

Analisis Sentimen Masyarakat di Twitter Terkait Pandemi Covid-19

Yessi Yunitasari¹, Andi Rahman Putera²

¹Universitas PGRI Madiun

¹yessi@unipma.ac.id, ²andirahmanputera@unipma.ac.id

ABSTRAK

Media sosial merupakan tempat untuk mencari pertemanan baru serta tempat untuk mengeluarkan pendapat terhadap sesuatu secara bebas. Salah satu media sosial yang banyak digunakan saat ini adalah Twitter. Banyak masyarakat yang memanfaatkan Twitter untuk mengeluarkan pendapat terhadap Pandemi Covid-19 yang terjadi di berbagai negara termasuk di Indonesia. Pandemi Covid-19 atau korona virus di Indonesia diawali dengan temuan penderita penyakit korona virus 2019 (COVID-19) pada 2 Maret 2020 hingga 8 April 2020, telah terkonfirmasi 2.738 kasus positif COVID-19, dengan 221 kasus di antaranya meninggal dan 204 kasus sembuh. Tweet yang dituliskan oleh masyarakat nantinya dapat diklasifikasikan ke dalam sentimen positif maupun sentiment negatif menggunakan analisis sentimen. Hasil dari analisis sentiment kita dapat melihat bagaimana persepsi masyarakat Indonesia terkait pandemi Covid-19 yang terjadi di Indonesia. Proses klasifikasi analisis sentimen akan menggunakan metode naïve bayes. Pengujian sentimen yang dilakukan menggunakan *Cross Validation* meliputi pengujian 5 *Fold* dan 10 *Fold*. Dari masing-masing pengujian tersebut akan dilihat nilai *accuracy*, *presisi* dan *recall*. Hasil *Cross Validation* pengujian 5 *Fold* didapatkan hasil dari rata-rata akurasi sebesar 0.756364 (75%). Hasil *Cross Validation* pengujian 10 *Fold* didapatkan hasil dari rata-rata akurasi sebesar 0.76 (76%).

Kata Kunci: Analisis sentimen, Covid-19, Naïve Bayes, Klasifikasi

ABSTRACT

Social media is a place to look for new friends and a place to express opinions about something freely. One social media that is widely used today is Twitter. Many people use Twitter to issue opinions on the Covid-19 Pandemic that occurred in various countries, including in Indonesia. The Covid-19 pandemic or coronavirus in Indonesia begins with the discovery of a 2019 coronavirus sufferer (COVID-19) on March 2, 2020, to April 8, 2020, 2,738 positive cases of COVID-19 have been confirmed, with 221 cases of which have died and 204 cases have recovered. Tweets written by the public can later be classified into positive and negative sentiments using sentiment analysis. The results of our sentiment analysis can see how the perceptions of the Indonesian people regarding the Covid-19 pandemic that occurred in Indonesia. The sentiment analysis classification process will use the naïve Bayes method. Sentiment testing carried out using Cross-Validation includes 5 Fold and 10 Fold testing. From each of these tests, the accuracy, precision, and recall values will be seen. the results of the Cross Validation 5 Fold test obtained results from an average accuracy of 0.756364 (75%). The results of the Cross Validation test for 10 Fold obtained results from an average accuracy of 0.76 (76%)

Keywords: Sentiment analysis, Covid-19, Naïve Bayes, Classification

1. PENDAHULUAN

Media sosial merupakan tempat untuk mencari pertemanan baru serta tempat untuk mengeluarkan pendapat terhadap sesuatu secara bebas. Setiap tahunnya, pengguna media sosial terus meningkat [1]. Salah satu media sosial yang banyak digunakan saat ini adalah Twitter. Twitter memungkinkan penggunaannya untuk menulis dan membaca pesan yang biasa disebut dengan tweet. Jutaan tweet ditulis oleh lebih dari 285 juta pengguna aktif setiap harinya [2].

