Deteksi Perilaku Depresi dengan *Sentiment Analysis* pada Media Sosial Reddit Menggunakan *Feature Selection Information Gain* dan *Categorical Proportional Difference* dengan Algoritma Multinomial Naïve Bayes

Tugas Akhir

diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

dari Program Studi S1 Informatika

Fakultas Informatika

Universitas Telkom

1301161771

M Reza Prawira S

Program Studi Sarjana S1 Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2020

LEMBAR PENGESAHAN

Deteksi Perilaku Depresi dengan *Sentiment Analysis* pada Media Sosial Reddit Menggunakan *Feature Selection Information Gain* dan *Categorical Proportional Difference* dengan Algoritma *Multinomial Naïve Bayes*

Detection of Depression Behavior with Sentiment Analysis on Reddit Social Media Using *Information Gain* and *Categorical Proportional Difference* Feature Selection with Multinomial Naïve Bayes Algorithm

NIM : 1301161771

M Reza Prawira S

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi Sarjana S1 Informatika

Fakultas Informatika

Universitas Telkom

Bandung, Tanggal/Juli/2020

Menyetujui

Pembimbing I,	Pembimbing II,
Prof. Dr. Adiwijaya, S.Si, M.Si	Said Al Faraby, S.T., M.Sc
00740046	15890019

Ketua Program Studi
Sarjana S1 Informatika

Niken Dwi W.C., S.T., M.Kom., Ph.D.
NIP. 00750052

LEMBAR PERNYATAAN

Dengan ini saya, M Reza Prawira S, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Deteksi Perilaku Depresi dengan *Sentiment Analysis* pada Media Sosial Reddit Menggunakan *Feature Selection Information Gain* dan *Categorical Proportional Difference* dengan Algoritma *Multinomial Naïve Bayes*, beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang belaku dalam masyarakat keilmuan. Saya siap menanggung resiko/sanksi yang diberikan jika di kemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam buku TA atau jika ada klaim dari pihak lain terhadap keaslian karya,

Bandung, 25 Agustus 2020

Yang Menyatakan

M Reza Prawira S

Deteksi Perilaku Depresi dengan Sentiment Analysis pada Media Sosial Reddit Menggunakan Feature Selection *Information Gain* dan *Categorical Proportional Difference* dengan Algoritma *Multinomial Naïve Bayes*

M. Reza Prawira S[1], Adiwijaya[2], Said Al Faraby[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]mrezaprwr@students.telkomuniversity.ac.id, [2]adiwijaya@telkomuniversity.ac.id,
[3]saidalfaraby@telkomuniversity.ac.id

Abstrak

Depresi merupakan penyakit mental yang umum ditemukan di dunia dan diasosiasikan menjadi penyebab utama disabilitas. Layanan pengobatan penyakit mental di dunia yang belum memadai memberikan motivasi untuk dilakukan upaya pencegahan. Penelitian di bidang kesehatan mental yang masih kekurangan data kuantitatif akibat kompleksitas penyakit mental menjadikan media sosial sebagai sumber data yang berpotensi dalam menciptakan upaya pencegahan penyakit mental khususnya depresi. Pada penelitian ini, *sentiment analysis* digunakan untuk mendeteksi perilaku depresi pada media sosial Reddit. Analisis kata ganti orang dan pola kata yang mengikuti dilakukan untuk melihat pola bahasa pada data depresi. Berikutnya, beberapa eksperimen dilakukan untuk mendapatkan performa model terbaik, seperti perbandingan jenis *preprocessing*, perbandingan *feature selection Information Gain* dan *Categorical Proportional Difference*, dan pencarian parameter *smoothing* terbaik dalam *Multinomial Naïve Bayes*. Hasil penelitian menunjukkan bahwa terdapat peningkatan performa model akibat dari *stopword removal* dan reduksi kata dengan *stemming*. Pada perbandingan *feature selection*, *Information Gain* menghasilkan subset fitur terbaik sebanyak 40% dari total fitur dan berhasil memberikan peningkatan akurasi terbaik sebanyak 5.59% menjadi 87.82% dan peningkatan f1-score sebanyak 3.91% menjadi 89.74%. Terakhir, tuning parameter *smoothing* dengan nilai alpha sebesar 0.1 pada *Multinomial Naïve Bayes* menghasilkan peningkatan akurasi terbaik sebanyak 2.54% menjadi 84.77% dan peningkatan f1-score sebanyak 1.56% menjadi 87.39%.

Kata kunci : Reddit, Sentiment Analysis, *Information Gain*, *Categorical Proportional Difference*, *Multinomial Naïve Bayes*

Abstract

Depression is a mental illness that is commonly found in the world and is associated with being the main cause of disability. Inadequate treatment services for mental illness in the world provide motivation for prevention efforts. Research in the field of mental health, which lacks quantitative data due to the complexity of mental illness, has made social media a potential source of data in creating efforts to prevent mental illness, especially depression. In this study, sentiment analysis was used to detect depressive behavior on Reddit social media. Analyzes of pronoun pronouns and the word patterns that follow were performed to look at language patterns in the depression data. Next, several experiments were carried out to get the best model performance, such as comparison of preprocessing types, comparison of the Information Gain and Categorical Proportional Difference feature selection, and the search for the best smoothing parameters in Multinomial Naïve Bayes. The results showed that there was an increase in model performance due to stopword removal and word reduction by stemming. In the feature selection comparison, Information Gain produced the best feature subset as much as 40% of the total features and managed to provide the best accuracy increase by 5.59% to 87.82% and an increase in f1-score by 3.91% to 89.74%. Finally, tuning the smoothing parameter with an alpha value of 0.1 on Multinomial Naïve Bayes resulted in an increase in the best accuracy of 2.54% to 84.77% and an increase in the f1-score by 1.56% to 87.39%.

Keywords: Reddit, Sentiment Analysis, *Information Gain*, *Categorical Proportional Difference*, *Multinomial Naïve Bayes*

1. Pendahuluan

Latar Belakang

Berdasarkan World Health Organization (WHO), penyakit depresi adalah penyakit mental yang paling umum ditemukan di dunia. Pada tahun 2017, lebih dari 300 juta orang memiliki penyakit depresi dengan peningkatan lebih dari 18% dari tahun 2005 hingga 2015 [1]. Peningkatan ini diasosiasikan dengan peningkatan penyakit lain

di dunia karena depresi menjadi penyebab utama disabilitas dan memiliki kontribusi besar terhadap keseluruhan penyakit global [2]. Selain itu, depresi juga diestimasi menjadi peringkat kedua sebagai penyebab utama disabilitas pada tahun 2020 [3].

Layanan pengobatan penyakit mental di dunia masih tergolong belum memadai [4]. Penderita penyakit mental di negara berkembang tercatat 76-85% tidak memiliki akses pengobatan yang tepat [4]. Faktanya, belum diketahui adanya uji laboratorium yang dapat diandalkan dalam melakukan diagnosis sebagian besar bentuk penyakit [5]. Hal ini menyebabkan upaya pencegahan lebih diutamakan.

Penelitian pada bidang kesehatan mental secara tradisional masih menggunakan survei, tes kepribadian dan wawancara akademik dalam pengumpulan data [6]. Penelitian di bidang ini masih kekurangan data kuantitatif yang tersedia karena adanya kompleksitas yang terdapat pada penyebab kesehatan mental dan stigma masyarakat yang masih memandang penyakit mental sebagai subjek yang tabu. Sebaliknya, media sosial telah banyak digunakan sebagai sumber data dalam banyak penelitian yang melibatkan hubungan antara penggunaan media sosial dan pola perilaku seperti stress ataupun depresi. Oleh karena itu pemanfaatan media sosial dapat menjadi alternatif baru dalam pencarian informasi mengenai kesehatan mental.

Pada penelitian ini dilakukan analisis sentimen yang bertujuan untuk mendeteksi perilaku depresi pada media sosial. Analisis sentimen atau *sentiment analysis* merupakan suatu pekerjaan yang mengidentifikasikan apakah sebuah opini yang disampaikan termasuk kategori positif atau negatif pada sebuah dokumen. Dalam konteks media sosial, *sentiment analysis* berperan dalam menentukan polaritas pada bahasa yang diekspresikan dengan menekankan pada identifikasi kecenderungan perilaku depresi.

Beberapa penelitian serupa telah dilakukan tetapi sebagian besar menggunakan domain *microblog* khususnya twitter seperti model index depresi populasi [7] dan model klasifikasi pembeda depresi, PTSD dan non depresi [8]. Penelitian lain pada *microblog* cina juga dilakukan menggunakan *sentiment analysis* untuk deteksi depresi [6]. Banyak penelitian yang dilakukan pada domain *microblog* khususnya twitter disebabkan penggunaan bahasa yang mendekati kehidupan sehari hari. Sebaliknya, penelitian pada domain forum belum banyak dilakukan. Forum tidak termasuk dalam *microblog* karena memungkinkan pengguna untuk berbagi informasi atau konten tanpa batasan karakter dan percakapan yang terjadi biasanya berpusat pada topik tertentu. Pada penelitian ini forum yang digunakan adalah Reddit. Reddit merupakan media sosial berorientasi agregasi berita, pemeringkatan konten, dan situs diskusi mengenai topik topik tertentu. Reddit dipilih karena memiliki kategori berdasarkan topik yang berkaitan dengan kesehatan mental dan kemudahan dalam menemukan pengguna yang telah didiagnosis penyakit kesehatan mental sehingga label data yang diberikan lebih valid. Walaupun begitu, penelitian pada reddit memiliki tantangan karena penggunaan bahasa dan karakter linguistik yang berbeda.

