutuhkan *stopword* sebagai kamusnya. *Stopword* merupakan kata – kata yang tidak deskriptif serta dilakukan penghapusan kata dalam pendekatan *bag-of-words*(Juang, 2016) atau terdapat juga istilah

stoplist yang berisikan kata yang penting. Sudah banyak *digital library* untuk dapat menggunakan *stopword* dengan mudah tetapi *stopword* tersebut juga masih terdapat kata yang penting untuk suatu data. Jika dihapus maka hasil *preprocessing* juga semakin berkurang. Dengan pembentukan *stopword* berdasarkan dataset yang ada dapat mengatasi permasalahan tersebut untuk dapat meningkatkan akurasi klasifikasi teks.

Klasifikasi teks merupakan salah satu permasalahan dalam *text mining* yang sering digunakan untuk menemukan informasi dalam suatu dokumen. *Text mining* merupakan proses ekstraksi berupa informasi serta pengetahuan yang berguna dari sejumlah besar sumber data seperti kutipan teks, dokumen Word, PDF, dan lain-lain(Chandra et al., 2019). Dalam sebuah dokumen teks terdapat banyak jenis kata yang dapat mempengaruhi klasifikasi seperti kata depan, kata sambung, kata ganti, kata sifat, dan lain sebagainya (Rahutomo & Ririd, 2019). Dokumen yang digunakan pada penelitian adalah *tweet* dari Twitter dengan topik tentang *new normal* dengan alasan karena pada dokumen tersebut terdapat opini dari masyarakat mengenai kebijakan pemerintah yang dapat cenderung ke positif, negatif, atau netral. Dalam dokumen tersebut juga terdapat *emoji* sebagai ungkapan ekspresi dari masyarakat tentang kebijakan yang digaungkan pemerintah. *Emoji* merupakan simbol pengekspresian yang dilakukan oleh seseorang dalam suatu *chat room*(Ari Kurnia Rakhman, 2020). Dengan pembentukan *stopword* dari dokumen serta pembobotan *emoji* dapat membantu proses klasifikasi teks.

Metode yang dapat membentuk *stopword* untuk digunakan di *stopword removal* adalah metode *Zipf Law*. Metode *Zipf Law* dilakukan untuk menentukan kata kunci pada suatu dokumen dari suatu pemeringkatan kata(Shaimah & Setyadi, 2019). Seperti pada penelitian sebelumnya yang telah dilakukan oleh Destin Eva Dila menggunakan metode *Zipf Law* sebagai metode pembentukan *stopword* dengan *augmented term frequency – probability term frequency* digunakan untuk pembobotan kata yang dimana penelitian tersebut menyimpulkan bahwa terdapat pengaruh terhadap hasil klasifikasi(Sari et al., 2020).

Selain pembentukan *stopword*, pembobotan *emoji* dilakukan dengan memanfaatkan *emoji* yang ada pada suatu dokumen. Seperti pada penelitian sebelumnya yang dilakukan oleh Lestari dengan berfokus pada pembobotan *emoji*. Pada penelitian ini didapatkan peningkatan akurasi yang dibandingkan dengan klasifikasi tanpa menggunakan pembobotan *emoji*(Lestari et al., 2017).

Berdasarkan penjabaran diatas, penelitian sebelumnya yang dilakukan oleh Destin Eva Dila melakukan pembentukan *stopword* dengan metode *Zipf Law*. Penelitian ini melakukan pembentukan *stopword* dengan menggunakan metode *Goffman Transition Point* yang merupakan pengembangan dari metode *Zipf Law* dan akan dibandingkan juga dengan metode *Zipf Law*. Salah satu algoritme yang dapat melakukan klasifikasi data tekstual dalam pengolahan dokumen secara otomatis adalah KNN dengan menggunakan jarak terdekat atau kemiripan dari data tersebut(Wahyono et al., 2020). Pada penelitian yang dilakukan Wahyono hanya menggunakan 4 macam *distance*, sedangkan pada penelitian ini menggunakan 6 macam *distance*. Alasan menggunakan metode KNN karena memiliki keunggulan dalam melakukan klasifikasi data Twitter yang tidak diketahui kelasnya dengan menggunakan data latih dan data uji dengan memprosedur dengan basis matematis untuk melakukan evaluasi nilai kelas tersebut menjadi sebuah keterangan klasifikasi. Dalam upaya meningkatkan akurasi analisis, ditambahkan pembobotan *emoji*. Selain itu, dilakukan perbandingan hasil akurasi dengan *stopword* Tala dengan hasil dari pembentukan *stopword*.

