ANALISIS SENTIMENT PADA TWITTER DENGAN MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER

Sigit Suryono[1], Ema Utami[2], Emha Taufiq Luthfi[3]
Mahasiswa Magister Teknik Informatika Universitas Amikom Yogyakarta[1], Dosen, Universitas Amikom Yogyakarta[2,3]
sigitharsy25@gmail.com, emma@nrar.net, emhataufiqluthfi@amikom.ac.id

ABSTRAK

Analisis Sentiment merupakan salah satu cabang dari bidang ilmu Text Mining. Analisis sentiment merupakan sumber penting dalam melakukan evaluasi dan pengambilan keputusan terhadap sebuah topik permasalahan. Tujuan utama dari analisis sentiment adalah untuk mengetahui polaritas dari sentiment positif, negatif ataupun netral. Sentiment-sentiment tersebut salah satunya didapatkan dari Twitter. Dalam tulisan ini, tweet-tweet yang berhubungan dengan kata kunci yang dicari dikumpulkan dari Twitter dengan menggunakan API Twitter dan data mentah yang didapatkan diolah dengan menggunakan Natural Language Toolkit pada bahasa pemrograman Python. Setelah diolah selanjutnya akan dilakukan klasifikasi dengan menggunakan Naïve Bayes Classifier untuk mengetahui tingkat akurasi dari proses klasifikasi yang dilakukan. Proses klasifikasi dilakukan dengan RapidMiner. Dari hasil uji coba sebanyak empat kali, didapatkan hasil tingkat akurasi pada percobaan pertama sebesar 62.98%, percobaan kedua sebesar 64.95%, percobaan ketiga sebesar 66.36%, dan percobaan keempat sebesar 66.79%. Dari hasil klasifikasi didapat tingkat persentase sentiment positif sebesar 28%, sentiment negatif sebesar 20% dan sentiment netral sebesar 52%.

Kata kunci *: Analisis Sentiment, Naive Bayes Classifier, Klasifikasi, Natural Language Toolkit, Opinion Mining*

ABSTRACT

Sentiment Analysis is the one branch of the field of Text Mining. Sentiment Analysis is an important source for evaluating and making decisions on a topic of concern. The main purpose of the sentiment analysis is to know the polarity of positive, negative and neutral sentiment. The sentiments are one of them obtained from Twitter. In this paper, tweets that related to the searched keyword are collected from Twitter using Twitter API and the raw data obtained is processed using Natural Language Toolkit in the Python Programming. Once processed then the data will be classified by using Naïve Bayes Classifier to determine the level of accuracy of the classification process undertaken. The process of classification is done by RapidMiner. Based on the experimental result for four times trial, the result obtained accuracy level in the first is 62.98%, the second is 64.95%, the third is 66.36%, and the fourth is 66.79%. Based on the classification result, the result obtained for percentage of sentiment are positive sentiment is 28%, negative sentiment is 20% and neutral sentiment is 52%.

***Keyword**: Sentiment Analysis, Naïve Bayes Classifier, Classification, Natural Language Toolkit, Opinion Mining*

PENDAHULUAN

Analisis sentiment merupakan salah satu cara untuk mengumpulkan pendapat orang banyak terhadap sesuatu seperti layanan public, isu, kinerja pemerintahan atau hal lain yang

berkaitan. Analisis sentiment dapat digunakan sebagai salah satu cara untuk melakukan evaluasi terhadap layanan yang telah diberikan. Analisis sentiment dapat dilakukan melalui berbagai cara salah satunya adalah dengan mengumpulkan pendapat orang banyak melalui media social.

Media social di Indonesia tidak hanya digunakan sebagai media untuk mengungkapkan apa yang sedang dirasakan tetapi juga digunakan sebagai media untuk memberikan saran atau kritik terhadap layanan yang telah diberikan oleh pemerintah. Melalui media social khususnya Twitter banyak orang yang memuji, menyalahkan atau tidak keduanya terhadap apa yang dilakukan oleh pemerintahan sekarang dalam hal ini pemerintahan preside Joko Widodo.