Topik dan Batasannya

Pada penelitian ini penulis melakukan percobaan mengidentifikasikan perilaku depresi pada sentimen pengguna media sosial Reddit. Analisis pola penggunaan bahasa pada data depresi dilakukan, akan tetapi fokus utama pada penelitian ini adalah untuk membangun klasifikasi teks yang paling tepat dalam mengidentifikasikan teks yang sudah dilabeli depresi dan non-depresi. Proses klasifikasi yang tepat didapatkan dari pemilihan jenis *preprocessing* data terbaik, metode pemilihan fitur dan parameter pada metode klasifikasi yang digunakan. Jenis *preprocessing* data pada penelitian ini terbagi menjadi *stopword removal*, *stemming* dan *lemmatiaztion*. Pada pemilihan fitur, metode yang digunakan adalah *Information Gain* (IG) dan *Categorical Proportional Difference* (CPD) dengan parameter fitur yang berbeda beda. Parameter jumlah fitur terbaik digunakan oleh IG, sedangkan batas *threshold* digunakan oleh CPD. Sedangkan metode klasifikasi yang digunakan adalah *Multinomial Naïve Bayes* dengan pengaturan parameter *laplace smoothing* berdasarkan beberapa nilai yang telah ditentukan.

Beberapa batasan masalah yang terdapat pada penelitian ini adalah, terbatasnya jumlah data pada dataset sebanyak 659, terbagi menjadi 343 data depresi dan 316 data non-depresi. Pengambilan data depresi dilakukan pada media sosial Reddit dengan cara manual berdasarkan cara yang dianggap paling optimal dalam memperoleh data yang paling tepat, sedangkan data non-depresi yang diperoleh dengan crawling otomatis dengan *library python* berasal dari subreddit diluar forum depresi yang berisi banyak sentimen positif. Data non-depresi yang diperoleh dianggap belum cukup ideal, karena tidak termasuk dalam lingkungan yang sama dengan data depresi. Hal ini disebabkan sulitnya mendapatkan data non-depresi yang tidak memiliki karakter yang sama dengan data depresi pada lingkungan atau forum depresi. Batasan lain adalah *splitting* data yang digunakan hanya sekali dan tidak menggunakan *cross validation.*

Tujuan

Penelitian ini bertujuan untuk menganalisis pola penggunaan bahasa pada data depresi dan membuat model dengan performa terbaik yang dapat mengidentifikasikan perilaku depresi pada teks media sosial. Penelitian diawali dengan melihat pola penggunaan kata pengganti orang pada dataset depresi. Berikutnya, eksperimen berupa perbandingan jenis *preprocessing*, *feature selection*, dan parameter metode klasifikasi dilakukan untuk mendapatkan performa model terbaik. Penelitian dimulai dengan analisis pengaruh *stopword removal* dan reduksi kata berdasarkan kata dasarnya berupa *stemming* dan *lemmatization*. Selanjutnya perbandingan *Information Gain*

(IG) dan *Categorical Proportional Difference* (CPD) dilakukan untuk mendapatkan subset yang menghasilkan performa model terbaik. Terakhir, tuning parameter *smoothing* dalam algoritma klasifikasi dilakukan untuk mendapatkan performa model terbaik.

2. Studi Terkait

Penelitian yang dilakukan oleh Rude, Gortner, dan Pennebaker [9] pada tahun 2004 berjudul *Language use of depressed and depression-vulnerable college students*, yang melakukan penelitian terhadap analisis teks essay yang ditulis oleh mahasiswa yang sedang depresi, pernah depresi dan tidak pernah depresi, ditemukan bahwa mahasiswa yang sedang terkena penyakit depresi menggunakan banyak kata ganti orang pertama, yaitu 'I' dibandingkan mahasiswa yang tidak pernah terkena penyakit depresi. Rude, Gortner dan Pennebaker menginterpretasi bahwa mahasiswa yang sedang terkena penyakit depresi lebih banyak menulis pengalaman yang lebih relevan dengan diri sendiri dan secara progressif lebih terjerat kepada dirinya sendiri dibandingkan mahasiswa yang pernah dan tidak pernah sama sekali terkena penyakit depresi.

Pada penelitian yang dilakukan oleh Andrew Yates, Arman Cohan dan Nazli Goharian [10] berjudul *Depression and Self-Harm Risk Assesment in Online Forums* tahun 2017 yang membahas metode untuk mengidentifikasi post yang dapat terindikasi membahayakan atau melukai diri sendiri pada komunitas yang memberikan support terhadap penderita penyakit kesehatan mental termasuk depresi. Penelitian ini membahas cara pengambilan data orang yang terindikasi penyakit depresi dengan melihat pola diagnosis akurasi tinggi, yaitu pengguna yang melakukan *self-reported diagnosis* dimana pengguna memang sudah dinyatakan terkena depresi dan memberitahukan mengenai diagnosis tersebut dengan pola kata seperti, 'I've been diagnosed with', 'I'm recently diagnosed', 'I had gotten the diagnosis', dan sebagainya. Pada penelitian ini juga ditentukan jumlah minimum post yang terindikasi depresi agar suatu pengguna dapat diklasifikasikan sebagai orang yang terkena penyakit depresi.

Penelitian yang dilakukan oleh O'Keefe, Tim dan Koprinska [11] berjudul *Feature Selection and Weighting Methods in Sentiment Analysis* pada tahun 2009 melakukan percobaan menggunakan beberapa seleksi fitur yang disertai pembobotan fitur dengan algoritma klasifikasi naive bayes dan support vector machine. Penelitian ini menemukan bahwa seleksi fitur yang digunakan dapat meningkatkan performa klasifikasi dengan akurasi sebesar 87.15% yang disertai penggunaan pembobotan fitur yang sesuai. Selain peningkatan performa klasifikasi, penggunaan seleksi fitur juga dapat mengurangi beban komputasi karena mereduksi penggunaan fitur hingga 36%.

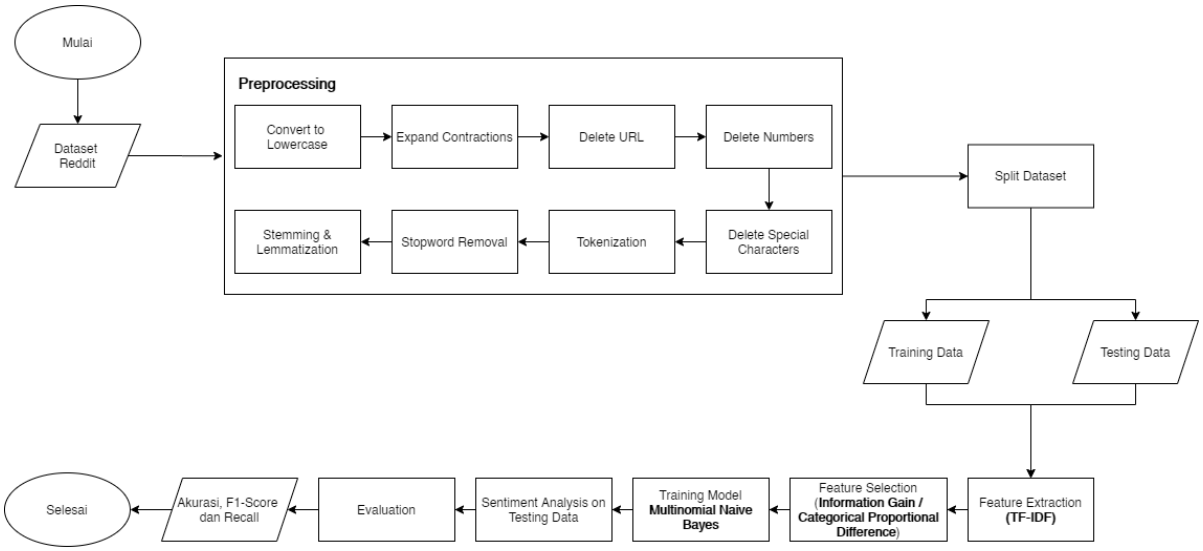
Penelitian yang dilakukan oleh Asriyanti Pratiwi dan Adiwijaya [12] berjudul *On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis* pada tahun 2017 dibahas mengenai *sentiment analysis* dengan seleksi fitur kombinasi antara *Information Gain* dan *Document Frequency* (IGDFFS) pada data ulasan film. Hasil penelitian ini adalah seleksi fitur dengan IGDFFS berhasil mengurangi fitur yang tidak relevan hingga 90%.

Pada penelitian yang dilakukan oleh Pang, Lee dan Vaithyanathan [13] yang berjudul *Thumbs up? Sentiment Classification using Machine Learning Techniques* pada tahun 2002 dibahas efektifitas dalam mengaplikasikan teknik machine learning sebagai metode klasifikasi pada analisis sentimen dibandingkan dengan klasifikasi teks berdasarkan topik. Algoritma klasifikasi yang digunakan pada penelitian ini adalah Naive Bayes, Maximum Entropy dan Support Vector Machine (SVM). Pada penelitian dilakukan tuning pada pilihan fitur fitur tertentu untuk melihat pengaruhnya terhadap performa ketiga algoritma tersebut. Hasil yang didapatkan adalah dalam segi performa, Naive Bayes cenderung memiliki performa terjelek dan SVM memiliki performa terbaik. Maximum entropy cenderung memberikan performa lebih baik dibandingkan Naive Bayes tetapi tidak selalu. Kemudian penggunaan fitur yang paling efektif adalah *unigram presence* karena fitur alternatif lain yang digunakan tidak memberikan performa yang lebih baik secara konsisten. Walaupun begitu akurasi pada klasifikasi sentimen tidak sebanding dengan akurasi yang didapatkan pada klasifikasi teks berdasarkan topik.

3. Sistem yang Dibangun

3.1. Rancangan Sistem

Pada penelitian ini, metode algoritma yang digunakan adalah algoritma Multinomial Naïve Bayes dengan *feature selection Information Gain* (IG) dan *Categorical Proportional Difference* (CPD). Data yang digunakan berasal dari beberapa forum dari media sosial reddit atau subreddit yang berkaitan dengan kesehatan mental untuk data depresi dan subreddit r/happy untuk data non-depresi. Sistem yang akan dibangun pada penelitian ini dapat digambarkan prosesnya sebagai berikut.



