
Analisis Sentimen Masyarakat terhadap Hasil *Quick Count* Pemilihan Presiden Indonesia 2019 pada Media Sosial *Twitter* Menggunakan Metode *Naive Bayes Classifier*

Lingga Aji Andika¹, Pratiwi Amalia Nur Azizah², and Respatiwulan³
^{1,2,3} Program Studi Statistika FMIPA Universitas Sebelas Maret

linggaajiandika@gmail.com, pratiwiamalia06@gmail.com, rrespatiwulan@yahoo.com

Abstract. Indonesia is one of the countries that adheres to a democratic system. In the course of a democratic system it is marked by periodic general elections. In 2019 Indonesia held a general election simultaneously to elect the President, DPR, DPRD and DPD. After the election, a lot of opinion arise within the community, including on social media twitter. One of the topics discussed was the results of the quick count of the presidential election. Therefore, a method that can be used to analyze sentiment from the quick count opinion is needed, that is naive Bayes method. The aims of this study are to find the best naive Bayes model and to classify sentiments. The result shows the best accuracy of 82.90% with $\alpha = 0.05$. The classification obtained is 34.5% (471) positive tweets and 65.5% (895) negative tweets on the results of the quick count.

Keywords : sentiment analysis, naive Bayes classifier, elections, quick count

1. Pendahuluan

Demokrasi adalah salah satu bentuk pemerintahan dimana semua warga negaranya memiliki hak yang sama dalam pengambilan keputusan yang dapat mengubah hidup mereka. Demokrasi mengizinkan warga negara berpartisipasi, baik secara langsung atau melalui perwakilan dalam perumusan, pengembangan, dan pembuatan hukum [1]. Indonesia adalah salah satu negara yang menganut sistem demokrasi. Dalam keberjalanan sistem demokrasi ditandai dengan diadakannya pemilihan umum secara periodik. Pada tahun 2019 Indonesia melaksanakan pemilihan umum serentak untuk memilih Presiden, DPR, DPRD, dan DPD.

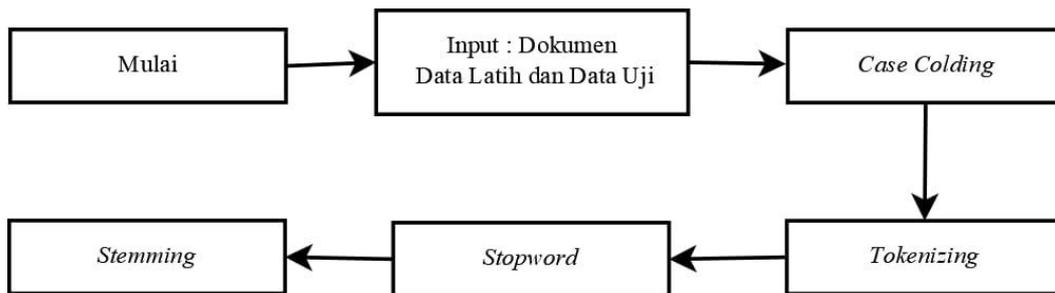
Pasca-dilaksanakannya pemilihan umum serentak pada tanggal 17 April 2019, muncul masalah di tengah-tengah masyarakat terkait hasil *quick count*, terutama untuk pemilihan presiden. Timbul berbagai opini terhadap hasil *quick count* baik menanggapi hasil *quick count* secara positif maupun negatif.

Media sosial adalah salah satu tempat bagi seseorang untuk menuangkan ide, gagasan, maupun opini sehingga dapat diketahui oleh orang lain. Salah satu media sosial yang sering digunakan masyarakat adalah media sosial *twitter*. Dalam laman *twitter* pasca-pemilu, banyak perbincangan di dunia maya terkait hasil *quick count* di media sosial. Opini-opini yang ada di laman *twitter* menarik untuk diklasifikasikan ke dalam

sentimen positif ataupun sentimen negatif. Klasifikasi sebuah opini merupakan bagian dari *data mining* khususnya *text mining* atau penambangan teks. Pada penelitian ini penulis menggunakan metode *naive Bayes* untuk mengklasifikasikan sentimen dari hasil *crawling twitter* dengan topik hasil *quick count* pemilu di Indonesia tahun 2019. *Naive Bayes* adalah salah satu algoritma pembelajaran induktif yang paling efektif dan efisien untuk *machine learning* dan *data mining*.