Banyak masyarakat yang memanfaatkan Twitter untuk mengeluarkan pendapat terhadap Pandemi Covid-19 yang terjadi di berbagai negara termasuk di Indonesia. Pandemi korona virus 2019-2020 adalah pandemi yang sedang berlangsung pada tahun 2019 (COVID-19), yang disebabkan oleh sindrom pernapasan akut coronavirus. Wabah

dimulai di Wuhan , provinsi Hubei , Cina, pada Desember 2019. Organisasi Kesehatan Dunia (WHO) menyatakan wabah itu sebagai Kesehatan Masyarakat Darurat Kepedulian Internasional pada 30 Januari 2020 dan mengakui itu sebagai pandemi pada 11 Maret 2020. Pada 8 April 2020, sekitar 1,42 juta kasus COVID-19 telah dilaporkan di 209 negara dan wilayah, yang mengakibatkan sekitar 82.000 kematian . Sekitar 300.000 orang telah pulih[3].

Pandemi Covid-19 atau korona virus di Indonesia diawali dengan temuan penderita penyakit korona virus 2019 (COVID-19) pada 2 Maret 2020 hingga 8 April 2020, telah terkonfirmasi 2.738 kasus positif COVID-19, dengan 221 kasus di antaranya meninggal dan 204 kasus sembuh. Kasus ini dinyatakan tersebar di 32 provinsi dari total 34 provinsi di Indonesia[3].

Banyak hal yang didiskusikan masyarakat terkait pandemi Covid-19 diantaranya mengenai *lockdown*, di rumah saja, koordinasi presiden dengan Menkes, peran tenaga medis dan lain sebagainya. Selain itu di twitter juga banyak *hashtag-hashtag* terkait pandemi Covid-19 diantaranya “#LawanCovid19, #IndonesiaMelawanCovid19, #dirumahaja, #COVID19, #coronavirus, #CoronaIndonesia, #COVID19Indonesia, dan lain-lain”. Tweet yang dituliskan oleh masyarakat nantinya dapat diklasifikasikan ke dalam sentimen positif maupun sentiment negatif menggunakan analisis sentiment.

Analisis sentimen merupakan salah satu cabang penelitian dari *text minning* yang melakukan proses klasifikasi pada dokumen teks. Data tweet dapat diklasifikasikan berdasarkan sentimennya, yaitu sentimen positif dan negatif. Beberapa metode untuk analisis sentimen telah digunakan pada penelitian sebelumnya, diantaranya Naïve bayes, SVM (*Support Vector Machine*), dan KNN (*K-Nearest Neighbor*)

Penelitian Habibi (2017) melakukan analisis sentimen dan klasifikasi komentar mahasiswa pada sistem evaluasi pembelajaran. Metode yang digunakan dalam penelitian ini adalah metode K-NN berbasis *Cosine Similarity*. Pengujian kombinasi KNN berbasis *Cosine Similarity* dan *Naïve Bayes Classifier* menghasilkan akurasi sebesar 68.33% sedangkan kombinasi KNN berbasis *Cosine Similarity* dan *Support Vector Machine* menghasilkan akurasi sebesar 75.83%

Menganalisis opini masyarakat yang berada di sosial media bukanlah hal yang mudah, opini yang berada di sosial media kebanyakan dituliskan dalam kata yang tidak baku [4]. Berdasarkan paparan di atas, penelitian ini akan membahas bagaimana melakukan Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19.

2. METODOLOGI PENELITIAN

Adapun alur penelitian yang digunakan dalam penelitian ini terbagi menjadi tujuh tahapan diantaranya sebagai berikut :

Identifikasi permasalahan

Pada tahap ini terdapat kegiatan observasi dari permasalahan yang ada melalui media sosial. Peneliti melakukan observasi tepatnya menggunakan twitter.

Penelusuran Pustaka / Pencarian Literatur

Pada tahap ini terdapat kegiatan pencarian literatur berupa buku, artikel ilmiah, jurnal dan hasil seminar yang berkaitan dengan penelitian yang akan dibahas mengenai Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19.

Perumusan masalah

Pada tahap ini terdapat kegiatan menguraikan permasalahan yang bersifat umum ke masalah yang lebih khusus (spesifik). Pada penelitian ini didapat rumusan masalah berupa “Bagaimana melakukan Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19?”