Gambar 1. Skema Rancangan Sistem

3.2. Pengumpulan *Dataset*

Dataset yang digunakan pada penelitian ini berasal dari pengumpulan data secara mandiri oleh penulis. Pada data depresi, pengumpulan data dilakukan dengan mencari data secara menual yang berasal dari *post post* pada beberapa subreddit depresi, seperti r/depression, r/mentalhealth, r/SuicideWatch, dan sebagainya. Pada data non-depresi, pengumpulan dilakukan dengan cara *crawling* python menggunakan API Reddit yang berasal dari subreddit r/happy.

Pengumpulan data depresi dilakukan dengan mencari post yang memiliki pola diagnosis dengan akurasi tinggi. Pola ini ditandai dengan adanya ungkapan ungkapan seperti 'I was diagnosed with', 'I have been diagnosed', 'I've already been diagnosed' dan masih banyak lagi. Pola ini menandakan user yang melakukan posting telah terdiagnosis depresi sehingga terdapat diagnosis yang telah dilaporkan sendiri dan bukan berasal dari asumsi diri sendiri. Setelah post yang mengandung pola tersebut ditemukan, *user* yang mengirimkan post tersebut ditelusuri halaman profilenya untuk dilihat berapa banyak post post yang dikirim ke subreddit depresi. Jika total post yang dikirim ke subreddit depresi berjumlah lebih dari atau sama dengan dua, maka semua post pada subreddit depresi tersebut diambil. Semua post yang diambil adalah post post yang dikirim dari tanggal 1 Januari 2017 sampai 20 Maret 2020 dengan jumlah 343 post.

Pada pengumpulan data non-depresi *crawling* python dilakukan pada subreddit r/happy yang merupakan forum berisi post post cerita bahagia atau sesuatu yang positif. Data yang diambil pada subreddit ini sebanyak 315 post dari tanggal 1 Januari 2020 sampai 20 Maret 2020.

Setelah kedua jenis data dikumpulkan dan dihapus duplikatnya, terdapat sebanyak 343 data depresi dan 311 data non-depresi, sehingga total data pada dataset berjumlah 654 data. Data yang telah dikumpulkan kemudian akan diamati statistik deskripsinya berdasarkan panjang teks, jumlah kata dan persentase stopword, sehingga dihasilkan informasi pada tabel berikut

Tabel 1. Panjang teks pada dataset

Statistik	Depresi	Non-Depresi
Rata-rata	467.46	189.42
Minimum	29	41
25%	273	84.5
50%	450	135
75%	631.5	244
Maximum	1203	815

Tabel 2. Jumlah kata pada dataset

Statistik	Depresi	Non-Depresi
Rata-rata	89.05	35.32
Minimum	6	6
25%	50.5	16
50%	87	25

75%	121.5	45.5
Maximum	221	149

Table 3. Persentase stopword pada dataset

Statistik	**Depresi**	**Non-Depresi**
Rata-rata	40.67	37.73
Minimum	12.5	0
25%	37.09	31.91
50%	41.12	38.71
75%	44.74	43.84
Maximum	61.76	63.16

Berdasarkan statistik panjang teks, data depresi memiliki panjang teks yang lebih signifikan dengan panjang hampir tiga kali dari panjang teks data non-depresi. Berikutnya, jika dilihat dari jumlah kata, data depresi juga memiliki jumlah kata yang signifikan lebih banyak dengan jumlah kata lebih banyak hampir tiga kali lebih banyak dibandingkan data non-depresi. Terakhir, jika dilihat dari persentase jumlah stopword terhadap keseluruhan jumlah kata, keduanya tidak memiliki perbedaan yang signifikan karena 75% dari keseluruhan data pada depresi dan non-depresi memiliki persentase kurang lebih 43% jumlah stopword dibandingkan jumlah kata keseluruhan.

3.3. *Preprocessing*

Setelah pengambilan data dilakukan, proses *preprocessing* dilakukan. *Preprocessing* adalah proses transformasi data mentah ke bentuk yang lebih terstruktur. Proses ini diperlukan untuk mengubah data yang menjadi bentuk yang dapat lebih mudah dimengerti model.

Tahap pertama dalam *preprocessing* adalah melakukan konversi teks ke dalam bentuk standar yaitu pengubahan menjadi huruf kecil. Proses ini dilakukan untuk menciptakan konsistensi teks pada data agar fitur yang didapatkan lebih optimal. Berikutnya ungkapan pada bahasa inggris yang disingkat, contohnya ‘I’m’, ‘They haven’t’, ‘you should’ve’ diperluas menjadi bentuk bakunya sehingga menjadi ‘I am’, ‘They have not’, dan ‘you should have’. Kemudian URL dan angka dihapus pada data agar tidak dijadikan fitur. Terakhir, karakter spesial yang bukan termasuk alfabet dan numerik, seperti tanda baca dan simbol simbol asing dihapus. Berikutnya, proses tokenization yang memotong kalimat menjadi kata kata terpisah dilakukan agar dapat dihapus stopword di dalamnya. Stopword adalah kata kata umum yang tidak memiliki arti signifikan dan hanya berguna sebagai koherensi kalimat, contohnya adalah ‘the’, ‘a’, ‘on’ dan sebagainya. Setelah stopword dihapus, teknik reduksi kata berupa *stemming* dan *lemmatization* dilakukan untuk mereduksi kata menjadi bentuk kata dasarnya.

3.4. *Feature Extraction*

Feature extraction atau ekstraksi fitur berguna sebagai pembobotan teks menjadi bentuk vektor vektor fitur yang diperlukan sebagai input proses klasifikasi. Pada penelitian ini, *Term Frequency Inverse Document Frequency* (TF-IDF) digunakan sebagai *feature extraction* yang melakukan pembobotan dengan cara menentukan frekuensi relatif dari suatu kata yang berada dalam dokumen tertentu dibandingkan dengan proporsi jumlah kebalikan dari kata pada dokumen yang sama. *Term Frequency* (TF) didefinisikan sebagai jumlah kemunculan kata pada dokumen terhadap jumlah seluruh kata pada dokumen tersebut. TF diformulasikan sebagai berikut

$$\mathrm{TF(t,d)} = \frac{jumlah\ kemunculan\ kata\ (t)\ pada\ dokumen\ (d)}{total\ jumlah\ kata\ pada\ dokumen\ (d)} \quad (1)$$

Inverse Document Frequency (IDF) merupakan nilai dari suatu kata terhadap keseluruhan dokumen. Kata yang jarang muncul memiliki nilai IDF yang tinggi sedangkan kata yang sering muncul memiliki nilai IDF yang rendah. IDF diformulasikan sebagai berikut

$$\mathrm{IDF(t,D)} = \log\left(\frac{total\ jumlah\ dokumen\ (D)}{jumlah\ dokumen\ yang\ terdapat\ kata\ (t)}\right) \quad (2)$$

Berikutnya, nilai TFIDF yang merupakan hasil perkalian dari TF dan IDF dapat diformulasikan sebagai berikut

$$\mathrm{TFIDF(t,d,D)} = \mathrm{TF(t,d)}\ .\ \mathrm{IDF(t,D)} \quad (3)$$

3.4. *Feature Selection*

Feature selection berperan dalam mereduksi jumlah fitur agar diperoleh subset fitur paling optimal yang dapat merepresentasikan keseluruhan data. *Feature selection* yang digunakan pada penelitian ini adalah *Information Gain* (IG) dan *Categorical Proportional Difference* (CPD).

Feature selection IG berguna dalam menghitung nilai reduksi dari entropy atau ketidapastian dari sebuah data. Entropy suatu data berguna sebagai pengukur seberapa banyak nilai informasi yang ada dalam suatu data atau tingkat keberagaman data. Data dengan keberagaman yang sedikit atau mudah ditebak memiliki nilai entropy yang rendah, sedangkan data dengan keberagaman yang banyak atau sulit ditebak memiliki nilai entropy yang tinggi. Formula information gain pada dokumen diformulasikan sebagai berikut

$$Info(D) \quad = -\sum_{i=1}^{m}(P_i)\log_2(P_i) \tag{4}$$

dengan m sebagai jumlah kelas pada dataset, P_i merepresentasikan probabilitas dari dokumen yang ingin ditentukan yang dilabeli sebagai kelas C_i atau dikalkulasikan sebagai C_i / D. Fungsi log basis dua menetapkan nilai informasi dalam bentuk bits. Jika sebuah data dalam dataset D akan diklasifikasikan dalam beberapa atribut A $\{a_1 \ldots., a_v\}$, maka dataset D akan terbagi menjadi partisi sebanyak v $\{ D_1, D_2, \ldots D_v\}$. Nilai informasi dari suatu kelompok atribut A ini diukur dengan formula:

$$Info_A(D) \quad = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \text{x Info } (D_j) \tag{5}$$

Nilai ini bisa disebut juga sebagai entropi kondisional dengan D_j / D sebagai bobot dari partisi ke-j dan Info D_j sebagai entropy dari partisi D_j. Terakhir, nilai *Information Gain* dari suatu atribut A dapat diformulasikan sebagai berikut

$$Information\ Gain\ (A) \quad = \text{Info } (D) - Info_A(D) \tag{6}$$

Categorical Proportional Difference merupakan *feature selection* yang diperkenalkan oleh Simeon & Hilderman [14] sebagai metrik yang memberitahu seberapa dekat kedua angka. CPD bekerja dengan cara mengkalkulasi selisih frekuensi dokumen positif dan frekuensi dokumen negatif yang terdapat kemunculan kata tertentu dibagi jumlah kedua jenis dokumen yang terdapat kemunculan kata terebut. Formula CPD dapat dikalkulasikan sebagai berikut

$$\text{CPD} = \frac{|PositiveDF - Negative\ DF|}{PositiveDF + Negative\ DF} \tag{7}$$

Pada perhitungan nilai CPD, jika suatu fitur muncul dominan di dalam dokumen positif atau negative saja maka nilai CPD-nya mendekati 1, sedangkan apabila suatu fitur muncul hampir sama banyaknya dalam kedua jenis dokumen maka nilai CPD mendekati 0.

