

Pembentukan Daftar *Stopword* menggunakan *Term Based Random Sampling* pada Analisis Sentimen dengan metode Naïve Bayes

(Studi Kasus: Kuliah Daring di Masa Pandemi)

Raditya Rinandyaswara¹, Yuita Arum Sari², Muhammad Tanzil Furqon³

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya

Email: ¹radityarin@gmail.com, ²yuita@ub.ac.id, ³m.tanzil.furqon@ub.ac.id

Abstrak

Stopword Removal merupakan bagian dari tahapan *preprocessing* teks yang bertujuan untuk menghapus kata yang tidak relevan didalam suatu kalimat berdasarkan daftar *stopword*. Daftar *stopword* yang biasa digunakan berbentuk *digital library* yang daftarnya sudah tersedia sebelumnya, namun tidak semua kata-kata yang terdapat didalam *digital library* merupakan kata yang tidak relevan dalam suatu data tertentu. Penelitian ini menggunakan daftar *stopword* yang dibentuk dengan algoritme *Term Based Random Sampling*. Dalam *Term Based Random Sampling* terdapat 3 parameter yaitu Y untuk jumlah perulangan pengambilan kata random, X untuk jumlah pengambilan bobot terendah dalam perulangan Y, dan L sebagai persentase jumlah *stopword* yang ingin digunakan. Sehingga penelitian ini ditujukan untuk mencari tahu kombinasi terbaik dari 3 parameter tersebut serta perbandingan *stopword Term Based Random Sampling* dengan *stopword* Tala dan tanpa proses *stopword removal* dalam analisis sentimen *tweet* mengenai kuliah daring dengan menggunakan metode *Naïve Bayes*. Hasil evaluasi dengan *stopword Term Based Random Sampling* mendapatkan akurasi tertinggi dengan kombinasi X, Y, L sebesar 10, 10, 40 dengan *macroaverage accuracy* sebesar 0,758, *macroaverage precision* sebesar 0,658, *macroaverage recall* sebesar 0,636, dan *macroaverage f-measure* sebesar 0,647. Berdasarkan hasil pengujian disimpulkan bahwa semakin besar nilai X, Y, L maka semakin tinggi kemungkinannya untuk *accuracy*, *precision*, *recall*, dan *f-measure* turun. Hasil pengujian membuktikan bahwa *Term Based Random Sampling* berhasil mendapatkan akurasi lebih tinggi dibandingkan dengan *stopword* Tala maupun tanpa menggunakan proses *stopword removal*.

Kata kunci: *sentimen analisis, kuliah daring, Naïve Bayes, Term Based Random Sampling, stopwords*

Abstract

Stopword Removal is part of the text preprocessing stage which aims to remove irrelevant words in a sentence based on the *stopword* list. The *stopword* list that is commonly used is in the form of a *digital library* whose list is already available, but not all words contained in the *digital library* are irrelevant words in certain data. This study uses a *stopword* list formed by the *Term Based Random Sampling* algorithm. In *Term Based Random Sampling*, there are 3 parameters, namely Y for the number of random word retrieval repetitions, X for the lowest number of weights in Y repetitions, and L as the percentage of the number of *stopwords* you want to use. So this research is aimed at finding out the best combination of these 3 parameters as well as the comparison of term based random sampling *stopword* with *stopword* tuning and without *stopword removal* process in the analysis of *tweet* sentiment about online lectures using the *Naïve Bayes* method. The results of the evaluation using the *Term Based Random Sampling* *stopword* get the highest accuracy with a combination of X, Y, L of 10, 10, 40 with a *macroaverage accuracy* of 0.758, a *macroaverage precision* of 0.658, a *macroaverage recall* of 0.636, and a *macroaverage f-measure* of 0.647. Based on the test results, it is concluded that the greater the X, Y, L values, the higher the probability that the accuracy, precision, recall, and f-measure will decrease. The test results prove that *Term Based Random Sampling* is successful in obtaining higher accuracy than *stopword* tuning or without using the *stopword removal* process.