2. METODE PENELITIAN

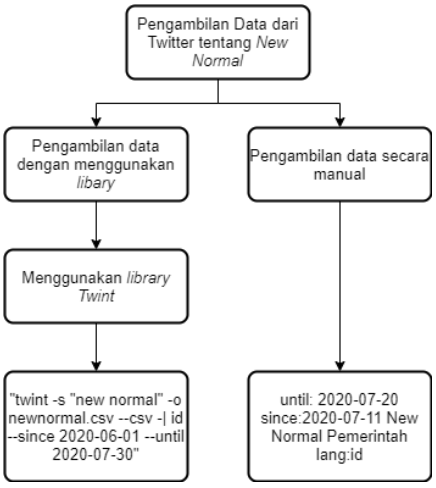
Penelitian ini merupakan lanjutan dari penelitian yang pernah dilakukan oleh (Sari et al., 2020) yang berfokus pada pembentukan daftar *stopword* dengan menggunakan metode *Zipf Law* dan pada penelitian ini menggunakan metode pengembangan terhadap metode *Zipf Law* yaitu *Goffman Transition Point*.

2.1. Pengumpulan Data

Pengumpulan data bertujuan sebagai pendukung serta bahan untuk implementasi sistem pada penelitian ini. Pengumpulan data dilakukan dengan mengambil data dari Twitter berbahasa Indonesia dengan topik "*new normal*" yang diambil dari data bulan Juni hingga bulan Agustus dimana pada saat itu sedang terdapat pro dan kontra terhadap kebijakan pemerintah tersebut dengan berbagai *tweet* yang diberikan seperti komentar negatif, positif, maupun netral sebanyak 300 data yang diambil dari bulan juni hingga bulan agustus, diantaranya 60 data uji dan 240 data latih. Dalam melakukan pelabelan dilakukan dengan cara memberikan kumpulan data *tweet* kepada 3 (tiga) mahasiswa untuk dilakukan pelabelan terhadap data *tweet* tersebut. Skema pengambilan data dapat dilihat pada Gambar 1.

2.2. Diagram Alir Sistem

Pada diagram alir sistem ini menggambarkan jalannya sistem secara umum yang digunakan pada penelitian. Sistem ini melakukan pembentukan *stopword* menggunakan metode *Goffman Transition Point* serta menggunakan *emoji* untuk meningkatkan akurasi dengan melakukan pembobotan terhadap *emoji* tersebut. Proses pertama yang dilakukan adalah menginputkan dataset yang nantinya dibagi menjadi data latih dan data uji.



Gambar 1. Skema Pengambilan Data

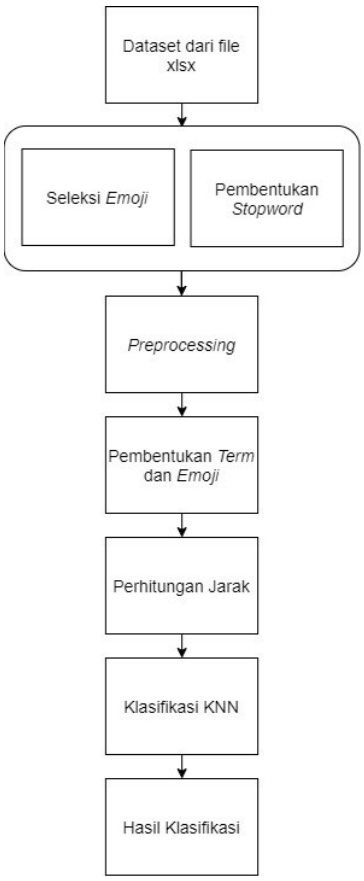
Setelah diinputkan data latih, masuk pada tahapan pembentukan *stopword* menggunakan metode *Zipf Law* dan *Goffman Transition Point* untuk digunakan pada tahapan *preprocessing*. Selanjutnya tahapan preprocessing menggunakan TF-IDF dengan menambahkan *emoji* dalam proses pembobotan tersebut. Tahapan selanjutnya adalah melakukan perhitungan jarak dengan berbagai macam formulanya seperti *cosine distance*, *braycurtis distance*, *minkowski distance*, *manhattan distance*, *supremum distance*, dan *ecludience distance*. Setelah didapatkan nilai dari proses perhitungan jarak, dapat dilakukan klasifikasi dengan menentukan nilai K yang digunakan dalam proses klasifikasi menggunakan KNN. Pada Gambar 2 merupakan diagram alir sistem secara umum.

