Identifikasi Keberpihakan *Tweet* pada *Twitter* Menggunakan *Naive Bayes Classifier* Berdasarkan Klasifikasi Emosi Menggunakan *Class Sequential Rules* (Studi Kasus: Pemilihan Presiden 2019)

Rizky Wahyu Kurniawati[1], Anisa Herdiani[2], Indra Lukmana Sardi[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]rwkiky@students.telkomuniversity.ac.id, [2]anisaherdiani@telkomuniversity.ac.id,
[3]indraluk@telkomuniversity.ac.id

Abstrak

Penggunaan media sosial untuk analisis politik menjadi hal yang umum terjadi, terutama selama pemilihan presiden (pilpres). Banyak peneliti dan media mencoba menggunakan media sosial untuk memahami opini dan tren publik. *Twitter* merupakan media sosial yang digunakan sebagai tempat banyak masyarakat di *internet* memberikan opininya termasuk terkait pilpres. Beragam jenis emosi ditunjukkan oleh mereka melalui *tweet*nya dan suatu jenis emosi tertentu dapat menentukan kecenderungan keberpihakan seseorang terhadap suatu paslon. Klasifikasi emosi pada *tweet* diperlukan untuk mengetahui seberapa banyak masyarakat yang berpihak pada suatu paslon. Dalam satu *tweet* dapat terdiri lebih dari satu kalimat dan banyak kata. Susunan kata juga dapat mempengaruhi hasil emosi yang disimpulkan. Pada penelitian ini metode *Class Sequential Rules* (CSR) digunakan karena kemampuannya dalam pendekatan berbasis pola bahasa didukung dengan pendekatan berbasis leksikon. Selain itu, juga menggunakan *Naive Bayes Classifier* (NBC) untuk mengidentifikasikan keberpihakan *tweet* terhadap suatu pasangan calon. Dengan metode tersebut, hasil yang didapatkan dari sistem yaitu keberpihakan kepada Jokowi sebesar 67.5% sedangkan Prabowo sebesar 35.5% serta didapatkan *F1-Score* sebesar 67.83%.

Kata kunci : klasifikasi, emosi, prediksi, pilpres, CSR, *twitter*.

Abstract

The use of social media for political analysis is common, especially during presidential elections. Many researchers and media try to use social media to understand public opinion and trends. Twitter is a social media that is used as a place for many people on the internet to give their opinions, including those related to the presidential election. Various types of emotions are shown by them through their tweets and a certain type of emotion can determine a person's tendency to align with a paslon. Emotional classification on tweets is needed to find out how many people are in favor of a paslon. In one tweet can consist of more than one sentence and many words. Word order can also affect the outcome of emotions that are inferred. In this study the Class Sequential Rules (CSR) method is used because its ability in a language pattern-based approach is supported by a lexicon-based approach. In addition, it also uses Naive Bayes Classifier to identify tweet alignments towards a candidate pair. With this method, the results obtained from the system are alignments to Jokowi by 67.5% while Prabowo by 35.5% and F1-Score of 67.83%.

Keywords: classification, emotion, prediction, election, CSR, twitter

1. Pendahuluan

1.1 Latar Belakang

Pemilihan Presiden 2019 (Pilpres 2019) dilaksanakan pada 17 April 2019. Banyak pihak yang ramai membicarakannya baik melalui berita, media sosial atau secara langsung. Penggunaan media sosial untuk analisis politik menjadi hal yang umum terjadi untuk memahami opini dan tren publik [1]. Disamping itu, *twitter* merupakan salah satu media sosial yang dapat dijadikan sebagai tempat masyarakat di internet (*netizen*) bersuara memberikan opini terkait keberpihakan terhadap suatu pasangan calon (paslon). Beragam jenis emosi ditunjukkan oleh *netizen* melalui *tweet*-nya. Suatu jenis emosi dapat menentukan arah keberpihakan seorang terhadap paslon tertentu. Seperti yang telah dikatakan oleh Direktur Eksekutif lembaga survei Median, Rico, mengatakan jika seseorang yang teraduk-aduk emosinya akan menimbulkan suatu keberpihakan terhadap suatu paslon [2].

Pada penelitian ini yang diselidiki adalah bagaimana mengidentifikasi keberpihakan *tweet* di *twitter* pada pilpres 2019 berdasarkan klasifikasi emosi. Salah satu hal yang dapat mempengaruhi keberpihakan kepada salah satu paslon adalah opini yang diutarakan melalui *tweet*. Oleh karena itu dibutuhkan suatu cara atau metode untuk mengidentifikasi keberpihakan *tweet* terhadap paslon pada dunia maya melalui *twitter* yang memiliki tingkat performansi yang baik.

Pada penelitian ini, diimplementasikan pengklasifikasian emosi terhadap *tweet* pada *Twitter* menggunakan *Class Sequential Rules* (CSR) dengan pendekatan *Lexicon-Based* untuk menemukan emosi yang terkandung dalam

tweet secara terperinci. Sedangkan pengidentifikasian keberpihakan terhadap suatu pasangan calon menggunakan *Naive Bayes Classification.*

Data diambil dari *tweet* yang berhubungan dengan Pilpres 2019 pada *twitter* karena menurut Kompas [3] menyatakan bahwa pada Februari 2017 lalu, Indonesia masuk ke dalam lima pengguna aktif tertinggi di dunia dengan sekitar 29 juta pengguna sehingga memudahkan dalam pengambilan data. Dari data *tweet* tersebut, diklasifikasikan ke dalam lima parameter emosi dasar manusia antara lain senang, sedih, marah, takut dan benci [4]. Kelima parameter emosi ini berguna untuk mengetahui kecenderungan keberpihakan seseorang terhadap paslon tertentu. Proses pengklasifikasiannya menggunakan metode CSR, karena untuk penggalian fitur yang terdapat pada *tweet* yang biasanya terdiri dari frasa pendek atau kalimat tidak lengkap, membutuhkan suatu pendekatan yang berbasis pola bahasa, dan metode ini dapat menghasilkan pola bahasa [5]. Metode CSR didukung dengan pendekatan *Lexicon-Based* untuk menentukan kelas emosi yang digunakan serta menentukan kata-kata yang mengandung emosi kemudian diberikan label sesuai kelas yang telah ditetapkan terlebih dahulu. Selain itu, dibutuhkan metode yang lain yaitu *Naive Bayes Classification* untuk pengidentifikasian keberpihakan suatu *tweet* dengan menghitung probabilitasnya. Analisa hasil performansi klasifikasi juga dilakukan dengan pengujian perhitungan *F1-Score*.