Tweet yang berkaitan dengan pemerintahan presiden Joko Widodo akan dikumpulan lalu dilakukan analisis sentiment dengan menggunakan Naïve Bayes Classifier. Pengumpulan data dilakukan dengan menggunakan API yang telah disediakan oleh Twitter. Pengumpulan data dengan menggunakan API Twitter diimplementasikan kedalam bahasa pemrograman Python. Pelabelan data tweet yang telah didapatkan dilakukan dengan menggunakan sentistrengths yang diimplemetasikan dengan menggunakan bahasa pemrograman Python. Sebanyak 40% data yang telah diberikan label nantinya akan digunakan sebagai dasar untuk menentukan sentiment terhadap tweet yang belum diberikan label dengan menggunakan metode Naïve Bayes Classifier. Setelah itu akan diukur tingkat akurasi metode Naïve Bayes Classifier dengan menggunakan RapidMiner serta menghitung persentase untuk masin-masing sentiment berdasarkan tweet yang telah didapatkan sebelumnya.

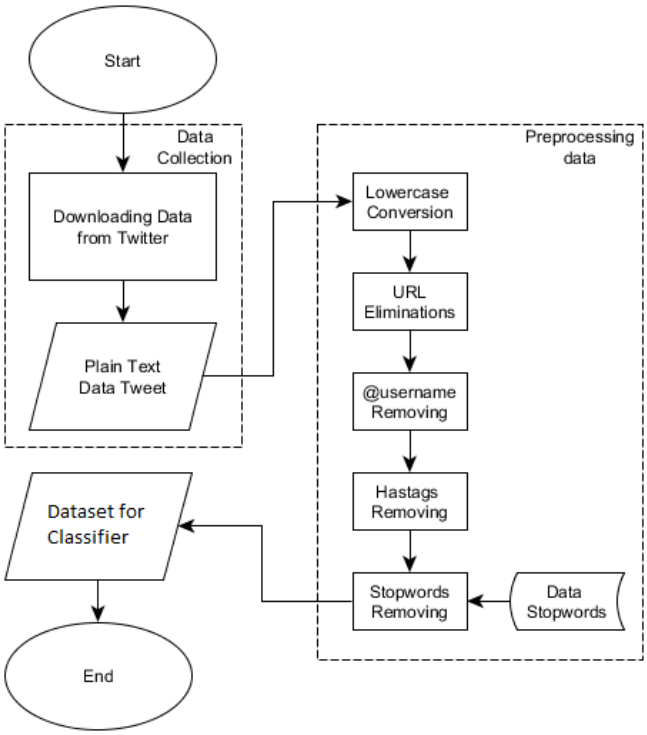
Ramadhan WP, Astri Novianty, Casi Setianingsih melakukan sebuah analisis sentiment dengan menggunakan Support Vector Machine dan Maximum Entropy [1]. Dalam penelitian ini analisis sentiment dilakukan terhadap masing-masing calon dalam pikada DKI. Penelitian bertujuan untuk mencari tahu metode serta kernel apa yang memilki tingkat akurasi yang baik. Hasilnya metode Support Vector Machine dan kernel linear yang memilki tingkat akurasi yang tinggi. Dalam penelitian yang dilakukan oleh Ramadhan WP, Astri Novianty, Casi Setianingsih tidak memberikan gambaran berapa hasil persentase untuk masing-masing sentiment yang ada.