2.2. *Text Mining*

Text mining adalah salah satu proses penambangan terhadap data yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru yang tidak diketahui sebelumnya dimana dapat menggali informasi yang tersirat secara implisit yang diekstrak dari berbagai sumber informasi data teks yang berbeda-beda. Tidak jauh berbeda dengan data mining, konsep dari *text mining* memiliki kesamaan yaitu menemukan pola, melakukan *preprocessing*, dan tampilan (Rofiqoh et al., 2017).

2.3. *Preprocessing*

Preprocessing merupakan salah satu tahapan dalam melakukan pemrosesan teks yaitu merubah teks menjadi sebuah term index. Tujuan dari dilakuannya *preprocessing* adalah untuk menghasilkan sebuah set *term* index yang dapat mewakili dokumen serta dambil informasi sebagai bagian dalam proses analisa sentimen. Terdapat beberapa tahapan yang biasa digunakan pada proses *preprocessing*, yaitu *parsing*, *lexical analysis*, *case folding*, *stopword removal*, dan *stemming*. Menurut (Mujilahwati, 2016) *preprocessing* text merupakan proses untuk melakukan persiapan dari data mentah sebelum dilakukan proses yang lainnya yang sangat penting pada media sosial karena terdapat kata-kata yang tidak terstruktur dan tidak formal.



Gambar 2. Diagram Alir Sistem

2.4. Pembentukan *Stopword*

Stopword merupakan kata yang sering muncul yang tidak begitu terpakai dalam melakukan *preprocessing* karena tidak membawa informasi(Budiman & Widjaja, 2020). *Stopword* juga disebut dengan *stoplist* tetapi pada *stoplist* berisikan kata yang penting. Pembentukan *stopword* diharapkan mampu memberikan dampak baik terhadap pemrosesan teks dan mengurangi waktu dalam proses *preprocessing* dengan tidak menggunakan *stopword* pada *digital library*. Tetapi *stopword* pada *digital library* terdapat kata – kata yang bukan menggabarkan *stopword* yang ada pada dokumen yang digunakan dalam pemrosesan teks tersebut yang dapat mengurangi kualitas dari *stopword* itu sendiri(Sari et al., 2020). Perlunya pembentukan *stopword* yang dibuat berdasarkan data latih untuk digunakan dalam pemrosesan teks. Metode yang digunakan bermacam-macam, salah satu metodenya adalah *Goffman Transition Point* yang merupakan pengembangan dari metode *Zipf Law*.

2.4.1 *Zipf Law*

Salah satu metode yang dapat digunakan sebagai pembentukan daftar *stopword* tersebut adala metode *Zipf law*. *Zipf Law* ditemukan oleh George Kingsley Zipf yang merupakan salah satu orang yang ahil bahasa di Universitas Harvard. Berdasarkan sumber, Zipf mempelajari frekuensi kata atau jumlah kemunculan kata pada setiap dokumen yang beranggapan bahwa manusia cenderung menggunakan kata – kata yang berulang – ulang yang mengakibatkan kalimat tersebut tidak efektif. Berdasarkan penelitian (Rani & Lobiyal, 2018) bahwa metode *Zipf Law* dapat menghemat waktu pemrosesan dan hasil dari stopwordnya tidak jauh berbeda dengan *stopword* yang menggunakan *digital library*.

Rumus yang digunakan sebagai pemeringkatan kata sebagai berikut pada Persamaan 1.

$$r\ x\ f = c \tag{1}$$

Keterangan:
r = peringkat dari suatu kata
f = frekuensi dari suatu kata
c = nilai konstanta

2.4.2 *Goffman Transition Point*

Hukum Zipf tersebut banyak dilakukan pengembangan salah satunya adalah *Goffman Transition Point* yang merupakan pengambilan nilai tengah yang beranggapan bahwa perubahan dari frekuensi tinggi ke frekuensi rendah memiliki titik teoritis yang didalamnya terdapat daerah transisi. Daerah tersebut yang berisi kata – kata yang dapat dijadikan suatu indeks atau isi dari suatu dokumen (Shaimah & Setyadi, 2019). Rumus yang digunakan untuk mencari nilai titik transisi yang nantinya dapat menentukan daerah transisinya sebagai berikut pada Persamaan 2.

$$n = \frac{-1 + \sqrt{1 + 8li}}{2} \tag{2}$$

Keterangan:
n = Titik transisi
li = Jumlah frekuensi yang bernilai 1 (satu)

2.5. Pembobotan *Term*

Pada tahapan sebelumnya yaitu *preprocessing* menghasilkan sekumpulan *term* dari sebuah dokumen yang dijadikan sebagai indeks. Setiap indeks tersebut perlu diberikan nilai atau bobot untuk melakukan pengubahan dari data ke numerik salah satu metodenya adalah *Term Frequency – Inverse Document Frequency* atau biasa disebut dengan TF-IDF(Herwijayanti et al., 2018). Rumus untuk menghitung TF-IDF sebagai berikut pada Persamaan 3.

$$W_{t,d} = log_{10}\left(\frac{N}{df_t}\right) \times\ 1 + log_{10} tf_{t,d} \tag{3}$$

Keterangan:
N = jumlah keseluruhan dokumen data latih.
df_t = jumlah dokumen yang mengandung *term*.
$tf_{t,d}$ = frekuensi *term* pada suatu dokumen.

