

**LAPORAN PROYEK AKHIR
PRAKTIKUM DATA SCIENCE**

**SENTIMEN ANALISIS TERHADAP
#JOKOWI3PERIODE BASED ON TWITTER
MENGUNAKAN METODE NAÏVE BAYES
CLASSIFIER**



MOCHAMMAD ALFAIN KURNIAWAN

123190065

DAMAI SUMURUNG HUTAURUK

123190074

**PROGRAM STUDI INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK INDUSTRI
UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”
YOGYAKARTA
2021**

1. PENDAHULUAN

Pemilihan Presiden Republik Indonesia yang akan dilangsungkan pada tahun 2024 sudah mulai banyak menjadi obrolan ditengan masyarakat. Sejak beberapa saat lalu muncul tokoh-tokoh tak terduga yang menyampaikan pernyataan ingin mencalonkan diri menjadi Presiden Republik Indonesia tahun 2024 nanti. Hal ini menyebabkan munculnya banyak pendapat dan opini tentang calon Presiden Republik Indonesia 2024, baik di dunia nyata maupun dunia maya. Opini-opini yang muncul sangat beragam, tidak hanya opini positif atau netral tapi juga yang bersifat negatif juga muncul.

Perkembangan media yang begitu pesat, memunculkan banyak media online dari media berita sampai media sosial. Media sosial saja sudah begitu banyak, dari Facebook, Twitter, Path, Instagram, Google+, Tumblr, LinkedIn dan sebagainya masih banyak lagi. Media sosial sekarang ini tidak hanya digunakan sebagai sarana pertemanan, mencari teman, tapi sudah banyak digunakan untuk kegiatan lain. Promo dagangan, jual beli apa saja sampai promo partai politik atau kampanye calon calon legislative dan presiden.

Analisis sentimen atau opinion mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Opini publik yang muncul selama ini menghasilkan sebuah tagar yang sempat viral di media sosial terutama twitter, yaitu #jokowi3periode. Tagar ini memunculkan pro dan kontra di dunia maya. Banyak yang mendukung isi dari tagar tersebut yang mengatakan bahwa Jokowi layak menjadi presiden untuk 3 periode. Namun tidak sedikit juga yang menyampaikan opini tidak setuju akan isi dari tagar tersebut. Oleh sebab itu, penelitian ini ingin melakukan analisis sentimen untuk melihat sebuah opini seseorang yang ditujukan kepada #jokowi3periode, opini itu bisa dimasukan kategori opini positif, netral atau negative. Besarnya sentimen yang ditujukan kepada #jokowi3periode bisa dijadikan sebuah parameter sebuah pertimbangan pemerintah untuk merombak peraturan perhal masa jabatan Presiden di Indonesia.

2. METODE

Langkah-langkah penelitian sesuai dengan alur penelitian adalah sebagai berikut:

1. Mengumpulkan data *tweet*

Data tweet diambil dengan metode Crawling dari media sosial Twitter. Data yang diambil hanya tweet dalam bahasa Indonesia, yaitu 100 tweet dengan kata kunci jokowi3periode. Data diambil secara acak baik dari user biasa ataupun media online di Twitter.

2. Preprocessing Data

Pada tahap preprocessing, dilakukan 4 langkah sebagai berikut.

1) Seleksi komentar

Pada tahap ini, dilakukan seleksi komentar yang mengandung kata atau hashtag(#) jokowi3periode karena Twitter terdapat fungsi retweet, yaitu memberikan komentar terhadap tweet komentar seseorang, karena komentar tweet akan mengganggu dalam proses Analisis Sentimen tweet. Jadi dalam preprocessing ini komentar tweet dihapus.

2) *Cleansing*

Kalimat yang didapat biasanya masih terdapat noise, yaitu kesalahan acak atau varian dalam variable terukur, karena itu harus menghilangkan noise tersebut. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, hashtag (#), username (@username), url (<http://website.com>), dan email (nama@website.com).

3) *Parsing*

Parsing adalah proses memecah dokumen menjadi sebuah kata dengan melakukan analisa terhadap kumpulan kata dengan memisahkan kata tersebut dan menentukan struktur sintaksis dari tiap kata tersebut.

4) Normalisasi Kalimat

Bertujuan untuk menormalkan kalimat sehingga kalimat gaul menjadi normal, sehingga bahasa gaul tersebut dapat dikenali sebagai bahasa yang sesuai dengan KBBI. Yang harus dilakukan untuk normalisasi kalimat adalah:

1. Meregangkan tanda baca (*punctuation*) dan simbol selain alfabet
2. Meregangkan tanda baca adalah memberikan jarak terhadap tanda baca dari kata-kata sesudah atau sebelumnya, tujuannya agar tanda baca dan simbol selain alphabet tidak menjadi satu dengan kata-kata pada saat proses tokenisasi.
3. Mengubah menjadi huruf kecil semua
4. Normalisasi kata

5) Menghilangkan huruf yang berulang

Ketika sedang senang atau kesal, seseorang bebas menuliskan opini berdasarkan emosinya, biasanya seseorang menuliskan dengan mengulang huruf yang sama. Contohnya: “tidaaak” untuk mengekspresikan kesenangan. Kata berulang seperti “tidaaak” akan di normalisasi menjadi “tidak”.