Fiktor Imanuel Tanesab, Irwan Sembiring and Hindriyanto Dwi Purnomo melakukan analisis sentiment terhadap komentar yang ada pada Youtube dengan menggunakan metode Support Vector Machine [2]. Komentar yang dikumpulkan adalah komentar tentang mantan gubernur DKI Jakarta Basuki Tjahaya Purnama. Penelitian ini bertujuan untuk mencari tingkat akurasi, presisi, recall, true positive dan true negative dari metode yang diajukan. Hasilnya akurasi 84%, presisi 91%, recall 80%, true positive 91.1% dan true negative 44.8%. penelitian yang dilakukan oleh Fiktor Imanuel Tanesab, Irwan Sembiring and Hindriyanto Dwi Purnomo belum memberikan gambaran berapa hasil persentase untuk masing-masing sentiment berdasarkan komentar yang telah diberikan sentiment sebelumnya.

Dalam tulisan ini, peneliti akan melakukan analisis sentiment pada Twitter dengan menggunakan metode Naïve Bayes Classifier terhadap tweet-tweet yang berkaitan dengan pemerintahan presiden Joko Widodo. Hasil dari penelitian ini adalah tingkat akurasi dari metode yang digunakan serta persentase untuk masin-masing sentiment berdasarkan tweet-tweet yang telah didapat sebelumnya.

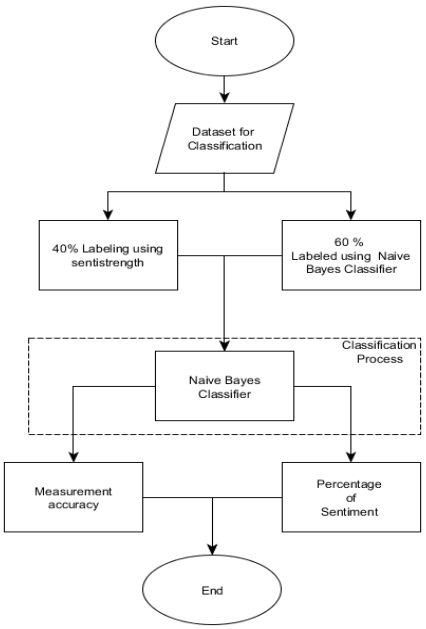
METODE

Penelitian ini dilakukan dengan membagi data menjadi 2 kelompok data yaitu data training dan data testing. Data yang digunakan pada penelitian ini sebanyak 3485 data. Data-data tersebut akan melalui beberapa proses sehingga menghasilkan data yang siap digunakan untuk melakukan klasifikasil. Proses pengumpulan, preprocessing data hingga menjadi data siap pakai akan ditunjukkan oleh gambar 1 berikut ini.



Gambar 1. Proses pengumpulan dan preprocessing data

Pada gambar 1 dapat dijelaskan proses pertama yang dilakukan adalah pengumpulan data. Pengumpulan data dilakukan dengan cara mengunduh data tweet dari Twitter dan menghasilkan data mentah. Data mentah yang telah didapatkan akan melalui proses preprocessing data. Dalam preprocessing data terdapat lima sub proses yang dilalui yaitu *lowercase conversion*, *URL eliminations*, *@username removing*, *hashtag removing* dan *stop words removing*. Setelah melalui proses preprocessing data maka akan menghasilkan dataset yang siap digunakan untuk melakukan klasifikasi dengan menggunakan Naïve Bayes Classifier. Proses klasifikasi dengan Naïve Bayes Classifier akan ditunjukkan pada gambar 2 berikut ini.



Gambar 2. Proses Naïve Bayes Classifer

Pada gambar 2 dapat dijelaskan bahwa data yang telah siap diolah (diklasifikasi) akan dilakukan pelabelan data dengan menggunakan sentistrength. Data yang dilabeli diambil sebesar 40% dari dataset. setelah dilakukan pelabelan, data yang telah dilabeli akan dijadikan acuan untuk menentukan klasifikasi sentiment untuk data tweet yang belum dilabeli dengan menggunakan Naïve Bayes Classifier. Setelah semua data memiliki label, selanjutnya akan diukur tingkat akurasi dari metode Naïve Bayes Classifier pada aplikasi RapidMiner.