1.2 Topik dan Batasan

Berdasarkan latar belakang yang sudah dijelaskan sebelumnya topik yang dibahas pada penelitian ini adalah sebagai berikut.

a. Bagaimana mengidentifikasi keberpihakan *tweet* berdasarkan klasifikasi emosi dalam Pilpres 2019?
b. Bagaimana mengukur tingkat performansi sistem dari hasil identifikasi keberpihakan *tweet* berdasarkan klasifikasi emosi dalam pemilihan presiden 2019?

Adapun beberapa batasan masalah terhadap sistem yang dibangun dalam penelitian ini adalah sebagai berikut.

a. Dataset yang digunakan berasal dari *tweet* pada media sosial *Twitter*.
b. Topik yang diambil dari dataset yaitu *tweet* seputar Pemilihan Presiden 2019 (Pilpres 2019).
c. Dataset yang diambil adalah *tweet* yang terdapat ketika masa kampanye Pilpres 2019 saja.
d. Dataset yang diambil hanya yang berbahasa Indonesia saja.
e. Jenis emosi yang digunakan untuk klasifikasi *tweet* di sosial media *Twitter* antara lain yaitu senang, sedih, marah, benci dan takut.
f. Jenis dataset yang diambil dari *tweet* yang berupa teks bukan berupa *web link*, gambar atau video.
g. Tidak dapat menangani *tweet* yang mengandung kalimat sindiran atau sarkasme.

1.3 Tujuan

Berdasarkan topik yang sudah dijelaskan sebelumnya, tujuan yang dicapai pada penelitian ini adalah sebagai berikut.

a. Mengindentifikasi keberpihakan *tweet* menggunakan *Naive Bayes Classification* berdasarkan klasifikasi emosi dengan menggunakan *Class Sequential Rules* dalam Pilpres 2019.
b. Mengukur tingkat performansi sistem dalam mengidentifikasi keberpihakan *tweet* berdasarkan klasifikasi emosi dalam pemilihan presiden 2019 dengan menggunakan *F1-Score.*

1.4 Sistematika Penulisan

Jurnal penelitian ini terdiri dari beberapa sub-bagian yang setiap bagian memiliki peranan untuk menjelaskan sesuatu. Bagian pertama dimulai dari bagian latar belakang, identifikasi masalah, tujuan serta sistematika penulisan. Bagian kedua terdapat studi terkait. Bagian ketiga terdapat metodologi yang menjelaskan rancangan dan sistem atau produk yang dihasilkan. Bagian keempat terdapat evaluasi berisi hasil dari pengujian dan analisis hasil pengujian dari penelitian. Serta, pada bagian terakhir terdapat kesimpulan yang diambil dari hasil pengujian dan analisis hasil pengujian sehingga tidak ada kesimpulan dari teori ataupun nalar semata.

2. **Studi Terkait**.

2.1 Klasifikasi Emosi

Pengklasifikasian *tweet* ke dalam beberapa parameter emosi digunakan untuk mengenali opini publik pada *tweet*. Parameter emosi yang digunakan yaitu senang, sedih, marah, takut dan benci. Kelima parameter emosi tersebut dipilih karena merupakan parameter emosi dasar menurut penelitian [4]. *Tweet* diklasifikasikan ke dalam beberapa parameter emosi karena dari setiap opini memiliki emosi yang ingin tersampaikan oleh penulisnya. Sehingga hal tersebut dapat dianalisis untuk pengklasifikasian agar lebih jelas maksud dari opini yang terdapat pada *tweet* tersebut mengenai Pilpres 2019 ini. Dari hasil pengklasifikasian emosi tersebut, selanjutnya ditarik kesimpulan seberapa besar keberpihakan *tweet* pada *Twitter* terhadap suatu paslon.

2.2 *Naive Bayes Classifier*

Naive Bayes Classifier merupakan metode *supervised learning* yang digunakan untuk pengklasifikasian teks berdasarkan teorema Bayes dengan asumsi "naif" tentang independensi bersyarat antara setiap pasangan fitur yang diberi nilai variabel kelas [6]. Pada rumusnya, kelas dokumen tidak ditentukan hanya berdasarkan kata yang muncul, namun berdasarkan jumlah kemunculannya juga [7]. Adapun Probabilitas dokumen *d* yang terletak di kelas c memiliki perhitungan [8].

$$P(c|d) \propto P(c) \prod_{1 \le k \le n_d} P(t_k|c) \quad (2.1)$$

Dimana $P(t_k|c)$ adalah probabilitas kondisional dari t_k yang berada pada dokumen yang dimiliki kelas *c*. Dari persamaan diatas dapat diketahui bahwasannya $P(t_k|c)$ merupakan *likehood probability* dari t_k yang terdapat pada kelas *c*, di sisi lain $P(c)$ merupakan *prior probability* dokumen yang berada pada kelas *c.* Hasil dari *posterior probability* akan dibandingkan untuk menentukan kelas, kelas yang memiliki nilai *posterior probability* terbesar merupakan kelas yang dipilih sebagai hasil prediksi [8]. Berikut formula dari *Prior probability*.

$$P(c) = \frac{N_c}{N} \quad (2.2)$$

N_c merupakan jumlah dari kategori dari *c*. *N* jumlah kategori keseluruhan. Untuk *likehood probability* akan dihitung kata atau fitur t_k pada seluruh dokumen latih pada c, dengan menggunakan *LaPlace Smoothing* memiliki Rumus [9].

$$P(t_k|c) = \frac{N_k + 1}{|V| + N'} \quad (2.3)$$

Dimana N_k adalah jumlah kemunculan t_k dalam dokumen latih pada suatu kelas c dan *N* adalah jumlah total kata atau fitur yang terdapat pada c dokumen latih. Penambahan angka 1 pada formula 2.3 berfungsi sebagai *LaPlace Smoothing* yang dilakukan agar terhindar dari zero probability pada ekstraksi fitur, sehingga hasil akhir yang diperoleh tidak bernilai nol [8].