Membuat Desain penelitian

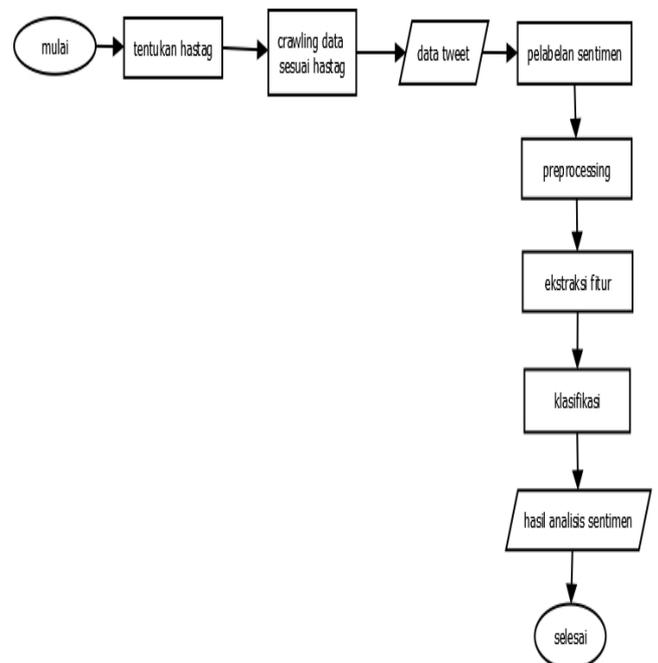
Pada tahap ini terdapat kegiatan menentukan alur penelitan serta metodologi yang digunakan untuk melakukan penelitian.

Pengumpulan Data

Tahapan pengumpulan data dilakukan dengan *crawling* data di twitter terkait penelitian. *Crawling* data dilakukan dengan penggunaan hashtag #Covid-19. Data yang dikumpulkan adalah tweet yang mengandung *hashtag* covid-19.

Pengolahan data

Pada tahap ini terdapat kegiatan analisa data & perancangan sistem. Perancangan sistem Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19. Metode kalsifikasi yang digunakan dalam penelitian ini menggunakan Naïve Bayes. Pada tahap ini terdapat kegiatan menyimpulkan hasil pengolahan data, dan dari hasil kesimpulan tersebut dapat diambil saran serta kebijakan untuk penelitian lanjutan berikut adalah alur penelitian pengolahan data secara keseluruhan pada Gambar 1.



Gambar 1. Alur Penelitian

3. HASIL DAN PEMBAHASAN

Proses pendeteksian sentimen yang dilakukan oleh sistem terdiri dari proses preprocessing, ekstraksi fitur, dan klasifikasi.

Data Penelitian

Tahapan yang harus dilakukan sebelum pembagian data adalah melakukan pengambilan data tweet. Data tweet diambil dari twitter.com dengan menggunakan API tweepy. Tahapan pertama yang dilakukan untuk pengambilan data tweet adalah mendapatkan API *access*. API *access* terdiri dari *access* token, *access* token *secret*, *consumer key*, dan *consumer secret*. Kemudian 4 informasi tersebut dimasukkan dalam *script* untuk proses *crawling* data. Data diambil pada bulan April-Agustus 2020. Pengujian dilakukan dengan *K-fold*. Dalam sistem ini akan digunakan 5 *fold* dan 10 *fold*.

Pelabelan

Data tweet yang telah dilakukan proses *crawling* akan dilabeli. Data yang telah dilakukan proses *crawling*, selanjutnya dilabeli dengan 2 label yaitu positif dan negatif. *Emoticon* juga diberikan label positif atau negatif dan label sarkasme atau tidak sarkasme. Selain itu *hashtag* juga perlu diberikan label positif dan negatif.