Adapun metode-metode yang digunakan dalam penelitian ini adalah sebagai berikut.

A. Analisis Sentiment

Analisis sentiment atau pada umumnya dikenal dengan *opinion mining* merupakan salah satu cabang studi tentang analisa pendapat ataupun pendapat seseorang sesuatu seperti layanan, produk, oraganisasi, dan lain sebagainya [3]. Analisis sentiment merupakan salaha satu penelitian yang cukup kompleks. Adapaun karakteristik dari analisis sentiment adalah sebagai berikut ini [4].

1. Pengkategorian sentiment yang akan membedakan antara kalimat subjektif dan objektif
2. Tingkatan analisis. Tingkatan analisis dibagi menjadi 3 bagian yaitu *message level, sentence level* dan *aspect level.*
3. Pendapat yang memberikan perbandingan terhadap sesuatu serta pendapat yang hanya sekedar pendapat. Ini memiliki maksud setiap orang dapat memberikan pendapat dengan membandingkan suatu hal dengan hal yang lain atau hanya sekedar memberikan pendapat.
4. Pembagian pendapat menjadi eksplisit dan implisit. Pendapat yang diungkapkan secara jujur, tegas serta lugas dan jelas atau pendapat yang diungkpan secara tidak jelas.

B. Natural Language Toolkit

Natural language toolkit merupakam sebuah tools yang dikembangkan khusus untuk bahasa pemrograman Python dan digunakan dalam proses yang berhubungan pemrosesan bahasa alamiah [5]. Natural language toolkit menyediakan tampilan antar muka yang mudah didapahi dan digunakan serta menyediakan lebih dari 50 data yang dapat digunakan seperti pemrosesan bahasa alamiah seperti WordNet dan libary TextProcessing serta untuk pemrosesan klasifikasi seperti *tokenization, stemming, tagging, parsing* dan *semantic reasoning.*

C. Python

Merupakan bahasa pemrograman yang dapat dijalankan pada berbagai sistem operasi baik Linux, MacOS ataupun Windows dengan melakukan konfigurasi terlebih dahulu [6]. Python merupakan bahasa pemrograman tingkat tinggi dikarenakan kode-kode yang ditulis akan di compile menjadi byte code serta dieksekusi sehingga membuat Python cocok digunakan untuk bahasa pemrograman scripting, aplikasi web dan lain sebagainya.

D. Naive Bayes Classfier

Naive bayes classifier merupakan salah satu metode atau model yang dapat melakukan proses klasifikasi secara baik [7]. Naive bayes melakukan klasifikasi dengan memggunakan dua buah proses yang membagi data menjadi data training serta data testing. Proses klasifikasi dalam Naive Bayes terhadap data dilakukan dengan merepresentasikan setiap data kedalam “X1, X2, X3,…,Xn”. Himpunan kategori direpresentasikan dengan K. Pada saat melakukan klasifikasi, Naive Bayes akan mencari nilai probablitias tertinggi dengan menggunakan rumus sebagai berikut.

$$K_a = argmax P(X_1, X_2, X_3, \dots, X_n).P(K) \quad (1)$$

Keterangan :
K_a : *Semua kategori yang diujikan*
$X_1, X_2, X_3, \dots, X_n$: *setiap kata dalam tweet*
$P(K)$: *probabilitas kategori*

Dalam proses klasifikasi data, rumus yang digunakan dalah sebagai berikut.

$$P(X_i|K) = P(K) \frac{X_1, X_2, X_3, \dots, X_n}{\sum K} \quad (2)$$

Keterangan :
$P(X_i|K)$: *Kategori yang diujikan*
$P(K)$: *probabilitas kategori*
$X_1, X_2, X_3, \dots, X_n$: *setiap kata dalam tweet*
$\sum K$: *jumlah nilai dari setiap kategori*

E. Split Validation

Split validation merupakan sebuah nested operator yang terdapat didalam RapidMiner. Pada Split validation terdapat dua buah sub proses yaitu sub proses training dan sub proses testing yang mana sub proses training digunakan untuk proses learning atau membangun sebuah model dan model yang telah dibuat akan diterapkan didalam sub proses testing juga performa dari model yang dibangun akan diukur didalam fase testing [8].