2. Landasan Teori

2.1. Preprocessing. *Preprocessing* merupakan teknik *data mining* yang melibatkan perubahan data mentah menjadi sebuah format yang terstruktur dan dimengerti. Data mentah seringkali tidak lengkap, tidak konsisten, dan mungkin mengandung banyak kesalahan. Teknik *preprocessing* terbukti dapat menyelesaikan masalah tersebut. Ada beberapa tahapan dalam teknik *preprocessing* yaitu *tokenization*, *case folding*, *filtering*, dan *stemming* [2].



Gambar 1. Proses *Preprocessing*

Case folding merupakan tahapan untuk mengubah semua huruf dalam dokumen menjadi huruf kecil (*lowercase*). Tahapan kedua, yaitu *tokenizing* merupakan tahapan pemotongan *string input* berdasarkan tiap kata yang menyusunnya. *Tokenizing* juga digunakan untuk membuang beberapa karakter yang dianggap sebagai tanda baca. Selanjutnya, *stopword* adalah tahap mengambil kata penting dari hasil *tokenizing* dengan algoritma *stopword removal* (membuang kata yang kurang penting). *Stopword* merupakan kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag of words*. Tahap terakhir, *stemming* adalah proses normalisasi dalam sistem *information retrieval* yang digunakan untuk mencari kata dasar (*root word*) yang terdapat dalam suatu dokumen atau *term* dengan mengacu pada aturan-aturan tertentu.

2.2. Naive Bayes Classifier. *Naive Bayes classifier* merupakan sebuah metode klasifikasi dengan probabilitas sederhana yang mengaplikasikan teorema Bayes dengan tidak ketergantungan (*independen*) yang tinggi. Algoritma *naive Bayes* digunakan

sebagai penggolong dalam berbagai masalah dunia nyata seperti analisis sentimen, deteksi *spam email*, pengelompokan otomatis *email*, pengurutan *email* berdasarkan prioritas, dan kategorisasi dokumen. Model klasifikasi *naive Bayes* menghitung probabilitas posterior suatu kelas, berdasarkan pada distribusi kata-kata dalam dokumen [3]. Klasifikasi *naive Bayes* dibangun oleh data pelatihan untuk memperkirakan probabilitas dari setiap kategori yang terdapat pada ciri dokumen yang diuji. Sistem akan dilatih dengan menggunakan data baru (data latih dan data uji) dan selanjutnya diberi tugas untuk menebak nilai fungsi target dari data tersebut [4]. Secara umum, proses klasifikasi dengan menggunakan *naive Bayes* dapat dilihat dari persamaan :

$$P(c_j|w_i) = \frac{P(w_i|c_j) \times P(c_j)}{P(w_i)}$$

$P(c_j|w_i)$: peluang kategori j ketika terdapat kemunculan kata i

$P(w_i|c_j)$: peluang sebuah kata i masuk kedalam kategori j

$P(c_j)$: peluang kemunculan sebuah kategori j

$P(w_i)$: peluang kemunculan sebuah kata

Peluang kemunculan sebuah kata bisa dihilangkan pada perhitungan klasifikasi karena peluang kemunculan kata tidak akan berpengaruh pada perbandingan hasil klasifikasi setiap kategori. Proses klasifikasi dapat disederhanakan sebagai berikut :

$$P(c_j|w_i) = P(w_i|c_j) \times P(c_j)$$

Untuk menghitung prior atau peluang kemunculan suatu kategori pada semua dokumen dapat dilakukan dengan menggunakan persamaan :