Preprocessing Sentimen

Preprocessing data merupakan proses dimana teks dibersihkan untuk menghilangkan *noise* pada data. *Preprocessing* dilakukan untuk menghindari data yang kurang sempurna, gangguan pada data, dan data-data yang tidak konsisten[5]. Adapun tahapan *preprocessing* untuk analisis sentimen yang dilakukan dalam penelitian ini meliputi:

Casefolding

Menyeragamkan semua huruf menjadi huruf kecil.

Stopword

Menghilangkan kata yang tidak sesuai pada hasil parsing.

Konversi emoticon

Kumpulan tweets yang terdapat emoticon akan dikonversikan ke dalam string yang bersesuaian.

Konversi karakter angka ke huruf

Misalnya kata m4k4n menjadi makan.

Slang word

Proses ini mengganti kata-kata dalam tweet yang tidak baku menjadi kata baku yang telah dikenal dalam bahasa Indonesia.

Stemming

Mengubah kata berimbuhan menjadi kata dasar. (*stemming* menggunakan Sastrawi *stemming*). Hasil *preprocessing* dari sistem dapat dilihat pada Gambar 1. Contoh tweet yang digunakan “Mari kita

cegah virus Covid-19 dengan Dirumah Saja <https://t.co/mgazQ0YCU3>”.

```
PREPROCESSING SENTIMEN=====
Dibuat huruf kecil semua : mari kita cegah virus covid-19 dengan dirumah saja https://t.co/mgazq0ycu3
Emoji diganti kata : mari kita cegah virus covid-19 dengan dirumah saja https://t.co/mgazq0ycu3
Link dihilangkan : mari kita cegah virus covid-19 dengan dirumah saja
Tanda baca dihilangkan : mari kita cegah virus covid-19 dengan dirumah saja
Slang dibakukan : mari cegah virus covid-19 dirumah
Angka jadi huruf : mari cegah virus covid-ig dirumah
Stopwords dihilangkan : mari cegah virus covid-ig dirumah
Stemming : mari cegah virus covid-ig rumah
```

Gambar 2. Hasil Preprocessing dari Sistem

Ekstraksi Fitur

Proses ekstraksi fitur digunakan untuk menemukan fitur-fitur yang ada pada sebuah *text*, ekstraksi fitur yang digunakan pada analisis sentimen adalah unigram dan TF-IDF.

Metric yang umum dipergunakan dalam proses kategorisasi teks adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). TF-IDF terdiri dari dua buah nilai komponen. Komponen yg pertama adalah *term-frequency* dan komponen ke 2 adalah *inverse document frequency*. Skema pembobotan TF-IDF adalah pemberian bobot pada *term t* pada dokumen *d* seperti pada persamaan (1) [6].

$$tf.idf_{t,d} = tf_{t,d} \times idf_t \quad (1)$$

$$idf_t = \text{Log} (N/df_t) \quad (2)$$

Keterangan:

$tf.idf_{t,d}$: Bobot tiap kata dari term *t* yang terdapat dari dokumen *d*

$tf_{t,d}$: Bobot term *t* yang terdapat pada dokumen *d*

idf_t : Inverse document frequency yang terdapat pada term *t*

N : Jumlah semua dokumen

df_t : Banyak dokumen yang memuat term *t* didalamnya

t : Term atau kata

Hasil dari ekstraksi fitur dapat dilihat pada Gambar 2. Setiap kata memiliki nilai *tfidf*.

```
=====
BERDASARKAN HASIL TRAINING DIDAPATKAN
Kata 'cegah'' memiliki nilai tfidf : 0.6542013382394214
Kata 'mari'' memiliki nilai tfidf : 0.5730616538140181
Kata 'rumah'' memiliki nilai tfidf : 0.43991994345393637
=====
```

Gambar 3. Hasil Ekstraksi Fitur dari Sistem

Klasifikasi

Naïve Bayes Classifier merupakan salah satu metode *machine learning* menggunakan probabilitas yang dapat digunakan untuk klasifikasi suatu dokumen [7]. *Naïve Bayes Classifier* juga merupakan sebuah metode klasifikasi yang berakar pada teorema bayes. Ciri utama *Naïve Bayes Classifier* adalah asumsi yang sangat kuat (naïf) dari independensi masing-masing kondisi atau kejadian. Secara sederhana, NBC mengasumsikan bahwa ada atau tidaknya fitur tertentu dari suatu kelas tidak memiliki keterkaitan dengan keberadaan fitur lainnya. Keuntungan penggunaan NBC adalah hanya diperlukan sejumlah kecil data pelatihan untuk mengestimasi parameter yaitu rata-rata dan varian dari variabel yang diperlukan untuk klasifikasi[8].