6) Menghilangkan emoticon

Ketika sedang menulis status (tweet) seseorang kadang salah atau kurang tepat dalam penggunaan emoticon. Contohnya: Mereka hanya bisa memfitnah karena tidak bisa ketemu fakta buruk :), kata opini fitnah tapi emoticonnya senyum :), dengan begitu emotikon akan mengganggu dalam proses Sentimen Analisis tweet, jadi dalam proses ini emotikon dihapus atau diabaikan.

3. Tokenisasi

Setelah normalisasi kalimat, selanjutnya kalimat tersebut dipecah kedalam token-token menggunakan pembatas atau delimiter spasi. Token yang digunakan dalam penelitian ini yaitu: N-gram: token yang terdiri dari tiga kata setiap satu token, contohnya: Seluruh Warga Negara.

4. *Part of Speech (POS) Tagger*

POS *tagger* adalah sebuah proses untuk memberikan kelas pada sebuah kata. Dalam proses POS *tagger* dilakukan dengan cara *parsing*, kemudian ditentukan kelas tiap kata dengan menggunakan bantuan kamus yang di buat sendiri berdasarkan Kamus Besar Bahasa Indonesia (KBBI) menggunakan metode *Maximum Entropy*. Proses POS *tagging* terbagi ke dalam tiga proses yaitu pemisahan setiap token dalam dokumen dengan pengecekan setiap kata dalam dokumen, mengidentifikasi setiap kata dalam dokumen dengan pemberian jenis kata, pengecekan kata yang belum teridentifikasi terhadap bentuk imbuhan dan akhiran sehingga diperoleh kata dasar.

Berdasarkan aturan linguistik pada kata diperoleh sentimen sementara. Penentuan sentimen dilakukan dengan melihat adanya kata yang mengandung opini baik yang memiliki polarity positif maupun negatif dari tweet yang sudah dilabeli kelas katanya. Kelas kata yang dipilih adalah kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*) dan kata kerja (*verb*), sesuai dengan penelitian bahwa keempat jenis kata di atas merupakan jenis kata yang paling banyak mengandung sentimen. Dalam sistem ini jika suatu tweet terdapat kata benda (NN) pada sebelum atau setelah kata sifat (JJ) atau kata keterangan (RB) dan kata benda (memiliki polarity berlawanan dengan kata sifat atau kata keterangan maka polarity yang diperoleh berdasarkan kata sifat atau kata keterangan, karena kata sifat atau kata keterangan memberikan penegasan terhadap kata benda.

5. Penentuan Class Attribute

Data Twitter yang sudah dilakukan *Preprocessing* kemudian akan ditentukan *class attribute*. *Class attribute* yang dimunculkan dalam penelitian ini ada 3, diantaranya positif, netral, dan negatif. Dengan 3 *class attribute* ini diharapkan mampu memberi penilaian masyarakat secara akurat terhadap objek yang menjadi inti dari penelitian.

6. Load Dictionary

Setelah dilakukan tokenisasi dan ditentukan *class attribute*, langkah selanjutnya adalah *load dictionary*. Banyak jenis kamus yang dapat digunakan, contohnya: kamus kata kunci sentimen positif (*positive keywords*), kamus kata kunci sentimen negatif (*negative keywords*), kamus kata negasi (*negation*

keywords), dan kamus normalisasi bahasa gaul atau alay. Berikut adalah contoh kamus dan isinya :

- 1) *Positive keywords*: baik, hebat, jujur, cerdas, keren.
- 2) *Negative keywords*: bohong, korupsi, jahat, jelek.
- 3) *Negation keywords*: nggak, tidak, bukan, jauh.
- 4) Kamus konversi bahasa gaul ke KBBA: sp = siapa, spt = seperti, brp = berapa, hrg = harga, ciyus = serius.

7. Determine Sentiment

Pada proses ini metode *Lexicon Based* digunakan, yaitu saat menentukan sentimen suatu kalimat opini, penentuan dilakukan dengan penjumlahan n skor polaritas kata opini p yang mengomentari fitur f. Skor polaritas suatu kata opini p akan bernilai 1 jika kata tersebut adalah kata opini positif, dan bernilai -1 jika kata tersebut adalah kata opini negatif.