3.5. *Classification*

Pada penelitian ini, metode klasifikasi Multinomial Naïve bayes digunakan sebagai model klasifikasi untuk *sentiment analysis*. Multinomial Naïve bayes adalah algoritma yang merupakan bagian dari teorema bayes dengan asumsi naïve bahwa setiap fitur memiliki hubungan yang independent dari fitur lain. Multinomial Naïve bayes merupakan hasil produk dari *prior probability* dan *likelihood probability*. *Prior probability* merupakan distribusi probabilitas yang mengekspresikan keyakinan suatu hal sebelum beberapa bukti diperhitungkan. Sedangkan *Likelihood probability* merupakan sebuah probabilitas suatu *event* yang memiliki hubungan dengan *event* lain yang telah terjadi.

Sebelum dilakukan klasifikasi, data dipecah menjadi 70% data train dan 30% data test. Data train diproses pada perhitungan *prior probability* dan *likelihood probability* sedangkan data test diproses pada perhitungan *maximum a posteriori* (MAP). *Prior probability* diformulasikan sebagai berikut

$$\text{P(c)} = \frac{N_c}{N} \tag{8}$$

dengan N_c sebagai jumlah dokumen pada suatu kelas c dan N sebagai jumlah keseluruhan suatu dokumen. Sedangkan *Likelihood probability* diformulasikan sebagai berikut

$$P(t|c) = \frac{T_{ct} + \alpha}{\Sigma_{x=1}^{|V|} T_{ct} + \alpha|V|} \quad (9)$$

dengan T_{ct} sebagai jumlah kata t dalam dokumen yang berada pada kelas c, |V| merupakan ukuran vocabulary. *Laplace smoothing* yang ditandai dengan notasi α digunakan pada *likelihood probability* dengan cara menambahkan jumlah kemunculan kata t dengan satu angka positif ($\alpha > 0$) untuk menghidari kalkulasi probabilitas yang menghasilkan nilai 0. Umumnya nilai α yang digunakan adalah 1. *Maximum a posteriori* (MAP) atau disebut juga dengan c_{MAP} diformulasikan sebagai berikut

$$c_{MAP} = \text{argmax } \mathrm{P}(t|c)\mathrm{P}(c) \quad (10)$$

Dengan melakukan perhitungan c_{MAP} pada data test maka dicari hasil argument maximum nilai tertinggi dari *posterior probability*-nya. Perhitungan c_{MAP} dapat menghasilkan nilai 0 jika ada kata yang terdapat pada kategori positif tetapi tidak ada pada kategori negatif dan sebaliknya.

3.6. *Evaluation*

Setelah dilakukan proses klasifikasi, tahap terakhir adalah mengukur performa hasil kinerja klasifikasi atau disebut juga sebagai evaluasi. Pada penelitian ini, *confusion matrix* digunakan untuk memeriksa hasil klasifikasi. Dengan *confusion matrix*, hasil prediksi benar dan tidak benar yang dibuat model klasifikasi dibandingkan dengan klasifikasi aktual dalam data train. Berikut tabel *confusion matrix*.

Tabel 4. Confusion Matrix

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Dalam konteks penelitian ini, *True Positive* (TP) merupakan perilaku depresi yang diprediksi depresi, *False Positive* (FP) merupakan perilaku non-depresi yang diprediksi depresi, *False Negative* (FN) merupakan perilaku depresi yang diprediksi non-depresi, dan *True Negative* (TN) adalah perilaku non-depresi yang diprediksi non-depresi. Dari keempat matriks tersebut, dapat dibuat empat jenis metrik evaluasi, yaitu

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (11)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (12)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (13)$$

$$F1\ Score = \frac{2}{\left(\frac{1}{Recall} + \frac{1}{Precision}\right)} \quad (14)$$

Pada penelitian ini, akurasi dijadikan metrik evaluasi untuk mengukur performa model dalam melakukan klasifikasi. Selain akurasi, f1-score dan recall juga digunakan sebagai metrik evaluasi. F1-score berperan sebagai metrik yang melihat keseimbangan antara precision dan recall, sedangkan recall saja berperan untuk meminimalkan jumlah kasus *False Negative* (FN) dalam data hasil prediksi. Hasil prediksi yang termasuk FN adalah data depresi yang dideteksi sebagai non-depresi dimana data ini dianggap memiliki resiko yang lebih besar dalam pengaplikasiannya di dunia medis sehingga jumlahnya perlu diminimalkan.

4. Evaluasi

Analisis pola penggunaan bahasa

Sebelum dilakuan penelitian lebih lanjut, diperlukan eksplorasi awal pada data teks. Hal ini untuk melihat pola pola yang terdapat pada data yang telah diambil. Eksplorasi dilakukan dengan melihat posisi kata ganti orang terhadap pola kata depresi. Kata ganti orang yang dianalisis adalah kata kata yang mengawali kalimat seperti, ‘I’, ‘You’, ‘We’, ‘They’, ‘He’ dan ‘She’. Kata ‘It’ tidak termasuk karena sering muncul pada akhir kalimat.

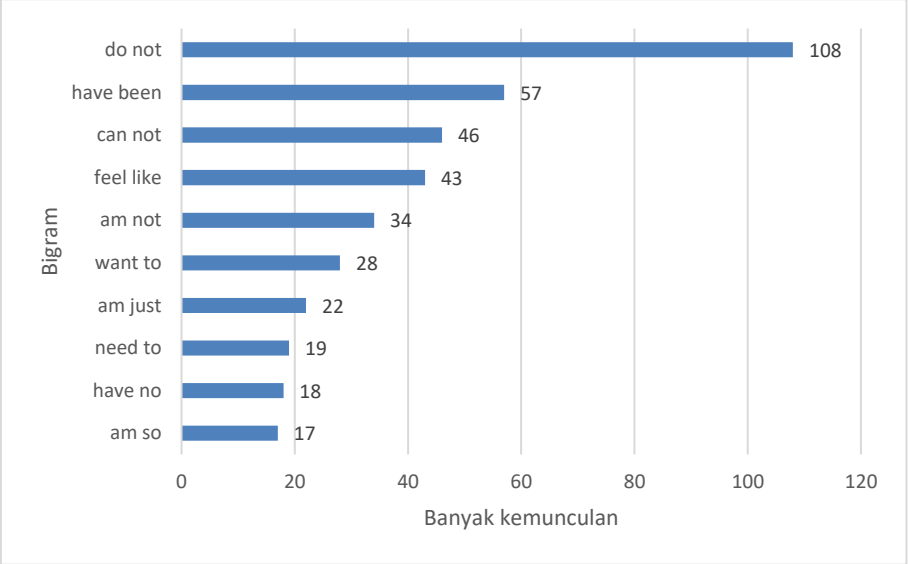
Dalam menentukan banyak kata ganti orang yang digunakan pada data depresi, penulis membuat n-gram kata dengan n=5 pada tiap data. Kemudian, penulis mengambil n-gram data yang hanya diawali oleh kata ganti yang telah ditentukan sebelumnya sehingga hasil yang didapatkan sebagai berikut.

Tabel 5. Daftar kata ganti orang

Feature	Total
I	2224
You	96
We	42
They	140
He	120
She	53

Berdasarkan penelitian oleh Rude tahun 2004 [9], individu yang pernah terkena depresi lebih dominan menggunakan kata ganti ‘I’ dengan jumlah yang signifikan. Hal ini juga juga dapat diamati pada dataset depresi reddit. Hasil eksplorasi menunjukkan bahwa terdapat 2224 jumlah kata ganti ‘I’ atau sebanyak 83.14% dari keseluruhan kata ganti yang dianalisis. Kata ganti orang terbanyak kedua adalah ‘They’ diikuti oleh ‘He’ pada posisi ketiga.

Analisis berikutnya adalah melihat pola kata pada setiap kata ganti orang yang diberikan. Untuk mengetahui pola tersebut, penulis mengambil bigram yang mengikuti setiap kata ganti orang dalam n-gram tadi. Bigram kemudian dikumpulkan dan dilihat distribusinya.



Grafik 1. Distribusi 10 bigram terbanyak yang mengikuti kata ganti

Grafik 1 menunjukkan distribusi sepuluh bigram terbanyak yang mengikuti kata ganti. Berdasarkan grafik, bigram ‘do not’ memiliki kemunculan terbanyak dengan jumlah 108 kali kemunculan. Bigram terbanyak kedua adalah ‘have been’ dan bigram terbanyak ketiga adalah ‘can not’. Hasil ini menunjukkan bahwa penggunaan kata ganti orang pada data depresi biasanya diikuti konteks negatif yang dapat berupa pernyataan tidak mampu, tidak tahu, tidak setuju dan sebagainya yang berkonotasi negatif. Karena mayoritas n-gram diawali dengan kata ganti ‘I’, maka diperlukan analisis lebih lanjut terhadap setiap kata pada bigram terhadap masing masing kata ganti orang agar dapat diketahui apakah setiap kata ganti orang diikuti dengan konteks negatif atau tidak. Oleh karena itu, penulis memisahkan bigram ini menjadi kedua kata berbeda yang dinamakan kata-1 dan kata-2 untuk melihat distribusi kata yang paling banyak.