Proses klasifikasi menggunakan *Naïve Bayes Classifier* diawali dengan menghitung probabilitas setiap kata dalam komentar. Setelah itu dilakukan perhitungan nilai probabilitas prior yaitu nilai kemungkinan munculnya kelas tertentu dalam sekumpulan data [9].

Hasil dari proses klasifikasi menggunakan *Naïve Bayes Classifier* dapat dilihat pada Gambar 3. Proses klasifikasi diawali dengan menghitung probabilitas setiap kata dalam tweet. Setiap kata dihitung probabilitas positif dan probabilitas negatifnya.

```

=====
BERDASARKAN HASIL TRAINING & NILAI TFIDF SEBELUMNYA MAKA DIDAPKAN
probability p('cegah'|positif) adalah : 0.000109282676676
probability p('cegah'|negatif) adalah : 8.57312493118e-05
probability p('mari'|positif) adalah : 0.000201340906203
probability p('mari'|negatif) adalah : 0.000130473653746
probability p('rumah'|positif) adalah : 0.0004032926367
probability p('rumah'|negatif) adalah : 0.000465728758463
=====
    
```

Gambar 4. Perhitungan Probabilitas Positif dan Probabilitas Negatif

```

Aveardo Proxpt (Aveardo2)
SEHINGGA RULE-NYA ADALAH

p(positif|(cegah,mari,rumah)) =
p(positif)
* p('cegah'|positif)
* p('mari'|positif)
* p('rumah'|positif)

p(positif|(cegah,mari,rumah)) = 0.72 * 0.000109282676676 * 0.000201340906203 * 0.0004032926367
perhitungan probability positif:6.38904771929e-12

p(negatif|(cegah,mari,rumah)) =
p(negatif)
* p('cegah'|negatif)
* p('mari'|negatif)
* p('rumah'|negatif)

p(negatif|(cegah,mari,rumah)) = 0.28 * 8.57312493118e-05 * 0.000130473653746 * 0.000465728758463
perhitungan probability negatif:1.45865661013e-12

SENTIMEN : Probability lebih besar yang positif
=====
    
```

Gambar 5. Hasil Klasifikasi Sentimen Tweet

Selanjutnya hasil dari probabilitas positif dan negatif dibandingkan untuk tweet “Mari kita cegah virus Covid-19 dengan Dirumah Saja <https://t.co/mgazQ0YCU3>”. Hasil perhitungan dari sistem menunjukkan bahwa tweet tersebut merupakan sentiment positif. Hasil dari proses tersebut klasifikasi sentimen tweet dapat dilihat pada Gambar 5.

Pengujian Data

Pengujian sentimen yang dilakukan menggunakan *Cross Validation* meliputi pengujian *5 Fold* dan *10 Fold*. Dari masing-masing pengujian tersebut akan dilihat nilai *accuracy*, presisi dan *recall*. Berikut merupakan hasil dari pengujian data yang dilakukan untuk Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19. Tabel 1 menunjukkan hasil *Cross Validation* pengujian *5 Fold*. Sedangkan Tabel 2 menunjukkan hasil *Cross Validation* pengujian *10 Fold*.

Tabel 1. *Cross Validation* Pengujian *5 Fold*

SENTIMEN 5 FOLDS			
	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
Fold1	0.727273	0.727273	1
Fold2	0.727273	0.727273	1
Fold3	0.727273	0.727273	1
Fold4	0.8	0.8	1
Fold5	0.8	0.8	1
AVG	0.756364	0.756364	1

Tabel 1 yaitu hasil *Cross Validation* pengujian *5 Fold* didapatkan hasil dari rata-rata akurasi sebesar 0.756364 (75%).