2.6. Pembobotan *Emoji*

Emoji merupakan gambar atau simbol grafis Unicode yang digunakan untuk mengekspresikan dari perasaan seseorang (Lestari et al., 2017). Dalam pembobotan *emoji*, metode yang digunakan tidak berbeda dengan pembobotan *term*, yaitu dengan menggunakan pembobotan TF-IDF. *Emoji* yang terdapat pada dokumen tidak dihapus dan disimpan pada suatu variabel agar nantinya dimasukkan ke dalam hasil dari *preprocessing term*.

3. HASIL DAN PEMBAHASAN

3.1 Pengujian Terhadap Distance

Dalam pengujian ini menggunakan 6 (enam) macam distance yaitu *cosine distance*, *euclidean distance*, *manhattan distance*, *braycurtis distance*, *minkowski distance*, dan *supremum distance*. Pengujian ini dilakukan untuk mengetahui *distance* yang memiliki akurasi tertinggi serta digunakan pada penelitian. Nilai k yang digunakan pada pengujian ini adalah k=24 serta persentase pengambilan term untuk *stopword Zipf Law* 10%. Pada Tabel 1 merupakan pengujian terhadap 6 (enam) macam *distance*.

Tabel 1 Pengujian Terhadap *Distance*

Distance	**Pembentukan *Stopword***				
	***S.* Tala**	***Zipf Law* 10%**	***Zipf Law* 10% *Filter* S. Tala**	***GTP***	***GTP Filter* S. Tala**
Cosine	0,703	0,740	0,746	0,703	0,625
Euclidean	0,639	0,656	0,651	0,710	0,620
BrayCurtis	0,676	0,691	0,692	0,713	0,654
Manhattan	0,676	0,599	0,595	0,608	0,578
Supremum	0,647	0,616	0,602	0,664	0,624
Minkowski	0,667	0,638	0,652	0,645	0,617

Keterangan:
GTP : *Goffman Transition Point*

3.2. Pengujian Nilai K Untuk Klasifikasi KNN

Pada pengujian nilai k ini menggunakan *stopword* tala pada tahapan *stopword removal* dimana pada pengujian sebelumnya pembentukan *stopword* tersebut mendapatkan akurasi tertinggi pada *cosine distance* nya. Nilai k yang diuji, yaitu k dari 7 hingga 50 dengan setiap nilai k melakukan perhitungan

akurasi dengan *confusion matrix* dengan *k-fold* sebesar 10-*fold* dilakukan untuk mendapatkan nilai k yang optimal untuk pengujian selanjutnya. Pada Tabel 2 merupakan pengujian nilai k untuk digunakan dalam proses klasifikasi KNN.

Tabel 2. Pengujian Nilai K Pada Klasifikasi KNN

Nilai K	Rata – rata Akurasi
7	0,704
8	0,714
17	0,730
18	0,707
27	0,695
28	0,683
29	0,691
35	0,682
36	0,665
37	0,663
43	0,670
44	0,686
45	0,676
49	0,671
50	0,679

3.3. Pengujian Persentase Pengambilan *Term* Pada *Stopword Zipf law*

Pada pengujian persentase pengambilan term pada *stopword Zipf Law* dilakukan 2 (dua) macam keadaan, yaitu menggunakan *filter stopword* Tala dan tanpa *filter stopword* Tala. Pengujian ini dilakukan untuk menentukan banyaknya term yang dapat diambil dan digunakan untuk tahapan *stopword removal* yang isi dari *stopword* tersebut adalah *term* yang tidak penting dan diambil dari hasil terendah berdasarkan perhitungan frekuensi dengan peringkatnya. Pengujiannya menggunakan persentase sebesar 5% hingga 100% dengan nilai k=17 untuk klasifikasi KNN dan menggunakan perhitungan *cosine distance* untuk perhitungan jaraknya. Pada Pada Tabel 3 merupakan pengujian terhadap persentase pengambilan *term* untuk pembentukan *stopword Zipf Law*.