2.3 Pendekatan Berbasis Leksikon

Pendekatan berbasis leksikon ini sangat bergantung pada leksikon emosi [10]. Pada penelitian ini, leksikon emosi dibangun berdasarkan tiga sumber. Pertama, leksikon emosi diambil dari jurnal Johnson-Laird dan Keith Oatley [11] yang terdiri dari lima emosi dasar yaitu senang (*happy*), sedih (*sad*), takut (*fear*), marah (*anger*) dan benci (*disgust*). Kedua, mengumpulkan dan menggunakan beberapa *slang words* dari dokumen pada *GitHub* [12] yang dapat membantu dalam pengklasifikasian emosi. Ketiga, mengumpulkan daftar kata yang mengandung emosi yang sudah terdapat *corpus*-nya pada jurnal [11]. Namun ada beberapa kata yang mengandung emosi yang tidak berhubungan dengan penelitian sehingga tidak digunakan.

Berdasarkan leksikon yang dibangun, selanjutnya mengenali pola susunan kata untuk memprediksi label emosi yang sesuai pada setiap *tweet*.

2.4 *Class Sequential Rules*

Metode CSR digunakan untuk membuat aturan dasar pada suatu pola berurut dari kalimat. Suatu *tweet* akan diklasifikasikan berdasarkan banyaknya kata yang mengandung emosi dalam suatu kalimat dan urutan peletakan kata. Beberapa pemetaan umum yang biasa digunakan dalam CSR menurut jurnal Minqing Hu dan Bing Liu [5], adalah $I = \{i_1, i_2, \ldots, i_n\}$ adalah jenis kata apa aja yang digunakan dalam penelitian. Pada penelitian ini berisi label emosi, label kandidat, dan jenis kata. Selanjutnya terdapat *itemset* atau *subsequence* yang dipetakan dengan $<a_1, \ldots, a_i, \ldots, a_r>$ untuk memberikan identitas urutan setiap kata pada kalimat. Kemudian terdapat *sequence* yang dipetakan dengan $s = a_1 a_2 \ldots a_r$ yang menjelaskan identitas dari suatu *tweet* yang menampung urutan dari *itemset* dimana $a_i \in s$ dan $s \in I$. Data *instance* atau *D* merupakan himpunan pasangan yang menyatakan suatu *tweet* masuk ke dalam parameter emosi tertentu. Misal $D = \{(s_1, y_1), (s_2, y_2), \ldots, (s_n, y_n)\}$, dimana $y_i \in Y$ adalah label kelas. *Y* merupakan sekumpulan dari semua kelas. *Class Sequential Rules* (CSR) adalah implikasi dari bentuk $X \rightarrow y$, dimana *X* adalah *sequence*, dan $y \in Y$ [10]. Data *instance* (s_i, y_i) dikatakan *satisfy* CSR jika *X* adalah dari s_i (urutan kalimat) dan y_i (kelas emosi prediksi) = *y* (kelas emosi asli). Dikatakan *satisfy* jika antara *sequence* dan labelnya sesuai. Nilai *support* dari sebuah CSR adalah jumlah *sequence* yang mengandung Y di dalam *database* S.

Untuk mendapatkan aturan CSR, digunakan algoritma *Class Prefix-Span* yang merupakan hasil modifikasi algoritma *PrefixSpan.* Berdasarkan penelitian Pei [13], *Prefixspan* memiliki performansi yang tinggi dalam menggali pola sekuensial dibandingkan algoritma lainnya seperti GSP dan *FreeSpan*. *PrefixSpan* tidak membangkitkan semua kandidat, tapi hanya yang memenuhi syarat minimum *threshold* saja.

Pada penelitian ini mengadaptasi metode pertumbuhan pola (*pattern growth*) dari penelitian Pei [13]. Tahapan untuk penerapannya dimulai dari berfokus hanya pada pola yang mengandung kelas. Kemudian temukan pola yang memiliki kelas sebagai *suffix*. Selanjutnya mengambil pola yang dihasilkan sebagai *prefix*. Dari langkah umum tersebut, dapat menemukan semua aturan sekuensial kelas dengan pertumbuhan pola.

2.5 *Confusion Matrix*

Confussion matrix adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan [14]. Terdapat 2 jenis *confusion matrix* yaitu untuk klasifikasi biner dan untuk *multiple classes*. Confusion matrix biner digunakan apabila hanya terdapat 2 kelas, sedangkan *multiple classes* menggunakan lebih dari dua kelas. Berikut merupakan contoh dari *confusion matrix* untuk klasifikasi biner.

Tabel 1 *Confusion Matrix* Kelas *Biner* [14]

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Perhitungan *precision* merupakan perhitungan tingkat kecocokan nilai prediksi yang benar terhadap total jumlah data pada kelas prediksi. Perhitungan *precision* dapat dilihat pada persamaan (2.4) di bawah ini.

$$Precision = \frac{TP}{TP + FP} \quad (2.4)$$

Perhitungan *recall* merupakan perhitungan tingkat kecocokan nilai prediksi yang benar terhadap total jumlah data pada kelas sebenarnya. Perhitungan *recall* dapat dilihat pada persamaan (2.5) di bawah ini.

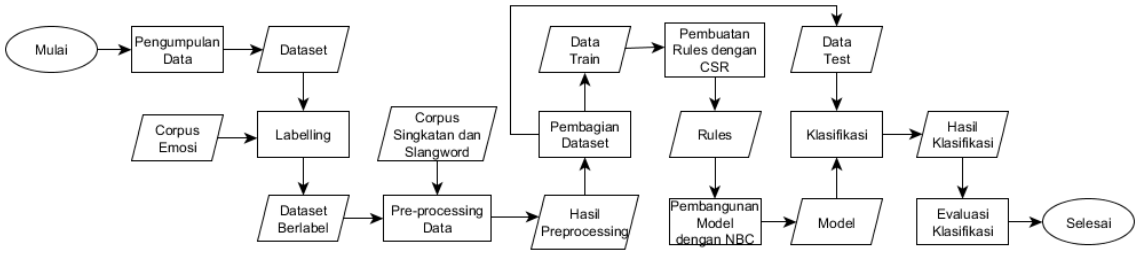
$$Recall = \frac{TP}{TP + FN} \quad (2.5)$$

Namun terkadang perhitungan antara *precision* dan *recall* memiliki perbedaan yang cukup tinggi, untuk itu dilakukan penyetaraan nilai precision dan recall menggunakan *F1-Score* [15]. Perhitungan *F1-Score* dapat dilihat pada persamaan (2.6) dibawah ini.

$$F1 - Score = 2\,x\,\frac{Precision\ x\ Recall}{Precision + recall} \quad (2.6)$$

3. Metodologi

Sistem yang dibangun pada penelitian mengadaptasi dari jurnal [10] untuk mengklasifikasi emosi dengan beberapa modifikasi. Gambar 1 menggambarkan tahapan proses secara garis besar pada penelitian.