B
erdasar
kan

Tabel 2. *Cross Validation* pengujian 10 Fold SENTIMEN 10 FOLDS

	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>
Fold1	0.666667	0.666667	1
Fold2	0.666667	0.666667	1
Fold3	0.666667	0.666667	1
Fold4	0.8	0.8	1
Fold5	0.8	0.8	1
Fold6	0.8	0.8	1
Fold7	0.8	0.8	1
Fold8	0.8	0.8	1
Fold9	0.8	0.8	1
Fold10	0.8	0.8	1
AVG	0.76	0.76	1

Sedangkan pada *Tabel 2* yaitu hasil *Cross Validation* pengujian 10 Fold didapatkan hasil dari rata-rata akurasi sebesar 0.76 (76%). Berdasarkan 2 perbandingan nilai *fold* yang digunakan hasil akurasi yang lebih tinggi diperoleh ketika nilai *fold* yang digunakan adalah 10 *fold*. Data set yang digunakan dibagi menjadi 10 bagian kemudian dicari hasil dari *fold 1* sampai dengan *fold10*.

4. KESIMPULAN DAN SARAN

Setelah dilakukan implementasi dan pembahasan, maka disimpulkan bahwa hasil dari pengujian sistem Analisis Sentimen Masyarakat Di Twitter Terkait Pandemi Covid-19 yang dilakukan menggunakan *Naive Bayes* menunjukkan akurasi sebesar 75% untuk analisis sentiment dengan 5 fold dan nilai akurasi sebesar 76% untuk analisis sentiment dengan 10 fold. Sistem yang telah dibangun juga dapat mendeteksi apakah suatu tweet tersebut merupakan sentiment negatif atau positif berdasarkan klasifikasi menggunakan metode *naive bayes*.

5. REFERENSI

[1] A. F. Hidayatullah and A. Sn, "Analisis Sentimen dan Klasifikasi Kategori Terhadap Tokoh Publik Pada Twitter," *Semin. Nas. Inform. 2014*, vol. 2014, no. August 2013, pp. 0–8, 2014.

[2] M. Bouazizi and T. Ohtsuki, "Sarcasm detection in twitter: »all your products are incredibly amazing!!!« - are they really?," in *2015 IEEE Global Communications Conference, GLOBECOM 2015*, 2015, doi: 10.1109/GLOCOM.2014.7417640.

[3] M. Arnani, "2.738 Orang Positif Covid-19 di Indonesia, Ini Rincian Kasus di 32 Provinsi," *Kompas*, Jakarta, 2020.

[4] G. A. Buntoro, "Analisis Sentimen Calon Presiden Indonesia 2014 Dengan Lima Class Attribute," Universitas Gadjah Mada, 2015.

[5] A. Faisal, Y. Alkhalifi, A. Rifai, and W. Gata, "Analisis Sentimen Dewan Perwakilan Rakyat Dengan Algoritma Klasifikasi Berbasis Particle Swarm Optimization," *JOINTECS*

(*Journal Inf. Technol. Comput. Sci.*, vol. 5, no. 2, p. 61, 2020, doi: 10.31328/jointecs.v5i2.1362.

[6] C. D. Manning, P. Ragahvan, and H. Schutze, "An Introduction to Information Retrieval," *Inf. Retr. Boston.*, no. c, pp. 1–18, 2009, doi: 10.1109/LPT.2009.2020494.

[7] T. A. M, Y. Alkhalifi, N. A. Mayangky, and W. Gata, "Analisis Sentimen Opini Publik Mengenai Larangan Mudik pada Twitter Menggunakan Naive Bayes," *CoreIT*, vol. 6, no. 2. 2020.

[8] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and D. R. Hadinata, "Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012085.

[9] M. Habibi, "Analisis Sentimen Dan Klasifikasi Komentar Mahasiswa Pada Sistem Evaluasi Pembelajaran Menggunakan Kombinasi Knn Berbasis Cosine Similarity Dan Supervised Model," Universitas Gadjah Mada, 2017.