3.4. Pengujian Terhadap Berbagai Jenis *Stopword*

Pada Pengujian ini dilakukan untuk mengetahui perbandingan akurasi pada setiap perlakukan pada jenis *stopword* yang digunakan, yaitu *stopword* Tala, *stopword Zipf Law*, dan *Goffman Transition Point* dengan nilai k yang digunakan sebesar 17 yang didapatkan dari pengujian sebelumnya. Formula jarak yang digunakan pada pengujian ini adalah *cosine distance* dan *braycurtis distance*. Pada Tabel 4 merupakan pengujian terhadap berbagai jenis stopword yang nantinya akan dibandingkan berdasarkan hasil akurasinya.

Tabel 3. Pengujian Persentase Pengambilan *Term* Pada *Stopword Zipf Law*

Persentase	*Stopword*	
	***Zipf Law* Tanpa *Filter* S. Tala**	***Zipf Law* Dengan *Filter* S. Tala**
5	0,731	0,732
15	0,727	0,726
25	0,731	0,725
35	0,736	0,724
45	0,729	0,723
55	0,722	0,716
65	0,712	0,726
75	0,705	0,726
85	0,682	0,703
95	0,651	0,715

Tabel 4. Pengujian Terhadap Berbagai Macam Stopword

Distance	*Stopword*	Hasil Evaluasi			
		Accuracy	*Precision*	*Recall*	*F-Measure*
Cosine Distance	*Stopword* Tala	0,730	0,614	0,596	0,604
	Zipf Law 35%	0,736	0,612	0,601	0,606
	Zipf Law 5% *Filter* S. Tala	0,732	0,624	0,590	0,606
	Goffman Transition Point	0,719	0,577	0,573	0,575
	Goffman Transition Point Filter S. Tala	0,661	0,496	0,487	0,488
BrayCurtis Distance	*Stopword* Tala	0,671	0,563	0,504	0,531
	Zipf Law 35%	0,717	0,582	0,570	0,576
	Zipf law 5% *Filter* S. Tala	0,704	0,572	0,558	0,565
	Goffman Transition Point	0,706	0,559	0,557	0,575
	Goffman Transition Point Filter S. Tala	0,627	0,429	0,451	0,438

Tabel 5. Pengujian Pembobotan *Emoji*

Cek *Emoji*	*Stopword*	Hasil Evaluasi			
		Accuracy	*Precision*	*Recall*	*F-Measure*
Dengan *Emoji*	*Stopword* Tala	0,730	0,614	0,596	0,604
	Zipf Law 35%	0,736	0,612	0,601	0,606
	Zipf Law 5% *Filter* S. Tala	0,732	0,624	0,590	0,606
	Goffman Transition Point	0,719	0,577	0,573	0,575
	Goffman Transition Point Filter S. Tala	0,661	0,496	0,487	0,488
Tanpa *Emoji*	*Stopword* Tala	0,719	0,592	0,580	0,585
	Zipf Law 35%	0,729	0,598	0,591	0,594
	Zipf Law 5% *Filter* S. Tala	0,727	0,613	0,582	0,596
	Goffman Transition Point	0,709	0,564	0,562	0,562
	Goffman Transition Point Filter S. Tala	0,643	0,472	0,463	0,462

3.5. Pengujian Terhadap Pembobotan *Emoji*

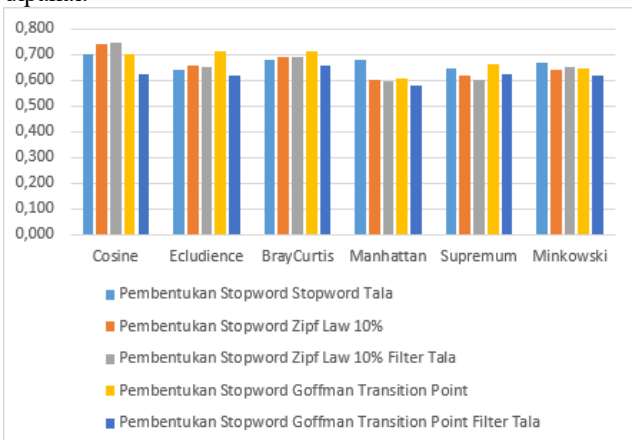
Pada pengujian ini dilakukan untuk mengetahui seberapa berpengaruh *emoji* terhadap analisis sentimen dengan menggunakan berbagai macam pembentukan *stopword* serta formula jarak yang digunakan adalah *cosine distance* dengan nilai k sebesar 17. Pada Tabel 5 merupakan pengujian untuk pembobotan *emoji*.