Tabel 6. Bigram kata-1 pada dataset depresi

Kata Ganti ‘I’		Kata Ganti ‘You’		Kata Ganti ‘We’		Kata Ganti ‘They’		Kata Ganti ‘He’		Kata Ganti ‘She’	
Kata	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**
am	326	are	10	have	6	are	29	was	11	is	8
have	262	can	7	re	4	were	9	is	9	was	4
was	166	have	4	are	3	do	8	would	6	asked	2
do	120	will	4	do	3	have	7	said	5	seemed	2
can	85	know	4	were	3	will	4	did	4	wants	2

feel	80	guys	4	could	2	can	4	will	4	can	2
just	79	with	3	live	2	all	3	does	4	also	2
know	52	do	2	had	2	could	3	had	4	would	2
will	50	for	2	met	1	just	3	got	4	just	2
would	45	and	2	first	1	did	3	told	2	thinks	1

Tabel 7. Bigram kata-2 pada dataset depresi

Kata Ganti 'I'		**Kata Ganti 'You'**		**Kata Ganti 'We'**		**Kata Ganti 'They'**		**Kata Ganti 'He'**		**Kata Ganti 'She'**	
Kata	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**	**Kata**	**Total**
not	236	not	5	not	7	not	15	not	10	me	11
to	130	i	4	to	5	to	7	me	9	to	3
a	61	just	4	in	2	me	4	to	7	does	2
i	60	a	4	trouble	1	down	3	and	3	the	2
been	59	to	3	ever	1	so	3	out	3	so	2
like	48	have	3	suffering	1	it	3	i	3	pretty	1
it	44	the	3	at	1	just	3	be	3	threatening	1
my	43	it	3	one	1	i	3	looking	3	did	1
so	37	what	2	financial	1	very	3	all	3	dead	1
just	35	in	2	know	1	do	2	the	2	now	1

Berdasarkan kedua tabel di atas, dapat terlihat pola, baik pada kata-1 dan kata-2 dari bigram. Pada kata-1, hampir semua kata ganti diikuti dengan *verb* to-be yang menunjukkan waktu yang sedang terjadi atau biasanya terjadi (*present tense*) kecuali kata ganti 'He' yang diikuti dengan to-be 'was'. Sedangkan pada kata-2, dapat dilihat bahwa hampir semua kata ganti orang menggunakan kata not setelah penggunaan *to-be*, kecuali pada kata ganti 'she'. Hal ini menunjukkan bahwa mayoritas data depresi yang diawali dengan beberapa kata ganti orang diatas menceritakan hal yang sedang terjadi atau biasanya terjadi dengan tambahan negasi terhadap *to-be* dan konteks setelahnya.

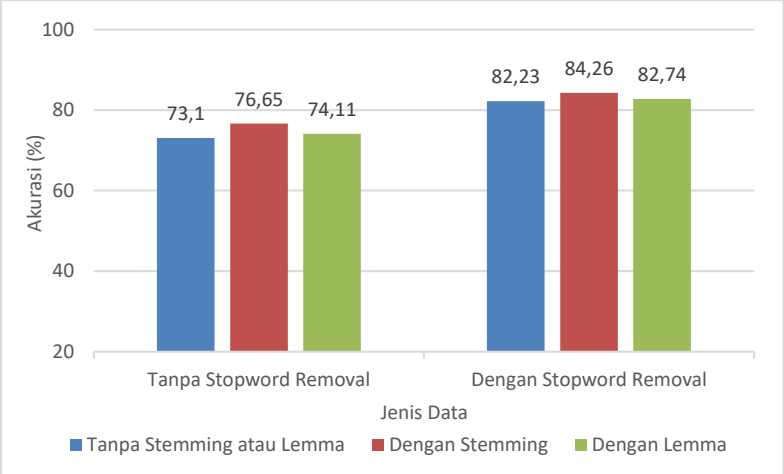
Berikutnya, ketiga skenario percobaan pada penelitian ini dilakukan untuk mendapatkan performa model klasifikasi terbaik. Skenario terbagi menjadi, pengaruh pengaturan berbagai jenis *preprocessing*, perbandingan metode seleksi fitur *Information Gain* (IG) dan *Categorical Proportional Difference* (CPD), dan pencarian parameter *smoothing* pada multinomial naïve bayes yang paling optimal.

Skenario pertama: Pengaruh jenis *preprocessing* terhadap performa model klasifikasi

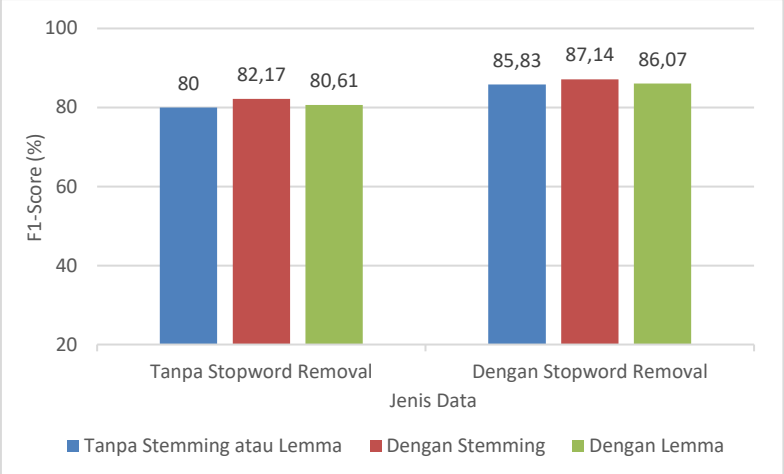
Jenis *preprocessing* yang dibedakan menjadi enam kelompok data hasil kombinasi tiga metode *preprocessing*, yaitu *stopword removal*, *stemming*, dan *lemmatization*. Perbandingan jenis *preprocessing* dilakukan pada kedua jenis dataset, yaitu

1. Dataset yang telah dinormalisasi dan tanpa *stopword removal*.
2. Dataset yang telah dinormalisasi dengan *stopword removal*.

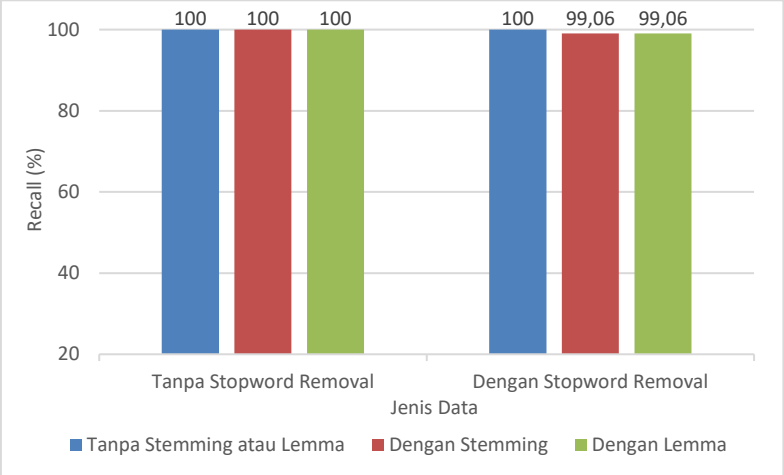
Klasifikasi menggunakan metode Multinomial Naïve Bayes tanpa *feature selection* didapatkan hasil sebagai berikut.



Grafik 2. Pengaruh jenis *preprocessing* terhadap akurasi model



Grafik 3. Pengaruh jenis *preprocessing* terhadap f1-score model



Grafik 4. Pengaruh jenis *preprocessing* terhadap recall model

Berdasarkan hasil pengujian pengaruh jenis *preprocessing* terhadap data yang telah dinormalisasi, data dengan *stopword removal* memberikan peningkatan performa model pada ketiga jenis data. Peningkatan akurasi sebanyak 9.13% dan f1-score sebanyak 5.83% pada data tanpa *stemming* atau *lemmatization.* Lalu, peningkatan akurasi sebanyak 7.61% dan f1-score sebanyak 4.97% pada data yang telah di-*stemming*. Terakhir, peningkatan akurasi sebanyak 8.63% dan f1-score sebanyak 5.46% pada data yang telah di-*lemmatization*. Walaupun terjadi peningkatan pada akurasi dan f1-score, recall mengalami penurunan sebanyak 0.94% pada data yang telah di-*stemming* dan *lemmatization*.

Pada reduksi kata berdasarkan kata dasarnya, *stemming* menghasilkan performa model yang lebih baik dibandingkan *lemmatization* sebagai metode reduksi fitur berdasarkan kata dasarnya. Pada data tanpa *stopword removal*, *stemming* memberikan peningkatan akurasi sebanyak 3.55% dan f1-score sebanyak 2.17% sedangkan lemmatization hanya memberikan peningkatan akurasi sebanyak 1.01% dan f1-score sebanyak 0.61%. Hal yang sama juga terjadi pada data dengan stopword removal. Pada data dengan *stopword removal*, stemming memberikan peningkatan akurasi sebanyak 2.03% dan f1-score sebanyak 1.31% sedangkan lemmatization hanya memberikan peningkatan akurasi sebanyak 0.51% dan f1-score sebanyak 0.24%. Akan tetapi, peningkatan nilai metrik evaluasi pada hasil dari *stemming* dan *lemmatization* pada data tidak dirasakan oleh recall. Recall mengalami penurunan sebanyak 0.94% akibat dari kedua jenis reduksi kata tersebut.

Berikutnya analisis terhadap kata kata yang terdapat pada data sebelum dan sesudah proses *stemming* dan *lemmatization* dilakukan untuk melihat kontribusi kedua jenis reduksi kata tersebut terhadap setiap kata dalam data.