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

N_c : banyak kategori c pada dokumen latih

N : banyak keseluruhan dokumen yang digunakan

Tabel 1. *Confusion matrix*

<i>Class</i>	<i>Positive</i>	<i>Negative</i>
<i>Positive</i>	<i>True positive (TP)</i>	<i>False negative (FN)</i>
<i>Negative</i>	<i>False positive (FP)</i>	<i>True negative (TN)</i>

2.3. Confusion Matrix. *Confusion matrix* adalah *matrix* 2x2 yang merepresentasikan hasil klasifikasi biner pada suatu dataset. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai *accuracy*, *precision* dan *recall* bisa ditampilkan dalam persentase.

a. *Accuracy*. *Accuracy* adalah jumlah proporsi prediksi yang benar. Adapun rumus perhitungan akurasi dapat dilihat dari persamaan di bawah ini [5].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

b. *Precision*. *Precision* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks yang terpilih oleh sistem. Rumus *precision* dapat dilihat pada persamaan di bawah ini [5]

$$Precision = \frac{TP}{TP + FP}$$

c. *Recall*. *Recall* adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi. Rumus *recall* [5] dapat dinyatakan dengan

$$Recall = \frac{TP}{TP + FN}$$

2.4. Analisis Sentimen. Analisis sentimen atau sering disebut dengan istilah *opinion mining* merupakan gabungan dari beberapa teknik, meliputi *natural language processing* (NLP), *information retrieval* (IR), dan *data mining* (DM) yang mengolah atau menganalisis opini, sentimen, dan emosi yang diekspresikan dalam bentuk teks pada suatu entitas. Pada analisis sentimen, model klasifikasi bisa digunakan untuk menentukan sentimen ke dalam dua kelas atau lebih. *Task sentiment analysis* bisa dilakukan menggunakan pendekatan teknik *machine learning* ataupun pendekatan berbasis *lexicon* [3].

2.5. Metode Penelitian. Dataset yang digunakan dalam penelitian ini adalah *public timeline tweet* bahasa Indonesia yang merupakan hasil pencarian berdasarkan *tweet user* yang mengandung kata *quick count*. Dataset didapatkan dengan cara *scraping* memanfaatkan API dari *twitter* kemudian disimpan dalam bentuk *database*. Data yang diambil merupakan data dari tanggal 17 April 2019 sampai tanggal 01 Mei 2019.

3. Hasil dan Pembahasan

Pada pembahasan ini diuraikan tentang tahapan *preprocessing* data, laporan klasifikasi analisis sentimen dari *quick count*.

3.1. Preprocessing data. Data tes merupakan data yang akan diklasifikasikan ke dalam klasifikasi positif atau negatif. Data tes didapat dari proses *crawling twitter* pada tanggal 17 April 2019 dimana bertepatan dengan pelaksanaan pemilihan umum serentak yang dilaksanakan di Indonesia hingga tanggal 01 Mei 2019. Dataset yang didapatkan memuat nama penulis, jumlah favorit, *retweet*, *hashtag*, teks. Data yang digunakan untuk analisis sentimen hanya dataset dengan kolom berisi teks saja. Data tes merupakan data mentah yang masih tidak terstruktur dari hasil penambangan. Oleh karena itu, data tes mentah perlu dilakukan proses *preprocessing* sebelum dianalisis sentimen. Data tes yang awalnya 2000 data berkurang menjadi 1322 data karena terindikasi sama atau hanya merupakan *retweet* dari pengguna lain.

Selanjutnya, adalah tahapan *case folding* yaitu mengubah semua huruf dalam data menjadi *lowercase* (huruf kecil) semua, hal ini dikarenakan analisis sentimen tidak melihat besar kecilnya huruf, sebagai contoh kata ‘Hitung’ dan ‘hitung’ dianggap sama. Kemudian *tokenizing*, yaitu tahap dimana setiap kalimat yang ada didalam data tes dipisah menjadi setiap kata. Tahapan selanjutnya adalah *stopword* atau mengubah kata yang tidak penting, dalam hal ini kata-kata yang tidak mengandung sentimen dengan contoh kata hubung. Terakhir adalah mengubah semua kata yang ada di dalam data tes ke dalam kata dasar (*root word*), tahapan ini disebut juga sebagai proses *stemming*. Setelah semua proses *preprocessing* selesai, maka data tes siap untuk diproses dalam analisis sentimen.