3.6. Analisis Terhadap *Distance*

Berdasarkan hasil pengujian didapatkan akurasi tertinggi rata-rata saat menggunakan cosine distance dengan akurasi terbaiknya sebesar 0,746 dengan stopword removal menggunakan *stopword Zipf Law filter stopword* Tala dan pengambilan *term* sebanyak 10%. Kesimpulannya *cosine distance* lebih baik dengan *distance* yang lainnya dimana pembobotan term berpengaruh pada *cosine distance*, yaitu semakin besar pembobotan pada suatu term maka semakin tinggi juga nilai *cosine distance* yang dihasilkan dengan melakukan perkalian nilai pembobotan dari data latih dan nilai IDF dari data uji. Pada Gambar 3 merupakan grafik perbandingan akurasi pada setiap *distance* dengan menggunakan bermacam-macam *stopword*. Seperti contoh pada kalimat "new-normal tapi tarif transportasi umum tidak-normal" dengan data negatif hanya dapat diklasifikasikan dengan tepat jika menggunakan formula jarak *cosine distance*.

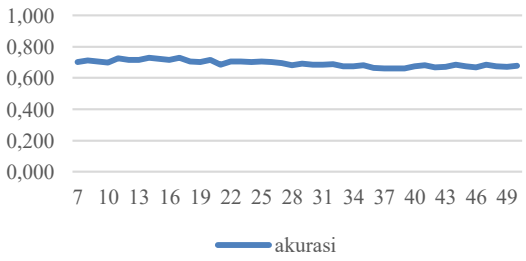
3.7. Analisis Terhadap Pengujian Nilai K Untuk Klasifikasi KNN

Berdasarkan pengujian didapatkan akurasi tertinggi pada nilai K=17 dengan akurasinya sebesar 0,730 serta nilai dengan akurasi terendah saat nilai K=37 dengan akurasi yang didapatkan sebesar 0,663. Nilai K pada KNN tidak dapat diambil nilai yang terkecil atau terbesar karena hal tersebut mempengaruhi hasil dari klasifikasinya dalam pengambilan nilai jarak sejumlah nilai K dan juga tergantung pada data yang dipakai.



Gambar 3. Grafik Perbandingan Setiap *Distance*

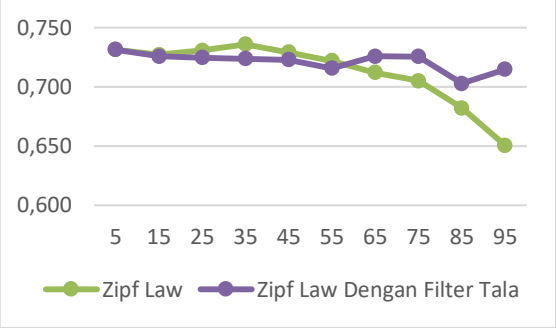
Semakin besar nilai K yang digunakan pada proses klasifikasi cenderung akurasi yang dihasilkan semakin kecil karena semakin besar nilai K, maka klasifikasinya menjadi samar dan kurang jelas dan hasil klasifikasi tersebut terlalu banyak data yang tidak dapat diklasifikasikan. Pada Gambar 4 merupakan grafik akurasi dari setiap kemungkinan nilai K.



Gambar 1 Grafik Akurasi Setiap Kemungkinan Nilai K

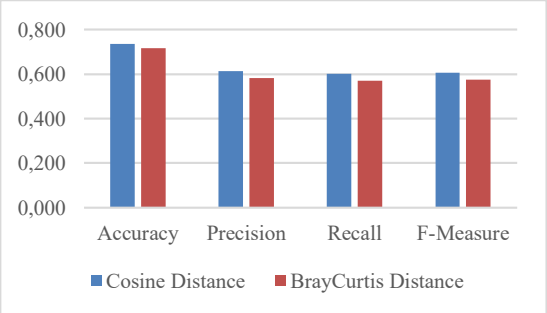
3.8. Analisis Terhadap Persentase Pengambilan *Term* Pada *Stopword Zipf Law*