Setelah diketahui kata yang mengandung positif, negatif dan netral di dalam sebuah kalimat, selanjutnya dihitung bobot nilai yang terkandung dalam kalimat tersebut yang dilakukan dengan menjumlahkan nilai kata opini. Jika jumlah nilai opini dalam kalimat tersebut 1, maka nilai sentimen dari kalimat tersebut adalah positif, jika nilai opini dalam kalimat tersebut = 0, maka nilai sentimen dari kalimat tersebut adalah netral, jika nilai opini dalam kalimat tersebut = -1, maka nilai sentimen dari kalimat tersebut adalah negatif.

8. Klasifikasi

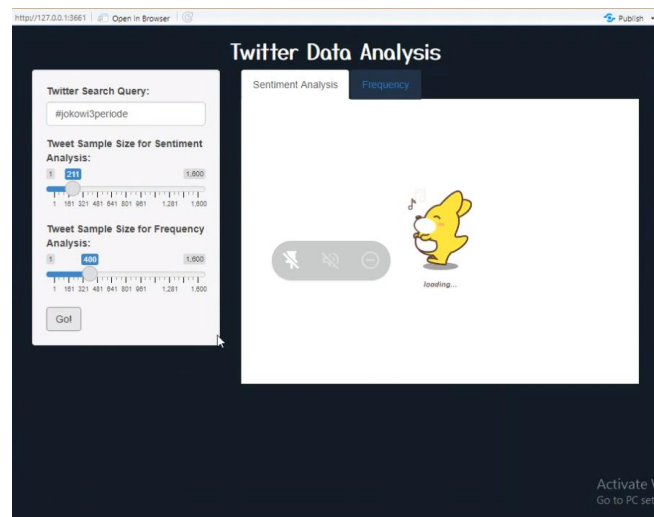
Metode klasifikasi yang digunakan dalam penelitian ini adalah Naïve Bayes Classifier (NBC). *Naïve Bayes Classifier* (NBC) adalah metode klasifikasi berdasarkan kemungkinan dan Teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas (*independence*). Dengan kata lain, *Naïve Bayesian Classifier* (NBC) mengansumsikan bahwa keberadaan sebuah variabel tidak ada kaitannya dengan beradaan variabel yang lain. Berikut adalah Rumusnya.

$$P(H|X) = (P(X|H)P(H))/P(X)$$

Pada proses klasifikasi data diuji dengan menggunakan metode *10 fold cross validation*. Jadi dataset akan dibagi menjadi dua, yaitu 10 bagian dengan 9/10 bagian digunakan untuk proses *training* dan 1/10 bagian digunakan untuk proses *testing*. Iterasi berlangsung 10 kali dengan variasi data *training* dan *testing* menggunakan kombinasi 10 bagian data.

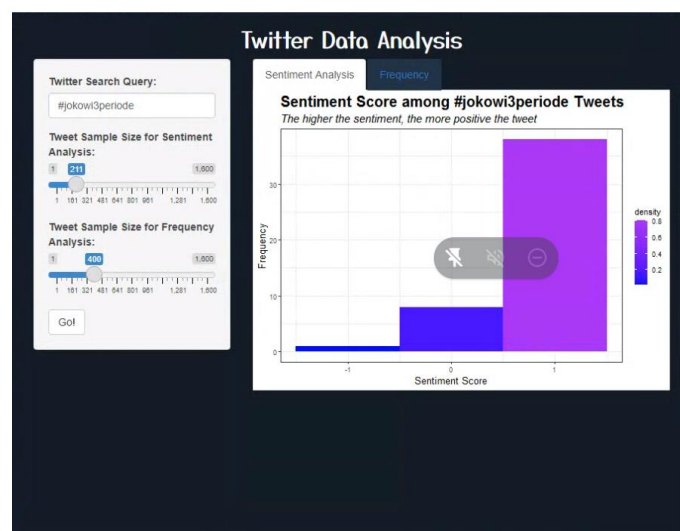
3. HASIL DAN PEMBAHASAN

Dataset pada penelitian ini menggunakan data opini yang dikumpulkan dari Twitter dengan metode Crawling dari media sosial Twitter. Data yang diambil hanya tweet dalam bahasa Indonesia, yaitu tweet dengan kata kunci jokowi3periode. Data diambil secara acak baik dari user biasa ataupun media online di Twitter.