Keywords: *sentiment analysis, online school, Naïve Bayes, Term Based Random Sampling, stopwords*

1. PENDAHULUAN

Pada akhir tahun 2019 lalu, dunia dikejutkan dengan adanya wabah yang diakibatkan oleh virus *corona* yang berasal dari kota Wuhan, China. Penyakit yang disebut sebagai (COVID-19) ini adalah penyakit yang menyerang sistem pernapasan virus manusia. Menurut data pemerintah China, penduduk Hubei menjadi kasus pertama Covid-19 pada 17 November 2019 (Arnani, 2020). Setelah kasus pertama Covid-19 di dunia itu terjadi peningkatan pasien tiap bulannya. Hingga saat ini Indonesia sudah melewati angka 190 ribu kasus Covid-19 yang sudah terkonfirmasi yang terhitung dari sejak pasien pertama (Ramadhan, et al., 2020). Dengan adanya pandemi Covid-19 ini, pemerintah Indonesia mengadakan sistem New Normal dengan tujuan untuk mempercepat penanganan Covid-19 (Putsanra, 2020). Dalam penerapannya banyak perilaku yang biasa dilakukan secara luring berubah menjadi daring. Salah satu contohnya adalah perkuliahan. Tentunya kuliah daring ini menyebabkan banyak pro dan kontra dan salah satu sarana masyarakat menuliskan opininya mengenai kuliah daring tersebut yaitu melalui Twitter. Namun twitter sendiri hanya berfungsi sebagai sosial media sehingga tidak menyediakan analisis sentimen dari tweet pengguna dan bagaimana masyarakat Indonesia menanggapi protokol baru tersebut.

Pada penelitian ini, peneliti akan membangun sebuah sistem analisis sentimen terhadap kuliah daring yang dituliskan masyarakat di Twitter menggunakan metode klasifikasi *Naïve Bayes*. Dalam penelitian ini klasifikasi akan dibuat menjadi 3 kelas yaitu opini netral, positif dan negatif sesuai dari saran penelitian analisis sentimen sebelumnya (Sa'rony, et al., 2019). Dalam proses analisis sentimen diperlukan preprocessing data agar data dapat dihitung dan diolah. Salah satu tahap yang penting dalam *preprocessing* adalah *Stopword Removal*, yaitu penghapusan kata yang tidak memiliki nilai keunikan dalam suatu dokumen. Biasanya dalam proses penghapusan *stopword* ini memanfaatkan *digital library* namun ada *stopword* yang kurang tepat sehingga hal itu dapat membuat kualitas proses *stopword removal* menjadi berkurang (Dila Purnama Sari, et al., 2020). Oleh karena itu perlu adanya suatu mekanisme sistem daftar *stopword* dinamik yang dapat menghasilkan

daftar *stopword* yang sesuai dengan yang diperlukan sesuai saran dari penelitian sebelumnya (Rahutomo & Ririd, 2018). Oleh karena itu pada penelitian ini, peneliti akan melakukan pembuatan daftar *stopword* menggunakan metode *Term Based Random Sampling*.

2. LANDASAN KEPUSTAKAAN

2.1 Text Preprocessing

Text Preprocessing merupakan langkah awal yang dilakukan dalam analisis sentimen untuk menyiapkan data yang berupa teks agar mudah untuk diproses nantinya (Gaddam, 2019). Teks *pre-processing* ini meliputi, *case folding*, *cleaning*, *tokenizing*, *stopword removal* dan *stemming*.

2.2 Term Based Random Sampling

Term Based Random Sampling adalah suatu metode yang dapat digunakan untuk menghasilkan daftar *stopword* secara otomatis berdasarkan seberapa informatif kata tertentu (Lo, et al., 2005). Kita dapat mengetahui apakah kata tersebut *stopword* atau bukan dengan melihat kepentingannya, semakin tidak penting kata tersebut, maka lebih tinggi pula kata tersebut memiliki kemungkinan menjadi *stopword*. Untuk mencari nilai kepentingan dari suatu *term* dapat dilakukan dengan perhitungan dengan rumus dari teori *Kullback-Leibler*. Dengan rumus tersebut kita dapat memberi bobot dari suatu *term* pada dokumen sampel. Berikut rumus dari *Kullback-Leibler* direpresentasikan dalam Persamaan 1.