Berdasarkan pengujian didapatkan hasil pada perlakukan *stopword Zipf Law* tanpa *filter stopword* Tala mendapatkan akurasi tertinggi saat persentasenya sebesar 35% dengan akurasi 0,736. Pada perlakukan *stopword Zipf Law* dengan *filter stopword* Tala mendapatkan akurasi tertinggi saat persentase pengambilan termnya sebesar 5% dengan akurasi 0,732. Penggunaan *filter stopword* Tala tidak terlalu berpengaruh terhadap pembentukan *stopword Zipf Law* dan hasil keseluruhan percobaan persentase, penggunaan *filter stopword* Tala lebih stabil akurasinya yaitu diatas 0,7 daripada tanpa melakukan *filter* karena setelah di *filter* isi dari *stopword* tersebut akan semakin sedikit jumlahnya dan berisikan kata tidak penting semuanya. Jika pada *Zipf Law* masih terdapat kata – kata yang penting tetapi dianggap tidak penting seperti term "pelongaran", "tidaknormal", "regulasi", dan sebagainya karena hal tersebut juga jika persentase semakin naik, maka akurasinya juga jelas semakin turun. Daftar *stopword* yang dihasilkan terlalu banyak, maka akurasinya akan menurun. Pada Gambar 5 merupakan grafik dari akurasi percobaan pada persentase pengambilan *term*.



Gambar 2 Grafik Akurasi Berdasarkan Persentase Pengambilan *Term*

3.9. Analisis Terhadap Berbagai Jenis *Stopword*

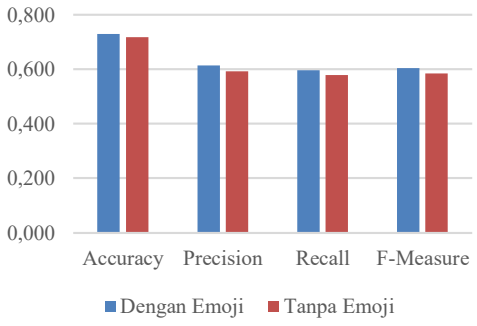


Gambar 3 Perbandingan Menggunakan *Stopword Zipf Law* 35%

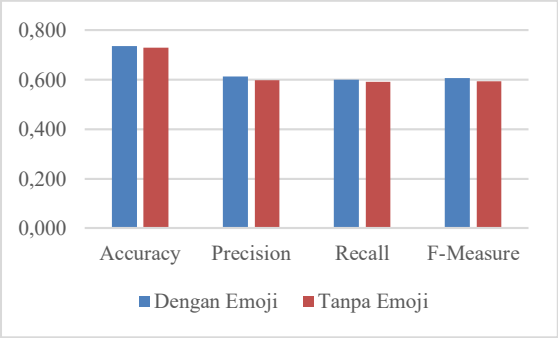
Berdasarkan pengujian didapatkan hasil akurasi tertinggi saat tahapan *stopword removal* menggunakan *stopword Zipf Law* serta penghitungan jarak menggunakan *cosine distance* dengan akurasi sebesar 0,736 dengan pengambilan term sebanyak 35% dari keseluruhan term dengan nilai *precision* sebesar 0,612, *recall* sebesar 0,601, dan *f-measure* sebesar 0,606. Pada *braycurtis distance* juga didapatkan hasil akurasi terbaik dengan menggunakan pembentukan *stopword Zipf Law* dengan akurasi sebesar 0,717. Sehingga dapat disimpulkan bahwa pembentukan *Zipf Law* dapat meningkatkan akurasi daripada hanya menggunakan *stopword* Tala yang masih sering banyak dipakai. Pada suatu *fold* yang memiliki akurasi tertinggi, *term* seperti "wkwkwk", "sableng", "domestic", "adik", dan "wayang" dianggap tidak penting pada penggunaan *stopword Zipf Law* dengan 35%. Jika difilterisasi dengan *stopword* Tala maka *term* tersebut menjadi dianggap kata penting. Pada pengujian ini juga mengalami kesulitan dalam melakukan klasifikasi terhadap data netral dimana pada saat *fold* tertentu akurasinya menurun dikarenakan data netral yang banyak tidak diprediksi dengan tepat.

3.10. Analisis Terhadap Pembobotan *Emoji*

Berdasarkan pengujian dapat dilihat bahwa akurasi saat menggunakan *emoji* dengan seluruh pengujian *emoji* mendapatkan hasil akurasi yang baik dibandingkan dengan tanpa menggunakan *emoji*. Emoji seperti "☺" yang sering digunakan pada sentiment positif dan "☹" yang seing digunakan pada sentiment negative. Contoh dari kedua emoji tersebut juga dapat meningkatkan hasil klasifikasi text. Akurasi tertinggi tanpa *emoji* saat menggunakan *stopword Zipf Law* dengan akurasi sebesar 0,729 dan nilai *precision* 0,598, *recall* 0,591, dan *f-measure* 0,594. Perbedaan akurasi yang tidak terlalu jauh dikarenakan *emoji* pada data pengujian minim jumlahnya. Sehingga dapat disimpulkan dengan *emoji* yang minim dapat meningkatkan akurasi klasifikasi. Pada Gambar 7 dan 8 pengujian dengan formula jarak *cosine distance*, dapat dilihat bahwa pembobotan *emoji* akurasinya selalu lebih tinggi jika dibandingkan dengan tanpa *emoji*.