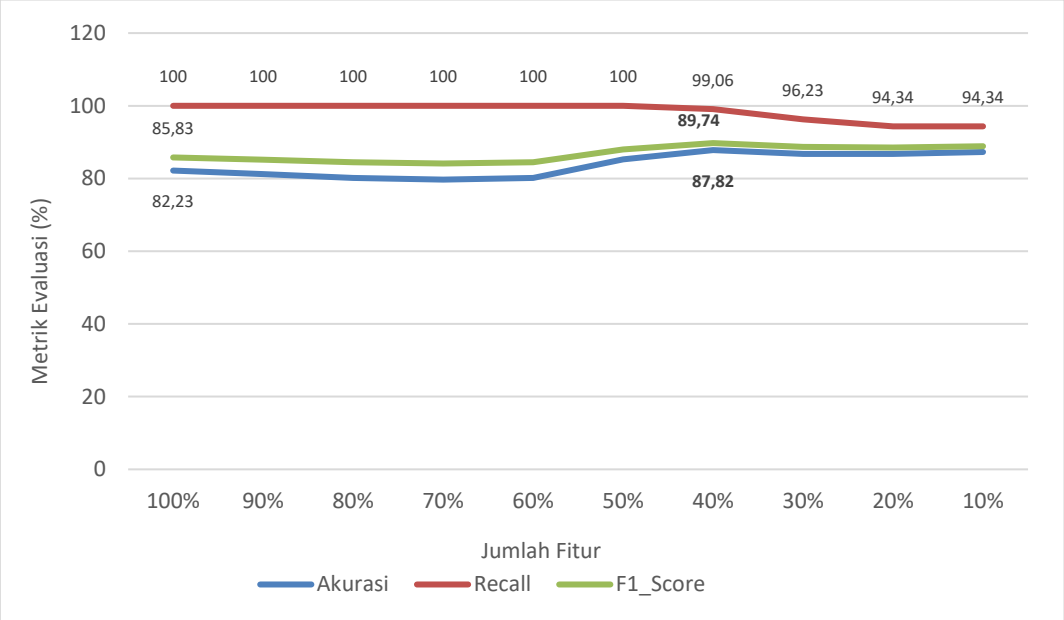
Tabel 8. Analisis kata yang terbuang dan muncul setelah *stemming* dan *lemmatization*

Data	Fitur yang terbuang setelah *Stemming*	Fitur baru yang muncul setelah *Stemming*	Fitur yang terbuang setelah *Lemmatization*	Fitur baru yang muncul setelah *Lemmatization*
Tanpa *Stopword removal*	2018 (-57.72%)	1173 (+33.55%)	443 (-12.67%)	174 (+4.97%)
Dengan *Stopword removal*	1996 (-59.12%)	1169 (+34.63%)	438 (-12.97%)	176 (+5.21%)

Setelah kedua jenis dataset yang telah melewati proses *stemming* dan *lemmatization* diamati, dapat terlihat bahwa proses *stemming* pada data tanpa *stopword removal* mengurangi hampir sebanyak 58% kata dari total kata tetapi menambah kata baru hingga 34% dari total kata. Sedangkan pada *lemmatization*, kata yang dikurangi sebanyak hampir 12% dari total kata dan menambah kata baru sebanyak hampir 5% dari total kata. Selisih reduksi kata oleh *stemming* masih jauh lebih besar sebanyak 24.17% dibandingkan *lemmatization* yang hanya sebesar 7.7%. Berdasarkan hasil analisis ini, dapat disimpulkan bahwa *stemming* memberikan reduksi kata yang cukup signifikan dibandingkan *lemmatization* walaupun terbentuk kata baru. Oleh karena itu performa model yang dihasilkan lebih baik pula.

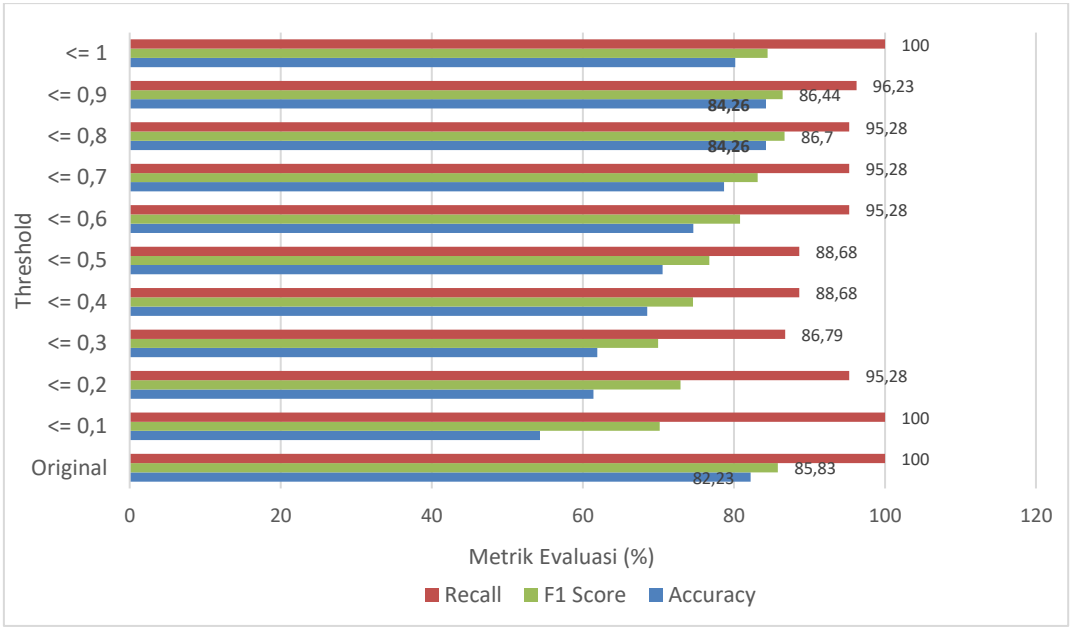
Skenario Kedua – Perbandingan metode seleksi fitur *Information Gain* (IG) dan *Categorical Proportional Difference* (CPD)

Skenario Kedua, yaitu perbandingan metode seleksi fitur *Information Gain* (IG) dan *Categorical Proportional Difference* (CPD). Percobaan ini dilakukan pada dataset dengan *stopword removal*. Pada IG, jumlah fitur terbanyak ke-n dengan nilai IG tertinggi digunakan sebagai parameter dengan variasi nilai n merupakan 10%, 20%, 30% … 100%. Sedangkan pada *feature selection* CPD, parameter yang digunakan adalah variasi nilai *threshold* 0.1, 0.2, 0.3 …1 tanpa menyertakan fitur yang memiliki nilai CPD=0.



Grafik 5. Performa model dengan *Information Gain*

Berdasarkan grafik pengujian seleksi fitur dengan IG dapat dilihat bahwa IG memberikan pengaruh berupa peningkatan akurasi dan f1-score dengan tren yang cenderung meningkat seiring berkurangnya jumlah fitur yang digunakan. Peningkatan akurasi maksimal yang didapatkan sebanyak 5.59% dan f1-score sebanyak 3.91% ketika jumlah fitur yang digunakan sebanyak 40% dari total keseluruhan fitur. Sedangkan pada recall, jumlah fitur 40% yang digunakan oleh model menyebabkan penurunan recall sebanyak 0.94%. Walaupun begitu jumlah fitur sebanyak 40% dari fitur total dianggap paling memberikan performa model yang optimal karena tidak mengorbankan nilai recall yang signifikan.