Gambar 1 Tahapan Proses dalam Penelitian

Berikut di bawah ini merupakan penjelasan dari Gambar 1 untuk setiap langkahnya.

3.1 Pengumpulan Data

Sesuai dengan judul penelitian ini, data diperoleh dari *tweet* pada *Twitter*. Cara mengumpulkan *tweet*s tersebut yaitu menggunakan suatu *library* pada bahasa pemrograman *Python* yaitu *Tweepy*. *Library* tersebut digunakan untuk pemanggilan, atau mengakses *Twitter* API itu sendiri. *Tweet* yang diambil yaitu yang di buat ketika masa kampanye saja yaitu dari tanggal 23 September 2019 hingga 13 April 2019. Pada pengambilan *tweet, keyword* yang digunakan adalah nama akun dari masing-masing calon yaitu @jokowi, @prabowo, @sandiuno, dan @Kiyai_MarufAmin. *Keyword* tersebut dapat mewakili masyarakat yang membicarakan mengenai pilpres kali ini. Setelah *tweets* yang terkumpul menjadi *dataset*, dilakukan pembersihan dengan melihat *tweet* yang relevan mengenai Pilpres 2019. Data yang diambil sejumlah 500 data. Berikut merupakan contoh dari *dataset* yang telah dikumpulkan.

Tabel 2 Contoh Data Set

Tweet
Pengusaha muda yang tergabung dalam Relawan Pengusaha Muda Nasional untuk @jokowi (Repnas) Sumatera Utara sepenuh hati mendukung dan siap memenangkan Paslon 01
@prabowo yaampun..terharu banget..begitu besar keinginan rakyat menginginkan bapak memimpin negeri ini..

3.2 *Corpus* Emosi

Corpus emosi ini didapatkan dari jurnal [11] yang kemudian diterjemahkan ke dalam bahasa Indonesia. Pada *corpus* tersebut terdapat 445 kata yang kemudian menjadi 367 kata yang mengandung emosi karena terdapat persamaan arti. *Corpus* yang telah diterjemahkan terdiri 148 kata beremosi senang, 83 kata beremosi sedih, 31 kata beremosi takut, 44 kata beremosi marah, dan 61 kata beremosi benci.

3.3 *Labelling*

Pendekatan *Lexicon-Based* dilakukan pada tahap ini. Sebelum terdapat *corpus*, terlebih dahulu menentukan parameter kelas emosi apa saja yang dipilih. Sesuai dengan jurnal mengenai emosi dasar manusia [11], emosi yang terpilih yaitu senang, sedih, takut, marah, dan benci. Pada tahap ini, *corpus* emosi yang telah ada menjadi pedoman dalam pelabelan dataset. Pelabelan atau *labelling* dilakukan oleh peneliti namun terdapat validasi terhadap psikolog dari UNISBA dan ilmuan psikologi UNPAD. Hasil dari proses ini yaitu dataset yang telah berlabel.

3.4 *Pre-procsessing* Data

Dataset yang telah terkumpul, selanjutnya dilakukan *pre-processing* yang di dalamnya terdapat serangkaian proses yang lebih rinci lagi. Selain itu, pada tahap ini juga membutuhkan *corpus* singkatan dan *slangword* yang diambil dari dokumen pada Github [12] yang dimodifikasi dengan menambahkan beberapa kata yang berkaitan dengan pilpres. Berikut tabel yang mendiskripsikan tahapan pada *pre-processing data*.

Tabel 3 Deskripsi Tahapan *Preprocessing*

Tahapan	Proses	Input	Output
Data Cleaning	Proses untuk penghapusan *website link*, simbol, *username*, dan angka. Selain melakukan penghapusan hal yang dianggap tidak penting, terdapat pula proses *casefolding*	Pengusaha muda yg tergabung dlm Relawan Pengusaha Muda Nasional untuk @jokowi (Repnas) Sumatera Utara sepenuh hati mendukung dan siap memenangkan Paslon 01	pengusaha muda yg tergabung dlm relawan pengusaha muda nasional untuk jokowi repnas sumatera utara sepenuh hati mendukung dan siap memenangkan paslon 01
Tokenization	Proses mengubah suatu kalimat, dalam hal ini *tweet* ke dalam bentuk *token* per kata	pengusaha muda yg tergabung dlm relawan pengusaha muda nasional untuk jokowi repnas sumatera utara sepenuh hati mendukung dan siap memenangkan paslon 01	['pengusaha', 'muda', 'yg', 'tergabung', 'dlm', 'relawan', 'pengusaha', 'muda', 'nasional', 'untuk', 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', 'dan', 'siap', 'memenangkan', 'paslon', '01]
Penanganan Singkatan dan *Slang Word*	Menyeragamkan setiap *term* pada *tweet*. Kata-kata yang disingkat memungkinkan memberikan pengaruh terhadap pemrosesan klasifikasi sehingga harus dilakukan konversi ke beantuk aslinya. Selain itu, juga dilakukan konversi *slang word* atau kata yang tidak baku menjadi kalimat baku.	['pengusaha', 'muda', 'yg', 'tergabung', 'dlm', 'relawan', 'pengusaha', 'muda', 'nasional', 'untuk', 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', 'dan', 'siap', 'memenangkan', 'paslon', '01]	['pengusaha', 'muda', 'yang', 'tergabung', 'dalam', 'relawan', 'pengusaha', 'muda', 'nasional', 'untuk', 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', 'dan', 'siap', 'memenangkan', 'pasangan', 'calon', 'jokowi']
Stopword Removal	Penghapusan kata-kata yang bersifat umum namun tidak memiliki makna atau informasi yang dibutuhkan.	['pengusaha', 'muda', 'yang', 'tergabung', 'dalam', 'relawan', 'pengusaha', 'muda', 'nasional', 'untuk', 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', 'dan', 'siap', 'memenangkan', 'pasangan', 'calon', 'jokowi']	['pengusaha', 'muda', ~~'yang'~~, 'tergabung', ~~'dalam'~~, 'relawan', 'pengusaha', 'muda', 'nasional', ~~'untuk'~~, 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', ~~'dan'~~, 'siap', 'memenangkan', 'pasangan', 'calon', 'jokowi']