Gambar 5 Grafik Akurasi Menggunakan *Stopword* Tala



Gambar 6 Grafik Akurasi Menggunakan *Stopword Zipf Law Filter Stopword* Tala *5%*

4. KESIMPULAN DAN SARAN

Kesimpulan yang dapat diambil dari penelitian diatas, yaitu pembentukan *stopword* berdasarkan dokumen latih dan pembobotan emoji dapat meningkatkan akurasi klasifikasi dengan akurasi sebesar 73,6% serta nilai *precision* 61,2%, *recall* 60,1%, dan *f-measure* 60,6% serta persentase pengambilan term untuk pembentukan stopword juga berpengaruh karena jika semakin sedikit akurasinya semakin kecil dan juga semakin banyak, maka klasifikasinya akan semakin susah diprediksi.

Saran yang dapat diberikan untuk penelitian selanjutnya, yaitu menggunakan *stopword* Bahasa Indonesia yang lain selain *stopword* Tala untuk dapat dilakukan *filter* dengan pembentukan *stopword zipf law*, menggunakan metode klasifikasi yang lain dengan jumlah dataset yang lebih banyak dengan jumlah emoji yang lebih banyak juga, dan menggunakan metode untuk melakukan penanganan terhadap kata yang tidak baku.

DAFTAR PUSTAKA

ARI KURNIA RAKHMAN. 2020. Emoji Pada Media Sosial Sebagai Komunikasi Antarbudaya. *Mozaik Komunikasi*, *2*, 1–11. http://jom.untidar.ac.id/index.php/mozaik/article/view/1194

BUDIMAN, A. E., & WIDJAJA, A. 2020. Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(3), 475–488. https://doi.org/10.28932/jutisi.v6i3.2892

CHANDRA, D. N., INDRAWAN, G., & SUKAJAYA, I. N. 2019. Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naive Bayes Dengan Fitur N-Gram. *10*(2), 11–19.

FATHAN, A. H. 2016. *Pengaruh Stopword Terhadap Performa Klasifikasi Tweet Berbahasa Indonesia* (pp. 1–4). JISKA.

HERWIJAYANTI, B., RATNAWATI, D. E., & MUFLIKHAH, L. 2018. Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(1), 306–312.

JUANG, D. 2016. Analisis Spam dengan Menggunakan Naïve Bayes. *Jurnal Teknovasi*, *3*(2), 51–57.

LESTARI, A. R. 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. In *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* (Vol. 1, Issue 12, pp. 1718–1724). http://j-ptiik.ub.ac.id

MUJILAHWATI, S. 2016. Pre-Processing Text Mining Pada Data Twitter. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, *2016*(Sentika), 2089–9815.

RAHUTOMO, F., & RIRID, A. R. T. H. 2019. Evaluasi Daftar Stopword Bahasa Indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *6*(1), 41. https://doi.org/10.25126/jtiik.2019611226

RANI, R., & LOBIYAL, D. K. 2018. Automatic Construction of Generic Stop Words List for Hindi Text. *Procedia Computer Science*, *132*(Iccids), 362–370. https://doi.org/10.1016/j.procs.2018.05.196

ROFIQOH, U., PERDANA, R. S., & FAUZI, M. A. 2017. Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature. In *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya* (Vol. 1, Issue 12, pp. 1725–1732). http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/628

SARI, D. E. D., SARI, Y. A., & FURQON, M. T. 2020. *Pembentukan Daftar Stopword menggunakan Zipf Law dan Pembobotan Augmented TF - Probability IDF pada Klasifikasi Dokumen Ulasan Produk* (Vol. 4, Issue 1, pp. 406–412).

SHAIMAH, L., & SETYADI, A. 2019. Relevansi Kata Kunci Hasil Pemeringkatan Zipf Pada Artikel Jurnal Berkala Ilmu Perpustakaan Dan Informasi Volume 13, No. 2, Tahun 2017. *Jurnal Ilmu Perpustakaan*, *8*.

WAHYONO, W., TRISNA, I. N. P., SARIWENING, S. L., FAJAR, M., & WIJAYANTO, D. 2020. Comparison of distance measurement on k-nearest neighbour in textual data classification. *Jurnal Teknologi Dan Sistem Komputer*, *8*(1), 54–58. https://doi.org/10.14710/jtsiskom.8.1.2020.54-58