$$w(t) = P_x \cdot \log_2 \left(\frac{P_x}{P_c} \right) \quad (1)$$

Yang dimana P_x direpresentasikan dalam Persamaan 2 dan P_c direpresentasikan dalam Persamaan 3 adalah sebagai berikut.

$$P_x = \frac{tf_x}{l_x} \quad (2)$$

$$P_c = \frac{F}{token_c} \quad (3)$$

Keterangan:

- $w(t)$: bobot *term* t pada dokumen sampel
- tf_x : frekuensi kueri *term* dalam dokumen sampel
- l_x : jumlah dari panjang dokumen sampel
- F : frekuensi kueri *term* dari keseluruhan dokumen

$token_c$: total token dari keseluruhan dokumen

Dalam perhitungannya dilakukan pemilihan acak term dari keseluruhan *term*, lalu kemudian ambil dokumen yang mengandung term tersebut dan cari semua *term* dalam dokumen tersebut. Setiap term dalam dokumen tersebut akan dilakukan perhitungan bobot nya menggunakan *Kullback-Leibler*. Lalu setelah perhitungan bobotnya diambil sejumlah X *term* yang diurutkan dari bobot terendah yang dimana X adalah parameter yang dapat diubah-ubah nantinya. Dalam prosesnya pemilihan *term* acak ini dilakukan sebanyak Y kali dimana Y adalah sebuah parameter yang dapat diisi secara manual untuk mendapatkan hasil yang terbaik. Setelah melakukan proses yang dijelaskan sebelumnya dihitung rata-rata keseluruhan bobot yang didapat oleh *term* lalu diambil sejumlah L dimana L adalah parameter yang dapat diubah nantinya. L adalah parameter yang menentukan berapa jumlah daftar *stopword* yang ingin digunakan.

2.3 Term Frequency – Inverse Document Frequency

Pada tahap ini dilakukan pembobotan kata yang mempresentasikan kata-kata tersebut untuk dilakukan perhitungan nantinya. Salah satu metode dalam *term weighting* yang sering digunakan adalah *Term Frequency – Inverse Document Frequency* (*tf. idf*) (Jones, 2004). Metode TF-IDF adalah penggabungan dua metode untuk melakukan pembobotan kata. TF atau *Term Frequency* adalah frekuensi kemunculan *term* pada suatu dokumen dan IDF atau *Inverse Document Frequency* adalah perhitungan *inverse* terhadap frekuensi dokumen yang mengandung kata tersebut (Prabowo, et al., 2016). Berikut rumus yang digunakan untuk perhitungan *tf-idf* direpresentasikan pada Persamaan 4 untuk *log tf* dan Persamaan 5 untuk *idf*.

$$tft, d = 1 + \log (ft, d) \quad (4)$$

$$idf(t) = \frac{\log (N)}{df_t} \quad (5)$$

Keterangan:

tft, d : frekuensi *term* pada dokumen d
 $idf(t)$: nilai *Inverse Document Frequency* suatu *term t*
 N : total dokumen
 $df(t)$: nilai *Document Frequency* suatu

term t

2.4 Naïve Bayes Classifier

Algoritme *Naïve Bayes* yaitu algoritme klasifikasi *supervised* yang berbasis dengan teorema bayes dengan asumsi independensi tiap fitur (Sawla, 2018). Algoritme ini menggunakan metode probabilistik dan statistik. Algoritme ini mencari probabilitas tertinggi untuk proses klasifikasi. Perhitungan Algoritme *Naïve Bayes* direpresentasikan pada Persamaan 6 dan perhitungan *prior* direpresentasikan pada Persamaan 7.

$$P(c|d) = P(c) * P(d|c) \quad (6)$$

$$P(c) = \frac{N_c}{N} \quad (7)$$

Keterangan:

$P(c|d)$: *Posterior* atau Probabilitas kelas c diberikan dokumen d
 $P(c)$: *Prior* atau Probabilitas awal muncul kategori c
 $P(d|c)$: *Likelihood*
 N_c : Jumlah dokumen kelas c
 N : Jumlah seluruh dokumen