Tahapan	Proses	Input	Output
Lemmatization	Proses penghilangan imbuhan pada suatu kata menjadi bentuk baku kata dasarnya.	['pengusaha', 'muda', 'tergabung', 'relawan', 'pengusaha', 'muda', 'nasional', 'jokowi', 'repnas', 'sumatera', 'utara', 'sepenuh', 'hati', 'mendukung', 'siap', 'memenangkan', 'pasangan', 'calon', 'jokowi']	['usaha', 'muda', 'gabung', 'rela', 'usaha', 'muda', 'nasional', 'jokowi', 'repnas', 'sumatera', 'utara', 'penuh', 'hati', 'dukung', 'menang', 'pasang', 'calon', 'jokowi']

3.5 Pembagian Dataset

Dataset yang telah dikumpulkan dan telah di *preprocessing* selanjutnya dibagi menjadi dua bagian. Dataset tersebut yang berjumlah 500 dataset kemudian dipecah 70% untuk data *train* dan 30% untuk data *test*. Jumlah untuk data *train* sebanyak 350 data dan jumlah data *test* sebanyak 150 data.

3.6 Pembuatan *Rules* dengan CSR

Setelah dataset dibagi, selanjutnya yaitu membuat *rule* dengan menggunakan data *train.* Pertama yang perlu dibuat yaitu *sequence*. Data yang telah dibersihkan pada tahap *preprocessing* selanjutnya dibuat *sequence* dengan melakukan POS *Tagging* namun hasil dari POS *Tagging* tersebut terdapat perubahan agar sesuai dengan kebutuhan pengklasifikasian [5]. Perubahan yang dilakukan yaitu untuk setiap kata yang mengandung emosi, diganti *tag*-nya menjadi c_1 untuk kata yang mengandung emosi senang, c_2 untuk emosi sedih, c_3 untuk emosi takut, c_4 untuk emosi marah dan c_5 untuk emosi benci. Selain itu, terdapat perubahan pula pada nama kandidat menjadi k_1 untuk pasangan Jokowi dan k_2 untuk pasangan Prabowo. Setelah semua *tweet* telah menjadi *sequence*, lalu dimasukkan ke dalam *Sequence database* sebagai penampung kumpulan *sequence*.

Tabel 4 Contoh Hasil Pembuatan *Sequence*

Preprocessing	['usaha', 'muda', 'gabung', 'rela', 'usaha', 'muda', 'nasional', 'jokowi', 'repnas', 'sumatera', 'utara', 'penuh', 'hati', 'dukung', 'menang', 'pasang', 'calon','jokowi'] ['jokowi', 'pokok', 'jokowi', 'saya', 'pilih', 'orang', 'jokowi', 'orang']
Sequence	['NN', 'JJ', 'NN', 'NN', 'NN', 'JJ', 'JJ', 'k1', 'NN', 'NN', 'NN', 'JJ', 'c1', 'c1', 'c1', 'NN', 'NN', 'k1'] ['k1', 'NN', 'k1', 'PRP', 'c1', 'NN', 'k1', 'NN']

Selanjutnya, setelah didapatkan *sequence database*, dilakukan pembuatan pola untuk pembentukan *rule* dengan melakukan beberapa langkah yang harus dilakukan. Pembangunan Algoritma *PrefixSpan* dengan menggunakan *library prefixspan* pada bahasa pemrograman *Python.* Pada tahap ini, *prefix* dan *suffix* dibentuk untuk pembangunan *rule* pada tahap selanjutnya. Tabel 5 menunjukkan pembuatan *subset* dari pola *suffix* untuk kelas c_1. Namun pada implementasi yang sesungguhnya, pembuatannya ditujukkan untuk setiap kelas.

Tabel 5 Pembuatan *subset* dari Pola *Suffix* untuk Setiap Kelas

Class	*Sequence*	*Projected db*	*Suffix Pattern*
c_1	['NN', 'JJ', 'NN', 'NN', 'NN', 'JJ', 'JJ', 'k1', 'NN', 'NN', 'NN', 'JJ', 'c1', 'c1', 'c1', 'NN', 'NN', 'k1'], ['k1', 'NN', 'k1', 'PRP', 'c1', 'NN', 'k1', 'NN']	['NN', 'JJ', 'NN', 'NN', 'NN', 'JJ', 'JJ', 'k1', 'NN', 'NN', 'NN', 'JJ'] ['k1', 'NN', 'k1', 'PRP']	['c_1'], ['NN', 'c_1'], [k_1', 'c_1'], ['k_1', 'NN', 'c_1']

Setelah proses pada Tabel 5 telah terbentuk, langkah selanjutnya yaitu pembuatan *rule* dengan menggunakan pola *suffix* yang dibangkitkan sebelumnya dengan *prefix* sebagai *output*-nya.