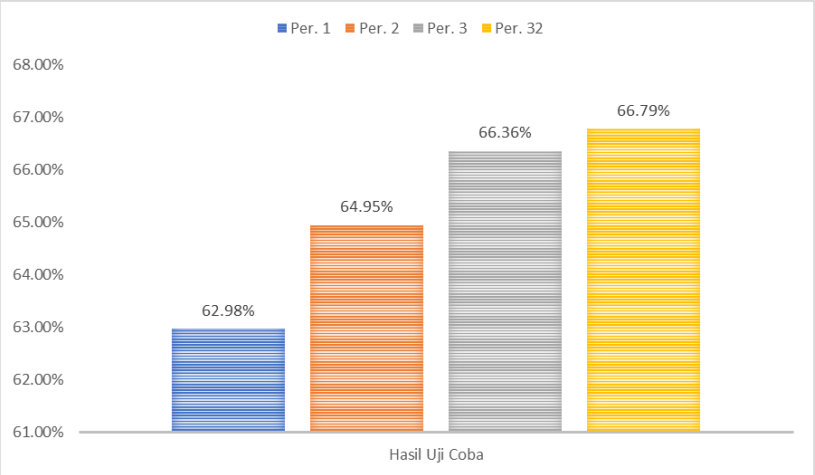
HASIL

Data yang telah didapatkan yaitu sebanyak 3485 tweet dan 40% dari data tersebut telah dilabeli dengan sentistrength dan sisanya dilabeli dengan Naïve Bayes Classifer. Dari data yang telah dilabeli seluruhnya adapun scenario uji coba untuk mengukur tingkat akurasi dari metode Naïve Bayes Classifier akan ditunjukkan pada table 1 berikut ini.

Tabel 1. Skenario Uji Coba

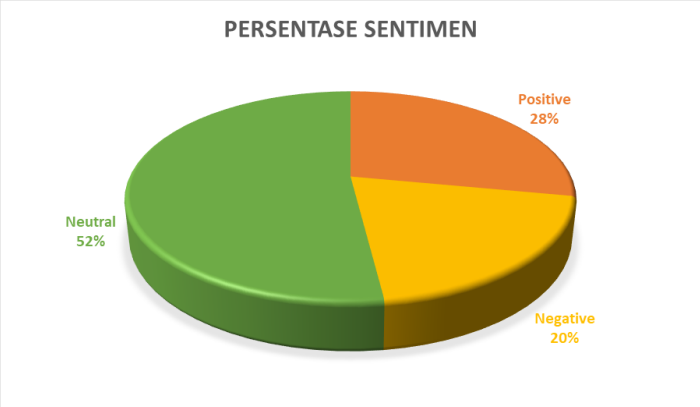
No	Data Set	Number of Trials	Data Training (%)	Data Testing (%)
1	Dataset for Classifier	1	20	80
		2	40	60
		3	50	50
		4	60	40

Berdasarkan hasil percobaan yang telah dilakukan sesuai dengan table 1, adapun hasil tingkat akurasi dari keempat percobaan akan ditunjukkan pada gambar 3 berikut ini.



Gambar 3. Tingkat Akurasi Berdasarkan table Uji Coba

Pada gambar 3 dapat dilihat bahwa tingkat akurasi tertinggi didapat pada percobaan keempat dengan pembagian data 60:40. Pada percobaan keempat tingkat akurasinya sebesar 66.79 %. Tingkat akurasi terendah didapat pada percobaan pertama dengan pembagian data 20:80. Pada percobaan pertama tingkat akurasi yang diperoleh adalah sebesar 62.98%. Berdasarkan hasil uji coba yang telah dilakukan dan hasil yang ditunjukkan, semakin besar data yang ditraining akan semakin meningkatkan akurasi. Hasil ini belum multak dikarenakan uji coba hanya dilakukan sebanyak empat kali. Berdasarkan hasil klasifikasi yang telah dilakukan adapun persentase sentiment dari dataset yang telah diperoleh akan ditunjukkan pada gambar 4 berikut ini.