Dalam perhitungan *likelihood* atau *conditional probability* salah satu metodenya adalah menggunakan *Multinomial*. Perhitungan *conditional probability* dengan *multinomial* direpresentasikan pada Persamaan 8.

$$P(w|c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \quad (8)$$

Keterangan:

$P(w|c)$: *Likelihood* w dalam kelas c
 $\text{count}(w, c)$: Jumlah kemunculan kata w pada kategori c
 $\text{count}(c)$: Jumlah semua total kemunculan kata pada kategori c
 $|V|$: Jumlah *term* unik atau fitur

Namun dalam penelitian kali ini digunakan *TF-IDF* sebagai pembobotan sehingga perhitungan *likelihood* atau *conditional probability* direpresentasikan pada Persamaan 9 (Rahman, et al., 2017).

$$P(t_n|c) = \frac{W_{ct} + 1}{(\sum_{W_t \in V} W'_{ct}) + B'} \quad (9)$$

Keterangan:

$P(w|c)$: *Likelihood* w dalam kelas c
 W_{ct} : Nilai pembobotan (W) *TF-IDF* dari *term t* di kategori c
 $\sum_{W_t \in V} W'_{ct}$: Jumlah bobot *TF-IDF* seluruh

B' : $term$ pada kelas C
: Jumlah IDF $term$ pada seluruh dokumen.

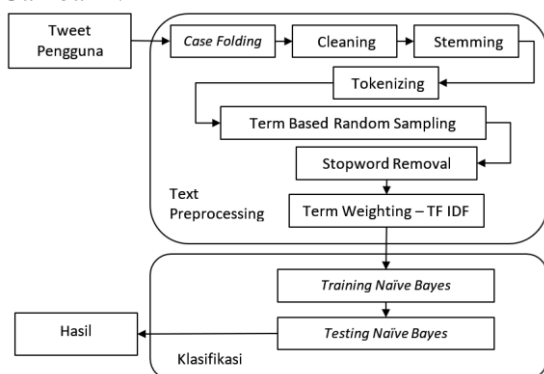
3. METODE PENELITIAN

3.1 Pengumpulan Data

Datasets yang digunakan dalam penelitian ini adalah berupa *Tweet* yang berasal dari Twitter. Data diambil menggunakan *library Twint* sebagai *data scrapper* Twitter untuk *Python* dengan kata kunci “Kuliah Daring” dan “Kuliah Online” yang dikumpulkan pada bulan April 2020 hingga November 2020. Data yang dikumpulkan melalui proses normalisasi secara manual terlebih dahulu. Data yang digunakan berjumlah 300 data yang dilabeli secara manual oleh 5 pakar. Dari 300 data tersebut terdiri dari 100 data negatif, 100 data netral, dan 100 data positif.

3.2 Perancangan Algoritme

Perancangan Algoritme dibagi menjadi beberapa tahapan yang akan ditampilkan pada Gambar 1.



Gambar 1. Perancangan Algoritme

Perancangan Algoritme diawali dengan melakukan pembuatan daftar *stopword* yang prosesnya diawali dengan *preprocessing* yang meliputi *case folding*, *cleaning*, *tokenizing*, dan *stemming*. *Preprocessing* ini bertujuan untuk merubah data latih berbentuk kumpulan dokumen menjadi *term* untuk dilakukan perhitungan algoritme *Term Based Random Sampling*. Setelah daftar *stopword* hasil dari algoritme *Term Based Random Sampling* didapatkan, selanjutnya data tersebut akan melalui tahap *stopword removal* atau penghapusan kata *stopword* dengan menggunakan daftar *stopword* yang telah dibuat sebelumnya. Setelah didapatkan daftar *term*, langkah selanjutnya adalah proses pembobotan

kata dengan menggunakan *tf. idf* untuk merubah kata tersebut menjadi suatu nilai yang nantinya dapat diproses oleh sistem untuk dilatih dan diklasifikasi menggunakan metode *Multinomial Naïve Bayes*.