Tabel 6 Pembuatan *Rule* dengan Pola *Suffix* yang Dibangkitkan sebagai *Prefix*

Prefix	*Projected db*	*Prefix patterns*
['c_1']	['NN', 'NN', 'k1'], ['NN', 'k1', 'NN']	['c_1', 'NN'], ['c_1', 'k_1'], ['c_1', 'NN', 'NN', 'k_1']
['NN', 'c_1']	['NN', 'NN', 'k1'], ['NN', 'k1', 'NN']	['NN', 'c_1', 'NN'], ['NN', 'c_1', 'k_1'], ['NN', 'c_1', 'NN', 'NN', 'k_1']
[k_1', 'c_1']	['NN', 'NN', 'k1'], ['NN', 'k1', 'NN']	['k_1', 'c_1', 'NN'], [' k_1', 'c_1', 'k_1'], [' k_1', 'c_1', 'NN', 'NN', 'k_1']
['k_1', 'NN', 'c_1']	['NN', 'NN', 'k1'], ['NN', 'k1', 'NN']	['k_1', 'NN', 'c_1', 'NN'], ['k_1', 'NN', 'c_1', 'k_1'], ['k_1', 'NN', 'c_1', 'NN', 'NN', 'k_1']

1) Penentuan *min_sup*

Minimum support atau *min_sup* merupakan jumlah paling sedikit *rule* yang ditinjau atau diambil pada penelitian ini. Diperlukan penentuan *min_sup* karena terdapat beberapa pola yang sering muncul dan ada yang

sangat jarang muncul. Hanya satu *min_sup* untuk mengendalikan prosedur pembangkitan tidak cukup. Sebab untuk menambang pola yang melibatkan kata yang mengandung kelas tertentu yang jarang terjadi, perlu mengatur nilai *min_sup* terendah yang akan menyebabkan seringnya kemunculan kata yang mengandung kelas tertentu. Cara penentuannya yaitu dengan mengalikan frekuensi kemunculan minimum dalam data *train* dari *sequence* pada *rule*. Dari *min_sup* yang ditentukan dapat mempengaruhi jumlah *rule* yang terambil.

Tabel 7 Pengaruh Nilai *min_sup* terhadap Jumlah *Rule* yang Terambil

Min_sup	**Jumlah *Rule***
25%	26
20%	48
15%	142
10%	486
5%	7605

2) Pemilihan *Rule*

Setelah beberapa *rule* terbangun, dan *min_sup* telah ditentukan, maka dilakukan pemilihan terhadap *rule* yang telah terbangun. *Rule* yang terbangun tidak terpilih semua, hanya *rule* yang yang mempunyai frekuensi kemunculan paling sering dan memenuhi nilai *min_sup* yang ditentukan. Beberapa *rule* yang tidak terambil antara lain karena jika *rule* tersebut mengandung kelas emosi terletak dipaling belakang maka tidak akan diambil. Begitu juga dengan kelas emosi yang berdiri sendiri, maka tidak akan terpilih.

Tabel 8 Detail Jumlah *Rule* yang Terambil

Min_sup	**Jumlah *Rule***	**Detail Jumlah *Rule***
25%	26	5 senang, 4 sedih, 1 takut, 0 marah, 2 benci, 8 jokowi, 6 prabowo
20%	48	9 senang, 10 sedih, 2 takut, 1 marah, 6 benci, 12 jokowi, 8 prabowo
15%	142	23 senang, 54 sedih, 4 takut, 1 marah, 16 benci, 28 jokowi, 16 prabowo
10%	486	59 senang, 227 sedih, 12 takut, 11 marah, 64 benci, 77 jokowi, 36 prabowo
5%	7605	245 senang, 6117 sedih, 183 takut, 105 marah, 397 benci, 407 jokowi, 148 prabowo

3) Pengecekan Data *Test*

Pada tahap ini, dilakukan pengecekan terhadap data *test* yang digunakan berdasarkan *rule* yang telah dipilih sebelumnya dengan melakukan serangkaian proses *preprocessing* hingga menghasilkan suatu *sequence*. Apabila terdapat kesamaan antara *sequence* pada data *test* dengan *rule* yang terdapat pada data *train*, maka data tersebut akan masuk kedalam tahap selanjutnya yaitu perhitungan kelas emosi.

3.8 Pembangunan Model dengan NBC

Pada tahap ini, menggunakan *rule* yang terpilih di tahap sebelumnya. Salah satu contoh *sequence* yang terbentuk oleh emosi senang (c_1) tersebut yaitu ['k_1', 'NN', 'c_1', 'NN'] yang masuk ke dalam salah satu *rule* ['k_1', 'NN', 'c_1', 'NN'] . Kemudian, *rule* tersebut dihitung frekuensi kemunculan pada setiap kelas emosinya. *Rule* yang sering mucul pada suatu kelas, menentukan *rule* tersebut masuk ke dalam kelas emosi yang memiliki jumlah frekuensi kemunculan terbanyak. Setelah itu, *rule* pada masing-masing kelas dihitung jumlahnya.

Tabel 9 Contoh Perhitungan Peluang Kelas Emosi

	Dokumen	*Rule*	Kelas
Data Latih	1	['k_2', 'NN', 'c_1', 'NN']	Senang
	2	['k_1', 'NN', 'c_2', 'k_1']	Sedih
	3	['k_2', 'NN', 'c_3', 'NN']	Takut
	4	['k_2', 'NN', 'c_4', 'NN']	Marah
	5	['k_1', 'NN', 'c_5', 'k_1']	Benci
	6	['k_1', 'NN', 'c_1', 'NN']	Senang
Data Uji	7	['k_2', 'NN', 'c_1', 'k_2', 'k_1']	?

Langkah-langkah untuk membangun model menggunakan NBC sebagai berikut.

- Menghitung *prior probability* dari kelima kelas yang ada.

$$P(senang) = \frac{2}{6} = 0{,}33 \qquad P(sedih) = \frac{1}{6} = 0{,}16 \qquad P(takut) = \frac{1}{6} = 0{,}16$$

$$P(marah) = \frac{1}{6} = 0{,}16 \qquad P(benci) = \frac{1}{6} = 0{,}16$$

- Setelah mendapatkan nilai *prior probability* untuk setiap kelas, selanjutnya menghitung *conditional probability*.