Gambar 4. Persentase Sentiment

Pada gambar 4 ditujukan bahwa persentase terbesar diperoleh oleh sentiment neutral sebesar 52%. Ditempat kedua diperoleh oleh sentiment positive dengan 28% dan diurutan terakhir diperoleh oleh sentiment negative sebesar 20%. Berdasarkan hasil persentase sentiment, sentiment yang dominan dalam data set yang telah diperoleh adalah sentiment neutral. Ini berarti banyak para pengguna Twitter yang tidak memuji apa yang telah dilakukan oleh presiden Joko Widodo dan tidak juga menyalahkan apa yang telah dilakukan oleh presiden Joko Widodo. Dalam hal ini pengguna Twitter didominasi oleh sentiment neutral.

SIMPULAN

Didasarkan pada hasil uji coba tingkat akurasi klasifikasi yang dilakukan oleh metode Naïve Bayes Classifier memiliki nila tertinggi pada percobaa keempat dengan 66.79% dengan perbandingan pembagian data yaitu 60% untuk data training dan 40% untuk data testing. Tingkat akurasi terendah diperoleh pada percobaan pertama dengan 62.98% dengan perbandingan pembagian data yaitu 20% untuk data training dan 80% untuk data testing. Berdasarkan hasil pecobaan didapatkan fakta bahwa semakin banyak porsi data untuk data training maka tingkat akurasinya juga akan meningkat. Hasil lain yang diperoleh yaitu tingkat persentase sentiment.

Persentase sentiment tertinggi didapat oleh sentiment neutral dengan 52% lalu sentiment positive dengan 28% dan yang terakhir adalah sentiment negative dengan 20%. Berdasarkan hasil persentase sentiment pengguna Twitter yang tidak memuji apa yang telah dilakukan oleh presiden Joko Widodo dan tidak juga menyalahkan apa yang telah dilakukan oleh presiden Joko Widodo. Dalam hal ini pengguna Twitter didominasi oleh sentiment neutral.

Untuk penelitian selanjutnya dapat menambahkan proses pada preprocessing data seperti proses untuk menangani singkatan-singkatan yang ada Tweet. Selain itu, penelitian ini dapat diterapkan terhadap metode pengklasifikasian data yang lainnya.

DAFTAR PUSTAKA

[1] R. Azuma, “A survey of augmented reality,” Presence Teleoperators Virtual Environ., vol. 6, no. 4, pp. 355–385, 1997.

[2] F. I. Tanesab, I. Sembiring, and H. D. Purnomo, "Sentiment Analysis Model Based on Youtube Comment Using Support Vector Machine," International Journal of Computer Science and Software Engineering, 2017.

[3] B. Liu, Sentiment Analysis and Opinion Mining, Toronto: Morgan & Claypool Publishers, 2012.

[4] F. A. Pozzi, E. Fersini, E. Messina and B. Liu, Sentiment Analysis in Social Networks, Cambridge: Todd Green, 2017.

[5] NLTK Project, "www.nltk.org," 02 November 2017. [Online]. Available: http://www.nltk.org/.

[6] C. Manning and H. Schuetze, Foundations of Statistical Natural Language Processing, London: MIT Press, 1999.

[7] F. Romano, D. Phillips and R. v. Hattlen, Python: Journey from Novice to Expert, Birmingham: Packt Publishing Ltd., 2016.

[8] Rapid Miner, Operator Refence Manual, Boston: Rapid Miner Inc., 2014.