4. PENGUJIAN DAN ANALISIS

4.1 Pengujian kombinasi parameter X, Y dan L terbaik

Pengujian kombinasi parameter X, Y, dan L dilakukan untuk mengetahui kombinasi terbaik parameter yang digunakan terhadap hasil evaluasi. Pada pengujian ini dilakukan 10 fold cross validation dengan pembagian data 80% data latih dan 20% data uji. Hasil evaluasi pengujian pengaruh parameter X, Y, dan L yang sudah diurutkan berdasarkan akurasi akan ditampilkan pada Tabel 1.

Tabel 1. 10 Hasil Pengujian pengaruh X, Y, dan L terbaik

X	Y	L	Accuracy	Precision	Recall	F-Measure
10	10	40	75,8%	65,8%	63,6%	64,7%
40	30	10	75,6%	65,3%	63,3%	64,3%
20	20	20	75,4%	64,7%	63,3%	64,3%
50	50	20	75,4%	64,6%	62,9%	63,7%
10	50	50	75,3%	64,7%	63%	63,8%
40	10	10	75,2%	65,1%	62,8%	63,9%
30	40	10	75,2%	64,9%	62,8%	63,8%
10	30	10	75,1%	64,7%	62,6%	63,6%
30	10	20	75,1%	64,4%	62,6%	63,4%
10	10	10	75%	64,4%	62,5%	63,4%

Berdasarkan pada Tabel 1 didapatkan bahwa kombinasi dengan nilai X bernilai 10, Y bernilai 10, dan L bernilai 40 memiliki akurasi terbaik dengan nilai 75.8%.

4.2 Pengujian Perbandingan *stopword Term Based Random Sampling* dengan tanpa *stopword removal*

Pada pengujian ini ditujukan untuk membandingkan hasil evaluasi yang didapatkan jika menggunakan *Stopword Term Based Random Sampling* dan dibandingkan dengan tanpa menggunakan *Stopword Removal*. Penggunaan parameter yang digunakan dalam proses pembuatan *stopword* dengan *Term Based Random Sampling* sesuai dengan pengujian sebelumnya yang memiliki nilai akurasi terbaik yaitu dengan X bernilai 10, Y bernilai 10, dan L bernilai 40. Hasil pengujian perbandingan *stopword Term Based Random Sampling* dengan tanpa proses *Stopword*

Removal akan ditampilkan pada Tabel 2.

Tabel 2 Hasil Pengujian tanpa *stopword* dan TBRS

Stopword	Accuracy	Precision	Recall	F-Measure
Tanpa Stopword	75,3%	64,7%	63%	63,8%
TBRS	75,8%	65,8%	63,6%	64,7%

Pada Tabel 2 didapatkan bahwa hasil evaluasi menggunakan *Term Based Random Sampling* sedikit lebih baik jika dibandingkan tanpa menggunakan *stopword*.

4.3 Pengujian perbandingan hasil evaluasi *stopword* Tala dan *stopword* *Term Based Random Sampling*

Pada pengujian ini ditujukan untuk membandingkan hasil evaluasi yang didapatkan jika menggunakan *stopword* Tala dan dibandingkan dengan *stopword* yang dihasilkan oleh algoritme *Term Based Random Sampling*. Penggunaan parameter yang digunakan dalam proses pembuatan *stopword* dengan *Term Based Random Sampling* sesuai dengan pengujian sebelumnya yang memiliki nilai akurasi terbaik yaitu dengan X bernilai 10, Y bernilai 10, dan L bernilai 40. Hasil pengujian perbandingan *stopword* *Term Based Random Sampling* dengan *stopword* Tala akan ditampilkan pada Tabel 3.

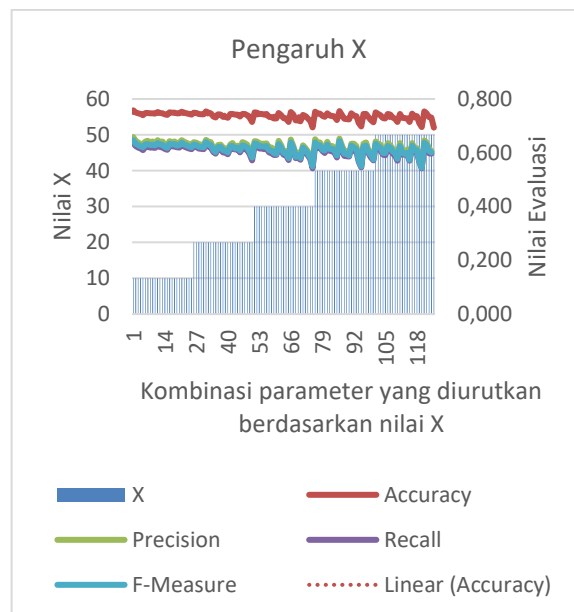
Tabel 3 Hasil Pengujian *stopword* Tala dan TBRS