$P(k2\|senang) = \frac{1+1}{8+24} = 0{,}06$ $P(NN\|senang) = \frac{4+1}{8+24} = 0{,}15$ $P(c1\|senang) = \frac{2+1}{8+24} = 0{,}09$ $P(k2\|senang) = \frac{1+1}{8+24} = 0{,}06$ $P(k1\|senang) = \frac{1+1}{8+24} = 0{,}06$	$P(k2\|sedih) = \frac{0+1}{4+24} = 0{,}03$ $P(NN\|sedih) = \frac{1+1}{4+24} = 0{,}07$ $P(c1\|sedih) = \frac{0+1}{4+24} = 0{,}03$ $P(k2\|sedih) = \frac{0+1}{4+24} = 0{,}03$ $P(k1\|sedih) = \frac{2+1}{4+24} = 0{,}1$
$P(k2\|takut) = \frac{1+1}{4+24} = 0{,}07$ $P(NN\|takut) = \frac{2+1}{4+24} = 0{,}1$ $P(c1\|takut) = \frac{0+1}{4+24} = 0{,}03$ $P(k2\|takut) = \frac{1+1}{4+24} = 0{,}07$ $P(k1\|takut) = \frac{0+1}{4+24} = 0{,}03$	$P(k2\|marah) = \frac{1+1}{4+24} = 0{,}07$ $P(NN\|marah) = \frac{2+1}{4+24} = 0{,}1$ $P(c1\|marah) = \frac{0+1}{4+24} = 0{,}03$ $P(k2\|marah) = \frac{1+1}{4+24} = 0{,}07$ $P(k1\|marah) = \frac{0+1}{4+24} = 0{,}03$

$$P(k2|benci) = \frac{0+1}{4+24} = 0{,}03$$
$$P(NN|benci) = \frac{1+1}{4+24} = 0{,}07$$
$$P(c1|benci) = \frac{0+1}{4+24} = 0{,}03$$
$$P(k2|benci) = \frac{0+1}{4+24} = 0{,}03$$
$$P(k1|benci) = \frac{2+1}{4+24} = 0{,}1$$

- Setelah didapatkan nilai *conditional probability* data pada masing-masing kelas, akan dilakukan perhitungan *likelihood* untuk menentukan data tersebut termasuk ke dalam kelas senang, sedih, takut, marah atau benci.

$$P(senang|Dokumen7) = 0{,}33 \times 0{,}06 \times 0{,}15 \times 0{,}09 \times 0{,}06 \times 0{,}06 = 1{,}6 \times 10^{-5}$$
$$P(sedih|Dokumen7) = 0{,}16 \times 0{,}15 \times 0{,}25 \times 0{,}15 \times 0{,}15 \times 0{,}05 = 3{,}02 \times 10^{-8}$$
$$P(takut|Dokumen7) = 0{,}16 \times 0{,}15 \times 0{,}25 \times 0{,}15 \times 0{,}15 \times 0{,}05 = 2{,}82 \times 10^{-5}$$
$$P(marah|Dokumen7) = 0{,}16 \times 0{,}15 \times 0{,}25 \times 0{,}15 \times 0{,}15 \times 0{,}05 = 2{,}82 \times 10^{-5}$$
$$P(benci|Dokumen7) = 0{,}16 \times 0{,}15 \times 0{,}25 \times 0{,}15 \times 0{,}15 \times 0{,}05 = 2{,}82 \times 10^{-5}$$

Hasil perhitungan akhir dari NBC sudah didapatkan yang kemudian akan dibandingkan nilai probabilitas mana yang paling besar diantara kedua kelas tersebut. Untuk hasil perhitungan dokumen 7 terlihat bahwa nilai probabilitasnya lebih besar di kelas sedih daripada kelas emosi yang lain. Maka dapat disimpulkan jika dokument 7 termasuk kelas sedih.

Kemudian contoh selanjutnya pada tahap ini, menggunakan *rule* yang masuk kelas emosi senang. Salah satu contoh *sequence* yang terbentuk pada data *test* yaitu ['k1', 'NN', 'k1', 'PRP', 'c1', 'NN', 'k1', 'NN'], maka dapat masuk ke dalam salah satu *rule* yang terbentuk oleh kandidat 1 (jokowi) yaitu ['k_1', 'NN', 'c_1', 'NN']. Kemudian, *rule* tersebut dihitung frekuensi kemunculan pada setiap kelas kandidatnya. *Rule* yang sering mucul pada suatau kelas, menentukan *rule* tersebut masuk ke dalam kelas kandidat yang memiliki jumlah frekuensi kemunculan terbanyak. Sedangkan *rule* yang masuk ke daam kelas selain emosi senang, menambah jumlah frekuensi kemunculan kandidat yang lain.

Tabel 10 Contoh Perhitungan Peluang Kelas Kandidat

	Dokumen	*Rule*	Kelas
Data Latih	1	['k2', 'NN', 'c_1', 'NN']	Prabowo
	2	['k_1', 'NN', 'c_1', 'k_1']	Jokowi
	3	['k2', 'NN', 'c_1', 'NN']	Prabowo
Data Uji	4	['k2', 'NN', 'c_1', 'k2', 'k1']	?

Langkah-langkah untuk membangun model menggunakan NBC sebagai berikut.

- Menghitung *prior probability* dari kedua kelas yang ada.

$$P(jokowi) = \frac{1}{3} = 0{,}33 \qquad P(prabowo) = \frac{2}{3} = 0{,}67$$

- Setelah mendapatkan nilai *prior probability* untuk setiap kelas, selanjutnya menghitung *conditional probability*.

$$P(k2|jokowi) = \frac{0+1}{4+12} = 0{,}06 \qquad P(k2|prabowo) = \frac{2+1}{8+12} = 0{,}15$$
$$P(NN|jokowi) = \frac{1+1}{4+12} = 0{,}12 \qquad P(NN|prabowo) = \frac{4+1}{8+12} = 0{,}25$$
$$P(c1|jokowi) = \frac{1+1}{4+12} = 0{,}12 \qquad P(c1|prabowo) = \frac{2+1}{8+12} = 0{,}15$$
$$P(k2|jokowi) = \frac{0+1}{4+12} = 0{,}06 \qquad P(k2|prabowo) = \frac{2+1}{8+12} = 0{,}15$$
$$P(k1|jokowi) = \frac{2+1}{4+12} = 0{,}18 \qquad P(k1|prabowo) = \frac{0+1}{8+12} = 0{,}05$$

- Setelah didapatkan nilai *conditional probability* data pada masing-masing kelas, akan dilakukan perhitungan *likelihood* untuk menentukan data tersebut ke dalam kelas jokowi atau prabowo

Hasil perhitungan akhir dari NBC sudah didapatkan yang kemudian akan dibandingkan nilai probabilitas mana yang paling besar diantara kedua kelas tersebut. Untuk hasil perhitungan dokumen 4 terlihat bahwa nilai probabilitasnya lebih besar di kelas prabowo daripada kelas jokowi. Maka, dapat disimpulkan dokumen 4 termasuk kelas prabowo.