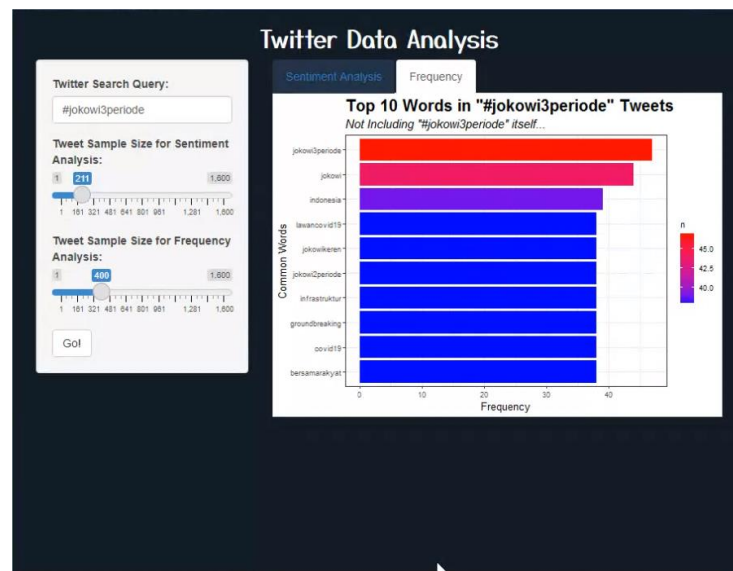
Gambar 1. Tampilan aplikasi yang sudah dibuat dalam bentuk shiny

Dataset yang akan kita gunakan bisa kita tentukan secara bebas. Kita dapat menentukan jumlah data opini yang untuk dilakukan sentimen analisis nantinya. Disini juga kita menambahkan fungsi untuk melakukan perangkikan untuk jumlah kata yang paling banyak berhubungan dengan #jokowi3periode. Kita juga dapat menentukan jumlah data opini yang akan di lakukan perangkikan. Dalam pengaplikasiannya, dervice harus terhubung dengan internet agar bisa mengambil data dari media twitter secara *real time*.



Gambar 2. Tampilan hasil penilaian sentiment

Dari Gambar 2. dapat dilihat bahwa nilai -1 adalah tanggapan yang bersifat negatif, nilai 0 adalah tanggapan yang bersifat netral, sedangkan nilai 1 adalah tanggapan yang bersifat positif. Pada Gambar 2. dapat dilihat bahwa data yang digunakan berjumlah 211 opini yang diambil dari twitter. Hasil yang diperoleh menunjukkan bahwa #jokowi3periode memperoleh tanggapan positif yang sangat unggul dibandingkan tanggapan yang bersifat netral maupun tanggapan negative. Sehingga, semakin tinggi sentimennya, semakin tinggi pula tweet positifnya.



Gambar 3. Tampilan hasil perangkingan kata

Dari Gambar 3. dapat dilihat bahwa dari 400 data opini yang digunakan, kata-kata yang paling banyak berhubungan dengan #jokowi3periode adalah jokowi3periode. Pengurutan dimulai dari kata paling atas sebagai kata yang paling berhubungan dan menampilkan 10 kata teratas yang paling banyak digunakan. Hasil ini dapat menjadi penelitian lanjutan untuk mengetahui bahasan yang terjadi seputaran #jokowi3periode. Kata yang diambil hanya yang berhubungan dengan #jokowi3periode, sehingga kata #jokowi3periode tidak termasuk di dalamnya.

4. KESIMPULAN

Dari penelitian yang telah dilakukan, maka dapat ditarik kesimpulan bahwa Analisis Sentimen dapat digunakan untuk mengetahui sentimen masyarakat khususnya netizen Twitter terhadap #jokowi3periode. Tujuannya membantu masyarakat menentukan sentimen yang terdapat pada tweet opini Bahasa Indonesia yang ada di Twitter. Setelah dilakukan analisis sentimen, terlihat berapa banyak sentimen yang ditujukan kepada #jokowi3periode. Hasil yang diperoleh dari penilaian sentiment seputar #jokowi3periode di twitter adalah semakin tinggi sentimennya, semakin tinggi pula tweet positifnya untuk #jokowi3periode. Untuk penelitian selanjutnya perlu dicoba dikembangkan dengan menggunakan data yang lebih banyak. Perlu dikembangkan juga stopword list dan stemmer Bahasa Indonesia yang mampu meningkatkan akurasi dalam analisis sentimen Bahasa Indonesia.

Referensi

- Franky dan Manurung, R., (2008). *Machine Learning-based Sentiment Analysis of Automatic Indonesian Translations of English Movie Reviews*. In Proceedings of the International Conference on Advanced Computational Intelligence and Its Applications.
- Go, A., Huang, L., & Bhayani, R. (2009). *Twitter Sentiment Analysis*. Final Project Report, Stanford University, Department of Computer Science.
- G. A. Buntoro, (2016). "*Sentiment Analysis Candidates of Indonesian Presiden 2014 with Five Class Attribute*" in International Journal of Computer Applications (0975 – 8887), Volume 136 – No.2, February 2016.
- G. A. Buntoro, T. B. Adji, and A. E. Purnamasari, (2014). "Sentiment Analysis Twitter dengan Kombinasi Lexicon Based dan Double Propagation," in *The 6th Conference on Information Technology and Electrical Engineering (CITEE)*, 2014, pp. 39-43