Stopword	Accuracy	Precision	Recall	F-Measure
Tala	73,8%	61,9%	60,7%	61,3%
TBRS	75,8%	65,8%	63,6%	64,7%

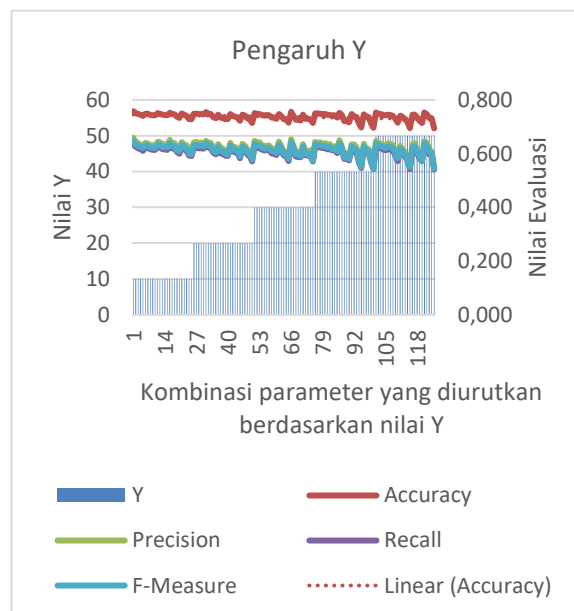
Pada Tabel 3 didapatkan bahwa akurasi keseluruhan dari penggunaan metode *Term Based Random Sampling* 2% lebih baik dibandingkan dengan penggunaan *stopword* Tala.

4.4 Hasil Analisis

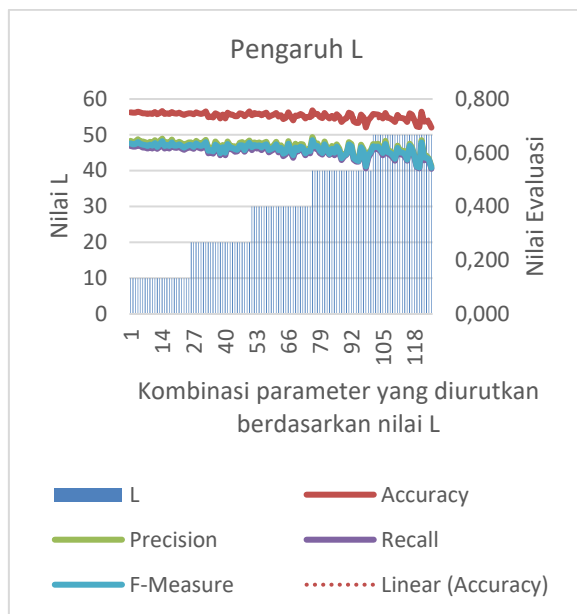
Pengujian kombinasi parameter X, Y, dan L dilakukan untuk mencari kombinasi parameter yang paling baik serta pengaruh tiap parameternya dalam evaluasi sistem. Grafik hasil pengujian parameter pengaruh X, Y, dan L dapat dilihat pada Gambar 2, Gambar 3, dan Gambar 4.