3.9 Pengklasifikasian Keberpihakan

Pada tahap ini, setelah semua data *test* telah masuk ke dalam kelas kandidat, selanjutnya dihitung persentase dari setiap kandidatnya. Perhitungannya dengan cara menghitung jumlah salah satu kandidat dibagi dengan jumlah keseluruhan data *test* dikali dengan 100% untuk mendapatkan hasil persentasenya.

4 Evaluasi

Pengujian sistem pada penelitian ini dilakukan untuk melihat tingkat performansi sistem yang digunakan untuk melakukan identifikasi keberpihakan *tweet* pada *Twitter*. Tingkat performansi diukur dari *precission* dan *recall*. Data *set* yang digunakan dalam pengujian ini berasal dari *tweet* yang diambil pada masa kampanye Pilpres 2019 berdasarkan kata kunci yang disesuaikan.

4.1 Hasil Pengujian

Hasil pengujian dalam penelitian ini didapat dari skenario pengklasifikasian keberpihakan *tweet* terhadap suatu paslon dengan menggunakan data *train* 350 dan data *test* 150 serta *min_sup* 20%. Adapun hasil pengujian dari sistem yang telah dibuat tersebut. Berikut adalah hasil perhitungan *precision*, *recall*, dan *F1-score* adalah seperti berikut.

Tabel 11 Hasil Perhitungan *Precision, Recall,* dan *F1-Score*

Precicion	*Recall*	*F1-score*
54.33%	78.41%	64.19%

Hasil tersebut menunjukkan bahwa perolehan tingkat dari performansi sistem sebesar 64.19%. Selain itu, dari 150 data *test* yang telah diuji, menghasilkan dua kelas keberpihakan *tweet* pada *Twitter* berdasarkan hasil klasifikasi emosi yang telah dilakukan sebelumnya, yaitu *tweet* yang berpihak kepada Jokowi dan *tweet* yang berpihak kepada Prabowo dengan presentase 58.67% untuk Jokowi dan 41.33% untuk Prabowo.

4.2 Analisis Hasil Pengujian

Dari hasil pengujian di atas ada beberapa faktor yang dapat mempengaruhi hasil performansi sistem, yaitu penentuan nilai *min_sup*. Hal tersebut dapat mempengaruhi banyaknya *rule* yang dihasilkan dan yang digunakan karena berdasarkan dengan enam kali percobaan nilai *min_sup* menghasilkan *F1-Score* yang berbeda-beda.

Tabel 12 Analisis Pengaruh Min_sup Terhadap *F1-Score*

Min_sup	**Jumlah *Rule***	***F1-Score***
25%	26	64.19%
20%	48	64.19%
15%	142	63.85%
10%	486	53.33%
5%	7605	40.91%

Penentuan *min_sup* dilakukan dengan mencoba dari nilai *minimum* terbesar hingga terkecil mana yang paling optimal untuk mengambil *rule* dari data *train* tersebut. Selanjutnya dipilih *min_sup* sebesar 20% dari jumlah *rule* tiap kelas. Sehingga setidaknya tedapat lima kelas emosi dan dua kelas kandidat yang masing-masing membangkitkan *rule* sebanyak 42 dari jumlah total semua *rule*. Pada penelitian ini telah menghasilkan jumlah *rule* yang terpilih dari setiap kelas dengan *rule* berjumlah 8 untuk kelas emosi senang, 9 untuk emosi sedih, 2 kelas

emosi takut, 1 kelas emosi marah, 3 kelas emosi benci, 11 kelas kandidat jokowi dan 6 kelas prabowo. *Min_sup* yang diambil yaitu sebesar 20% karena dilihat dari hasil performansi yang juga tinggi. *Min_sup* dengan nilai diatas 25% tidak digunakan karena pada pengambilan *rule* yang terbentuk, terdapat kelas yang kosong karena tidak terambil oleh *min_sup*.

Hasil dari *F1-Score* dipengaruhi oleh nilai *min_sup* yang ditentukan dan jumlah data pada dataset. Dataset tersebut sebelumnya telah melalui tahap *preprocessing*. Data yang digunakan untuk *training* sebanyak 350 dan *testing* sebanyak 150. Data tersebut sangat bervariasi sehingga menghasilkan kualitas yang berbeda-beda setiap datanya. Hasil dari *preprocessing* akan digunakan dalam pembuatan *rule* sehingga semakin mendetail *preprocessing*nya, dataset yang diolah juga semakin bersih dan menghasilkan *rule* yang sesuai.

5 Kesimpulan dan Saran

Berdasarkan hasil pengujian dan pembahasan analisis yang telah dipaparkan, terdapat beberapa hal yang dapat disimpulkan, yaitu sebagai berikut.

1. Hasil identifikasi keberpihakan yang didapatkan melalui sistem yaitu *tweet* yang berpihak kepada Jokowi dan kepada Prabowo dengan menggunakan *Naive Bayes Classifier* berdasarkan klasifikasi emosi menggunakan *Class Sequential Rules*.
2. Sistem menghasilkan *F1-Score* sebesar 67.83% pada *tweet* yang berkaitan dengan pilpres dengan pengambilan nilai *min_sup* sebesar 20% dan menghasilkan 42 *rule* dengan rincian 8 untuk kelas emosi senang, 9 untuk emosi sedih, 2 kelas emosi takut, 1 kelas emosi marah, 3 kelas emosi benci, 11 kelas kandidat jokowi dan 6 kelas prabowo.

Sedangkan untuk pengembangan penelitian selanjutnya, ada beberapa hal yang direkomendasikan yaitu sebagai berikut ini.

1. Masing-masing tahap *preprocessing* lebih diperdalam lagi agar dapat menghasilkan data yang berkualitas karena jika menggunakan *library* masih terbatas. Sehingga dengan pendalaman tahap tersebut, diharapkan dapat meningkatkan akurasi.