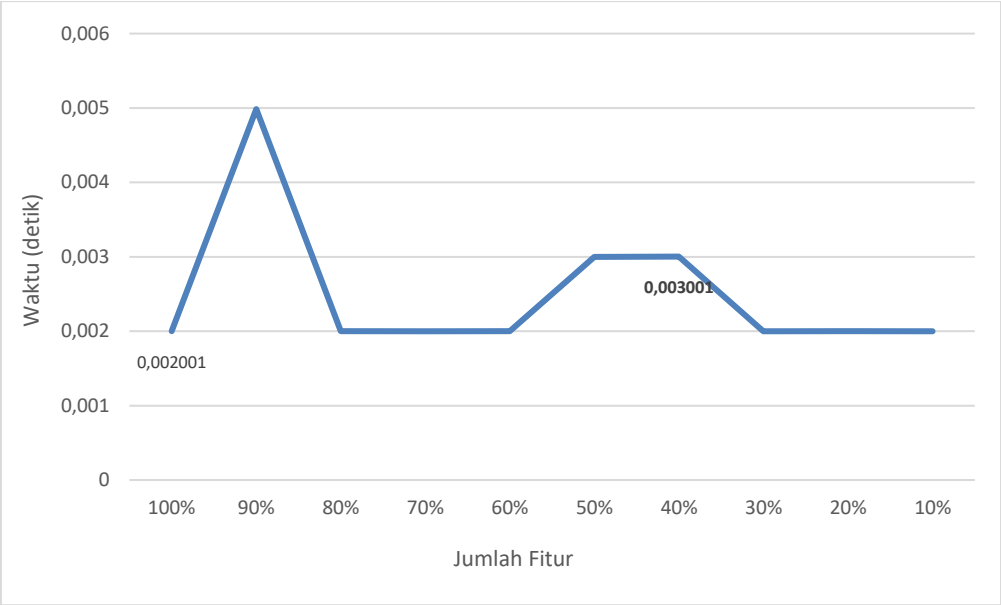


Grafik 6. Performa model dengan *Categorical Proportional Difference*

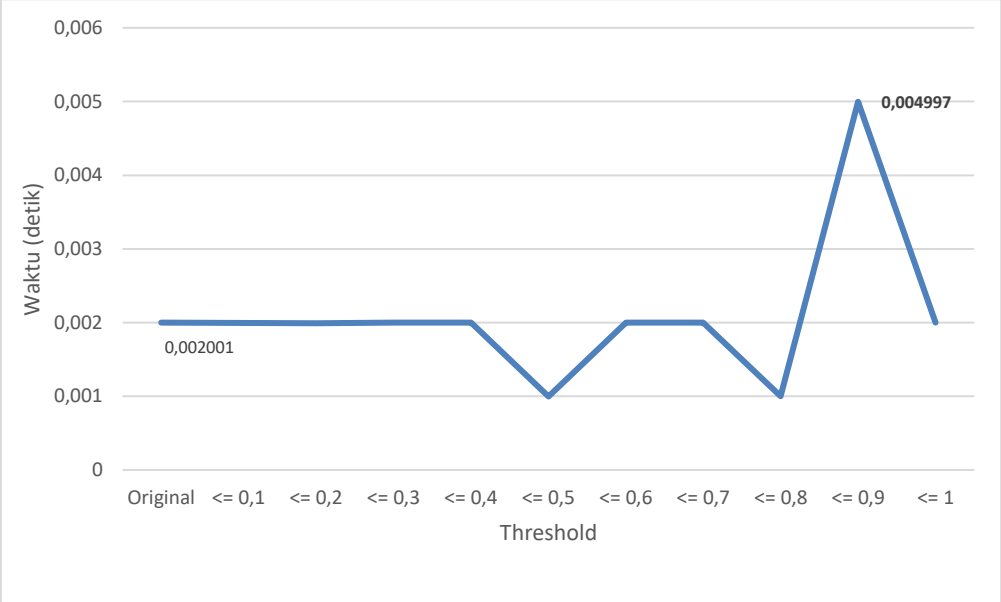
Berdasarkan grafik pengujian seleksi fitur dengan CPD, peningkatan akurasi terbaik didapatkan sebanyak 2.03% ketika nilai CPD <=0.9 dan CPD <= 0.8 dan peningkatan f1-score terbaik sebanyak 0.87% pada CPD<=0.8. Berdasarkan kedua metric evaluasi ini dapat disimpulkan bahwa peningkatan performa model terbaik dicapai ketika fitur dengan nilai CPD<=0.8 digunakan tetapi recall pada titik ini mengalami penurunan yang lebih besar sebanyak 4.72% dibandingkan ketika model menggunakan subset fitur dengan nilai CPD<=0.9, yaitu penurunan sebanyak 3.77%. Karena salah satu tujuan yang ingin didapatkan adalah performa model klasifikasi terbaik dengan meminimalkan jumlah *False Negative* (FN) maka titik CPD<=0.9 dianggap lebih optimum karena menghasilkan akurasi terbaik, penurunan recall lebih kecil dan selisih f1-score yang sangat sedikit, yaitu hanya sebesar 0.26% lebih kecil dibandingkan ketika CPD<=0.8.

Berdasarkan hasil perbandingan akurasi model kedua *feature selection*, dapat disimpulkan bahwa seleksi subset fitur dengan IG mendapatkan performa model terbaik karena peningkatan akurasi maksimum sebanyak 5.59% dibandingkan CPD yang hanya sebesar 2.03%. Jika dilihat dari nilai f1-score, IG juga unggul karena terdapat peningkatan maksimum sebanyak 3.91% sedangkan CPD hanya sebanyak 0.61%. Selain kedua metrik evaluasi tersebut, penurunan nilai recall yang pada model dengan *feature selection* IG juga lebih sedikit, yaitu penurunan recall sebanyak 0.94% dibandingkan CPD yang terjadi penurunan nilai recall sebanyak 3.77%. Penggunaan parameter pemilihan jumlah fitur terbaik pada IG juga dianggap lebih efektif karena subset fitur terbaik dapat ditentukan berdasarkan nilai IG terbesar. Sedangkan pada CPD, fitur dengan nilai CPD mendekati 1 yang menunjukkan dominansi suatu fitur pada kelas tertentu belum tentu memberikan performa model terbaik karena fitur tersebut belum tentu merepresentasikan keseluruhan dataset.

Selain membandingkan performa model berdasarkan akurasi, f1-score dan recall pada kedua *feature selection*, performa komputasi model dengan algoritma *Multinomial Naïve Bayes* juga dibandingkan berdasarkan parameter yang digunakan tiap *feature selection*. Algoritma model yang digunakan diambil dari *library* python yang bernama sklearn. Lalu, program dijalankan pada komputer dengan sistem operasi windows 10 64 bit dengan empat *core* CPU dengan clock speed sebesar 2.2GHz. Model dijalankan berdasarkan tiap feature selection yang digunakan dan dihitung waktunya sehingga didapatkan hasil sebagai berikut



Grafik 7. Waktu komputasi model dengan *feature selection* IG



Grafik 8. Waktu komputasi model dengan *feature selection* CPD

Berdasarkan grafik 7 dapat dilihat bahwa waktu komputasi model fluktuatif seiring dengan berkurangnya jumlah fitur yang digunakan. Grafik 8 juga menunjukkan waktu komputasi yang fluktuatif seiring meningkatnya batas threshold. Pada *feature selection* IG waktu komputasi dalam memproses klasifikasi model dengan jumlah fitur 40% menggunakan IG didapatkan sebesar 0.003001 detik. Sedangkan pada *feature selection* CPD, waktu komputasi ketika threshold CPD<=0.9 adalah 0.004997 detik. Model mengklasifikasikan data lebih lama 0.001996 detik ketika subset fitur yang digunakan adalah fitur-fitur dengan nilai CPD<=0.9 dibandingkan ketika menggunakan 40% fitur dengan nilai IG terbaik. Hal ini menandakan komputasi model dalam mengklasifikasikan data lebih cepat ketika menggunakan subset fitur hasil *feature selection* IG dibandingkan CPD. Selain waktu komputasi model dalam mengklasifikasikan data dicari, waktu komputasi dalam pencarian nilai IG dan CPD pada masing masing fitur juga dicari. Hasil pencarian waktu komputasi pada setiap fitur dirangkum dalam tabel statistik deskriptif sebagai berikut

Tabel 9. Statistik deskriptif waktu komputasi pencarian nilai *feature selection*

Statistik	IG	CPD
Rata-rata	2,395266	0,000004
Minimum	0.883066	0,000000
25%	0,991460	0,000000
50%	1,251403	0,000000
75%	2,132035	0,000000
Maximum	65,263348	0,001001

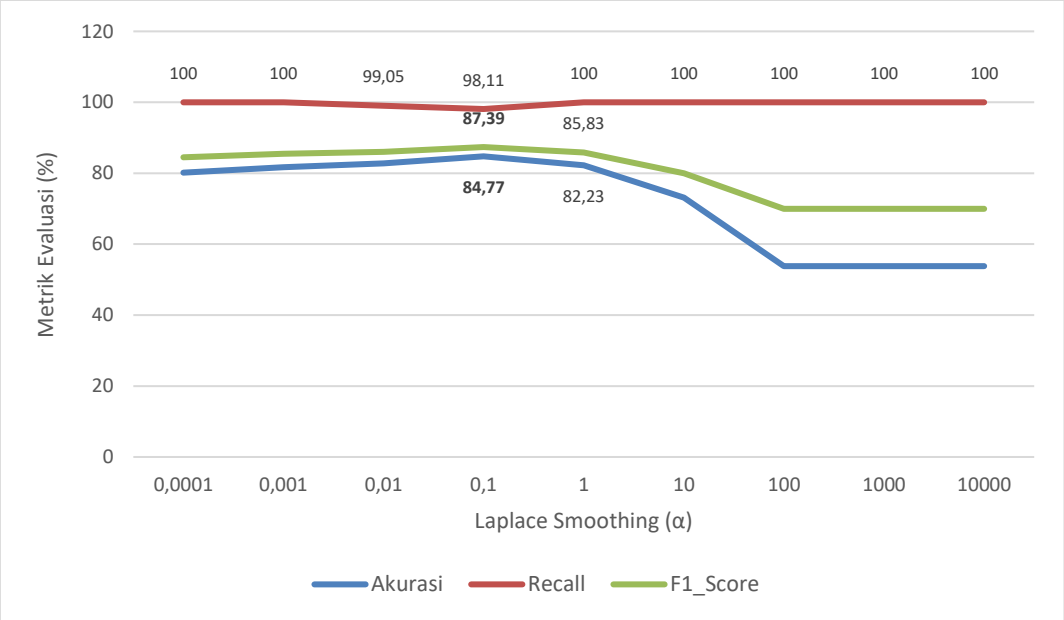
Berdasarkan tabel 9 yang menunjukkan statistic waktu komputasi setiap fitur pada kedua *feature selection*, didapatkan bahwa pencarian nilai IG pada setiap fitur jauh lebih lama karena memerlukan rata rata waktu 2.395266 detik, sedangkan pencarian nilai CPD pada setiap fitur hanya memakan rata rata waktu 0.000004 detik. Selain itu, total waktu yang diperlukan dalam mencari nilai IG untuk seluruh fitur adalah 8086.2 detik atau 134.77 menit sedangkan pada CPD total waktu yang diperlukan hanya 0,011993 detik. Sehingga dalam kemampuan komputasi pencarian nilai *feature selection* setiap fitur, CPD lebih unggul karena waktu yang diperlukan hampir mendekati 0 detik.

Skenario Ketiga – Peningkatan performa model dengan pengaturan parameter *laplace smoothing*.

Sebelum melakukan upaya peningkatan performa model dengan tuning parameter *smoothing*, kata kata yang terdapat pada kedua data testing diambil dan dilihat mana dari kata kata tersebut yang tidak termasuk *vocabulary*. Kata kata ini disebut *Out-of-Vocabulary* (OOV). Setelah diperiksa terdapat 815 OOV atau 38% dari keseluruhan kata pada data testing. Kata kata OOV ini akan menjadi acuan terhadap data hasil klasifikasi model yang telah mengalami peningkatan performa jika peningkatan tersebut terjadi.

Setelah mendapatkan informasi OOV dari setiap dataset, penulis melakukan skenario ketiga, yaitu pencarian parameter *laplace smoothing* atau bisa juga disebut dengan nilai α dalam pengaplikasiannya pada multinomial naïve bayes sklearn. Parameter *smoothing* ini adalah teknik yang berguna dalam mengatasi perhitungan sebuah probabilitas agar tidak menghasilkan nilai 0. Nilai α pada *likelihood probability* adalah angka positif (α > 0) dan nilai α yang umum digunakan adalah 1.