Gambar 2. Grafik Pengaruh X



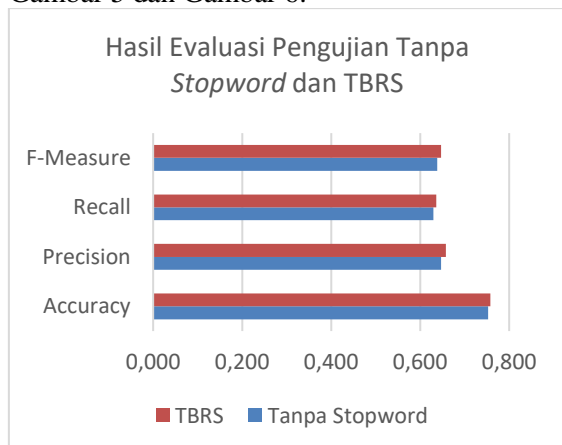
Gambar 3. Grafik Pengaruh Y



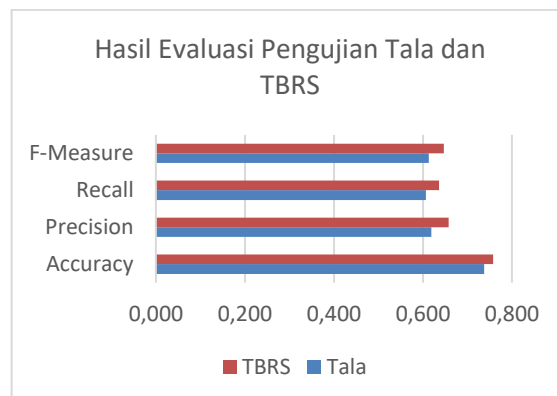
Gambar 4. Grafik Pengaruh L

Berdasarkan pada Gambar 4, 5, 6 dapat dilihat pada bagian bawah yang merupakan 125 kombinasi parameter yang sudah diurutkan berdasarkan nilai X, Y, dan L, dan diketahui bahwa garis *accuracy*, *precision*, *recall*, dan *f-measure* semakin kanan semakin menurun menunjukkan bahwa semakin besar nilai X, Y, dan L maka semakin tinggi kemungkinannya untuk *accuracy*, *precision*, *recall*, dan *f-measure* turun. Hal ini dibuktikan ketika X bernilai 10, garis *accuracy* lebih tinggi dan lebih stabil jika dibandingkan dengan X bernilai 50, hal ini berlaku juga untuk kedua parameter lainnya yaitu Y dan L.

Pengujian selanjutnya merupakan perbandingan antara algoritme *Term Based Random Sampling*, tanpa menggunakan *stopword*, dan menggunakan *stopword* Tala yang akan ditunjukkan dalam grafik pada Gambar 5 dan Gambar 6.



Gambar 5 Hasil Evaluasi Pengujian Tanpa Stopword dan TBRS



Gambar 6 Hasil Evaluasi Pengujian Tala dan TBRS

Pada Gambar 7 dan Gambar 8 didapatkan bahwa hasil evaluasi keseluruhan dari penggunaan metode *Term Based Random Sampling* ini sedikit lebih baik 0,5% dibandingkan dengan tanpa menggunakan proses *stopword removal* dan 2% lebih baik dibandingkan dengan penggunaan *stopword* Tala. Metode tanpa *stopword* memiliki *macroaverage accuracy* sebesar 75,3%, *macroaverage precision* sebesar 64,7%, *macroaverage recall* sebesar 63,0%, *macroaverage f-measure* sebesar 63,8% dan untuk metode dengan *stopword* Tala *macroaverage accuracy* sebesar 73,8%, *macroaverage precision* sebesar 61,9%, *macroaverage recall* sebesar 60,7%, *macroaverage f-measure* sebesar 61,3% memiliki sedangkan untuk metode dengan *Term Based Random Sampling* memiliki *macroaverage accuracy* sebesar 75,8%, *macroaverage precision* sebesar 65,8%, *macroaverage recall* sebesar 63,6%, *macroaverage f-measure* sebesar 64,7%. Sehingga dapat disimpulkan bahwa penggunaan *Term Based Random Sampling* sedikit lebih baik dalam kasus ini.

5. KESIMPULAN DAN SARAN

Berdasarkan pengujian yang sudah dilakukan terhadap parameter X, Y, dan L bahwa semakin besar nilai X, Y, dan L maka semakin tinggi pula kemungkinannya evaluasi sistem untuk menurun. Hal ini dibuktikan ketika X bernilai 10, hasil evaluasi lebih tinggi dan lebih stabil jika dibandingkan dengan X ketika bernilai 50, hal ini berlaku juga untuk kedua parameter lainnya yaitu Y dan L.