Tabel 2. Frekuensi Kemunculan Kata pada Dokumen

Kata	Frekuensi
twitter	164
com	155
curang	112
pic	100
kpu	89
rakyat	89
aja	64
salah	62
menang	61

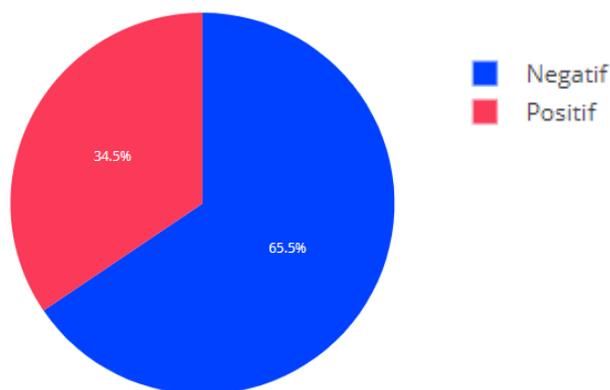
3.2. Frekuensi Kata. Frekuensi dari kata yang sering muncul sangat penting untuk digunakan sebagai pembobot dalam analisis sentimen. Sepuluh kata yang sering muncul

3.4. Laporan Klasifikasi. Laporan klasifikasi diperoleh dari *confusion matrix* yang memberikan informasi lebih rinci tentang kinerja model. *Output* laporan klasifikasi dari program *python* dijelaskan dalam Tabel 4.

Presisi: 80% untuk label 0 dan 98% untuk label 1. Angka ini dapat diartikan bahwa proporsi label yang diprediksi dengan benar dari total prediksi untuk kelas tersebut.

f1-score: rata-rata tertimbang dari presisi dan daya ingat untuk kelas itu. Secara umum memberikan gambaran yang lebih besar tentang bagaimana kinerja model untuk label itu dan jelas semakin tinggi angka ini semakin baik. 89% untuk label 0 dan 63% untuk label 1. 0 menjadi sentimen negatif dan 1 menjadi sentimen positif.

3.5. Analisis sentimen. Dari *output* diperoleh tingkat akurasi terbaik sebesar 82,90% dengan $\alpha = 0,05$ yang berarti bahwa data dapat diklasifikasikan kedalam positif ataupun negatif secara benar sebanyak 82,90% dengan tingkat signifikansi 0,05. Kemudian didapatkan data sebanyak 34,5% (471) *tweet* positif dan 65,5% (895) *tweet* negatif terhadap hasil *quick count*.



Gambar 3. Analisis sentimen

4. Kesimpulan

Model *naive Bayes* terbaik yang diperoleh memiliki tingkat akurasi sebesar 82,90% dengan $\alpha = 0,05$. Klasifikasi yang diperoleh masing-masing sebesar 34,5% (471) *tweet* positif dan 65,5% (895) *tweet* negatif terhadap hasil *quick count*.

Daftar Pustaka

- [1] Sumodiningrat, G., Agustian, A.G. *Mencintai Bangsa dan Negara Pegangan dalam Hidup Berbangsa dan Bernegara di Indonesia*. PT. Sarana Komunikasi Utama. Bogor. 2008.

-
- [2] Kiprono, K. W., and Abade, E. O. Comparative Twitter Sentiment Analysis Based on Linear and Probabilistic Models. *International Journal on Data Science and Technology*. 2(4): 41-45. 2016.
 - [3] Ravi, V., and Ravi, K. A survey on opinion mining and Sentiment Analysis: Tasks, approaches and applications. *Elsevier Knowledge-Based Systems*. 14-46. 2015.
 - [4] Destuardi dan Surya, S. *Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naive Bayes*. Surabaya: Institut Teknologi Sepuluh November. 2009.
 - [5] Manning, C., Raghavan, P. & Schütze, H., *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press. 2009.