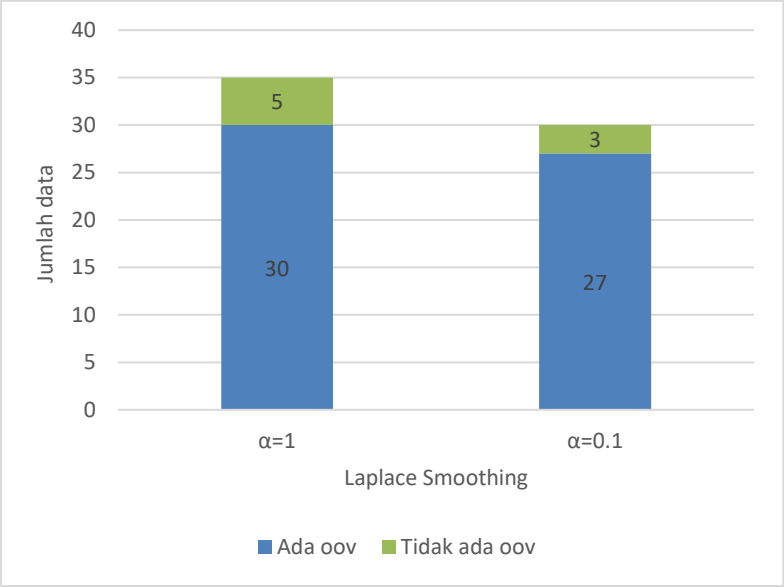
Pada percobaan ini pengaturan parameter nilai α merupakan masukan dari sembilan bilangan real, yaitu 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, dan 10000. Hasil dari percobaan performa model dengan beberapa nilai α tersebut didapatkan grafik berikut ini.



Grafik 9. Pengaruh *laplace smoothing* (α) terhadap performa model

Berdasarkan grafik 9, dapat dilihat bahwa peningkatan akurasi maksimum dicapai ketika nilai α = 0.1 yaitu sebanyak 2.54% dan f1-score sebanyak 1.56%. Walaupun terjadi peningkatan di kedua metrik tersebut, recall ketika α = 0.1 terjadi penurunan sebanyak 1.89%. Penurunan recall yang terjadi tidak signifikan karena masih memberikan keseluruhan nilai recall yang besar.

Untuk melihat bukti bahwa *smoothing* berhasil mengatasi data dengan OOV, penulis mengambil data data yang terklasifikasi salah ketika model menggunakan α = 1 dan ketika model menggunakan α =0.1.



Grafik 10. Jumlah data yang terklasifikasi salah pada kedua alpha

Berdasarkan informasi grafik 10 dapat diamati bahwa peningkatan performa model berhasil mengurangi data yang terklasifikasi salah dengan total sebanyak lima data. Adapun komposisi data yang mengandung fitur OOV berkurang sebanyak 10% dan data yang tidak terdapat OOV berkurang sebanyak 40%. Hal ini membuktikan bahwa *smoothing* parameter memberikan peningkatan performa model yang cukup baik dan berhasil mengurangi data yang mengandung fitur OOV walaupun tidak signifikan.

Penggunaan nilai α = 0.1 yang menghasilkan peningkatan performa ini sesuai dengan aplikasi *likelihood probability* pada formula 9 dimana α yang menjadi pembagi bernilai kecil dapat menyebabkan nilai *likelihood probability* nya besar. Sehingga hubungan antara *laplace smoothing* dan *likelihood probability* ini berbanding terbalik.

5. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa kesimpulan. Kesimpulan terbagi menjadi kesimpulan dalam pola penggunaan bahasa oleh perilaku depresi dan beberapa skenario penelitian yang bertujuan untuk membangun model klasifikasi dengan performa terbaik.

Dalam menentukan pola penggunaan bahasa pada data yang terklasifikasi depresi, penulis melakukan eksplorasi penggunaan kata ganti orang pada data depresi. Hasil yang didapatkan adalah, pada data depresi, penggunaan kata ganti orang pertama, yaitu kata ganti 'I' lebih dominan sebanyak 83.14% dibandingkan kata ganti lain. Selain itu, perilaku depresi juga ditandai dengan banyak penggunaan to-be yang termasuk dalam *present tense* atau grammar yang menceritakan kejadian yang sedang terjadi atau biasanya terjadi. Terakhir, perilaku depresi juga ditandai dengan adanya negasi terhadap sesuatu atau konteks yang sedang diceritakan.

Pada skenario pertama yang merupakan perbandingan jenis *preprocessing* untuk menghasilkan performa model terbaik didapatkan bahwa *stopword removal* memberikan pengaruh terhadap peningkatan performa model. Peningkatan akurasi yang didapatkan sebanyak 7-9% dan peningkatan f1-score sebesar 5-6% pada ketiga jenis data. Dalam penggunaan teknik reduksi kata, *stemming* terbukti memberikan performa model terbaik dengan peningkatan akurasi sebanyak 2-3.5% dan f1-score sebanyak 1-2% pada kedua jenis data. Sedangkan lemmatization hanya memberikan peningkatan akurasi sebanyak 0.5-1% dan f1-score sebanyak 0.2-0.6% pada kedua jenis data. Selain itu *stemming* juga mengurangi kata lebih banyak dibandingkan *lemmatization*, yaitu 24.17% lebih banyak daripada *lemmatization* yang hanya sebesar 7.7%.

Pada skenario kedua, *feature selection* yang menghasilkan performa model terbaik adalah *Information Gain* (IG) karena memberikan peningkatan akurasi maksimum sebanyak 5.59% menjadi 87.82% dibandingkan *Categorical Proportional Difference* (CPD) yang hanya memberikan peningkatan sebanyak 2.03% menjadi 84.26%. Selain itu peningkatan f1-score maksimum juga dirasakan oleh IG sebanyak 3.91% menjadi 89.74% dibandingkan CPD yang hanya memiliki peningkatan 0.61% menjadi 86.44%. Nilai recall yang dihasilkan oleh IG juga mengalami penurunan paling kecil, yaitu 0.94% dibandingkan CPD sebanyak 3.77%.

Pada IG, selain akurasi, f1-score dan recall yang dihasilkan lebih baik, penggunaan parameternya juga lebih efektif karena mengambil fitur dengan nilai IG terbesar sebanyak variasi jumlah fitur yang telah ditentukan, sedangkan pada CPD parameter yang digunakan adalah batas nilai *threshold* yang berdasarkan nilai CPD tertentu. Hal ini karena fitur dengan nilai CPD yang besar belum tentu memberikan performa model yang terbaik.

Walaupun IG memberikan performa model klasifikasi terbaik, perhitungan nilai IG pada setiap fitur memakan waktu dua kali lebih lama daripada CPD, yaitu 2.395266 detik dibandingkan CPD yang hanya 0.000004 detik. Akan tetapi, waktu komputasi model dalam mengklasifikasikan data berdasarkan subset fitur dari kedua *feature selection* tidak jauh berbeda karena hanya terjadi selisih waktu 0.001996 detik dengan keunggulan pada subset fitur IG.

Pada skenario ketiga yang melakukan pencarian parameter *laplace smoothing* atau nilai α terbaik yang berguna dalam menghindari sebuah probabilitas menghasilkan nilai 0 dan nilai α adalah bilangan positif. Hasil penelitian ini didapatkan performa model terbaik ketika model menggunakan nilai $\alpha = 0.1$. Penggunaan nilai alpha ini menghasilkan peningkatan akurasi maksimum sebanyak 2.54% menjadi 84.77% dan peningkatan f1-score sebanyak 1.56% menjadi 87.39%. Walaupun nilai recall yang didapatkan pada alpha ini mengalami penurunan sebesar 1.89% tetapi penurunan yang terjadi tidak signifikan karena data False Negative (FN) masih berjumlah sedikit. Selain itu, peningkatan performa model ini juga ditandai dengan kemampuan model mengatasi pengurangan data error yang terdapat kata out-of-vocabulary atau OOV. Terbukti peningkatan performa model mengurangi data error yang mengandung OOV sebanyak 10%.

Daftar Pustaka

[1] "Depression," 22 March 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] Burdisso, S. G., Errecalde, M., & Montes-y-Gómez, M.. 2019) A Text Classification Framework for Simple and Effective Early Depression Detection Over Social Media Streams. Expert Systems with Applications, 133, 182-197.

[3] Murray, C. J., & Lopez, A. D.. 1997. Alternative projections of mortality and disability by cause 1990–2020. Global Burden of Disease Study. The Lancet, 349(9064), 1498-1504.

[4] Detels, R.. 2009. The scope and concerns of public health. London. Oxford University Press.

[5] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E.. 2013. Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media.

[6] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z.. 2013. A depression detection model based on sentiment analysis in micro-blog social network. Berlin: In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 201-213).

[7] De Choudhury, M., Counts, S., & Horvitz, E.. 2013. Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47-56).

[8] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M.. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 31-39).

[9] Rude, S., Gortner, E. M., & Pennebaker, J.. 2004. Language use of depressed and depression-vulnerable college students. Cognition & Emotion, 18(8), 1121-1133.

[10] Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848.

[11] O'Keefe, T., & Koprinska, I.. 2009. Feature selection and weighting methods in sentiment analysis. In Proceedings of the 14th Australasian document computing symposium, Sydney (pp. 67-74).

[12] Pratiwi, A. I. (2018). On the feature selection and classification based on information gain for document sentiment analysis. Applied Computational Intelligence and Soft Computing, 2018

[13] Pang, B., Lee, L., & Vaithyanathan, S.. 2002. Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

[14] Simeon, M., & Hilderman, R. (2008, November). Categorical proportional difference: A feature selection method for text categorization. In Proceedings of the 7th Australasian Data Mining Conference-Volume 87 (pp. 201-208).