Penggunaan metode pembentukan *stopword* *Term Based Random Sampling* untuk analisis sentimen dengan *Naïve Bayes* dapat diterapkan dengan baik, hal ini dapat dilihat dengan

meningkatnya akurasi sistem yang dilakukan sebanyak 10-fold ketika menggunakan stopword *Term Based Random Sampling* sebesar 0,5% jika dibandingkan dengan tidak menggunakan proses *stopword removal*.

Berdasarkan pengujian perbandingan antara *Naïve Bayes* dan *stopword Term Based Random Sampling* mendapatkan rata-rata akurasi dari 10-fold, *stopword Term Based Random Sampling* memiliki akurasi sebesar 75,8% sedangkan jika menggunakan *stopword* Tala adalah sebesar 73,8%. Penggunaan *stopword Term Based Random Sampling* terbukti dapat meningkatkan akurasi pada analisis sentimen dengan *Naïve Bayes* sebesar 2%.

Adapun saran yang dapat membuat penelitian lebih baik adalah adanya normalisasi kata atau perbaikan kata pada tahap pre-processing terhadap *tweet* yang memiliki kesalahan penulisan kata. Selain itu *data* yang digunakan lebih baik dalam pemilihannya.

6. DAFTAR PUSTAKA

- Arnani, M., 2020. KOMPAS. [Online] Available at: <https://www.kompas.com/tren/read/2020/03/13/111245765/kasus-pertama-virus-corona-di-china-dilacak-hingga-17-november-2019>
- Ramadhan, A., Nugraheny, D. E. & Maharani, T., 2020. KOMPAS. [Online] Available at: <https://nasional.kompas.com/read/2020/09/05/15204581/update-kembali-bertambah-di-atas-3000-kasus-covid-19-lewati-190000?page=all>
- Putsanra, D. V., 2020. tirto. [Online] Available at: <https://tirto.id/apa-itu-new-normal-dan-bagaimana-penerapannya-saat-pandemi-corona-fCSg>
- Sa'rony, A., Adikara, P. P. & Wihandika, R. C., 2019. Analisis Sentimen Kebijakan Pemindahan Ibukota Republik Indonesia dengan Menggunakan Algoritme Term-Based Random Sampling dan Metode Klasifikasi Naïve Bayes. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, pp. 10086-10094.
- Dila Purnama Sari, D. E., Sari, Y. A. & Furqon, M. T., 2020. Pembentukan Daftar Stopword menggunakan Zipf Law dan Pembobotan Augmented TF - Probability IDF pada Klasifikasi Dokumen Ulasan Produk. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, pp. 406-412.
- Rahutomo, F. & Ririd, A. R. T. H., 2018. EVALUASI DAFTAR STOPWORD BAHASA INDONESIA. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, pp. 41-48.
- Gaddam, S. H. R., 2019. Text Preprocessing in Natural Language Processing. [Online] Available at: <https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8>
- Lo, R. T.-W., He, B. & Ounis, I., 2005. Automatically Building a Stopword List for an Information Retrieval System, Glasgow, UK: Department of Computing Science.
- Jones, S., 2004. A Statistical Interpretation of Term Specificity and Its Retrieval. *Journal Of Documentation*, 60(5), pp. 11-21.
- Prabowo, D. A., Fhadli, M., Najib, M. A. & Fauzi, H. A., 2016. TF-IDF-Enhanced Genetic Algorithm Untuk Extractive Automatic Text Summarization. *Jurnal Teknologi Informasi dan Ilmu Komputer*, 3(3), pp. 208-215.
- Sawla, S., 2018. Introduction to Naive Bayes for Classification. [Online] Available at: <https://medium.com/@srishtisawla/introduction-to-naive-Bayes-for-classification-baefefb43a2d>
- Rahman, A., Wiranto & Doewes, A., 2017. Online News Classification Using Multinomial Naive Bayes. *ITSMART: Jurnal Ilmiah Teknologi dan Informasi*.