KLASIFIKASI SENTIMEN MASYARAKAT PENGGUNA TWITTER MENGGUNAKAN METODE *NAÏVE BAYES* PADA KASUS PEMERINTAHAN DAERAH

Yulia Nur Kumala[1)] , Sekar Jati Cahyaning Wulan[2)], Rara Ayu Puspita[3)], Prizka Rismawati Arum[4)] , Indah Manfaati Nur[5)]
[1]Universitas Muhammadiyah Semarang (Yulia Nur Kumala)
email: yuliakumala04@gmail.com
[2]Universitas Muhammadiyah Semarang (Sekar Jati Cahyaning Wulan)
email: sekarjati29@gmail.com
[3]Universitas Muhammadiyah Semarang (Rara Ayu Puspita)
email: rarapuspita65@gmail.com
[4]Universitas Muhammadiyah Semarang (Prizka Rismawati Arum)
email: prizka.rismawati@gmail.com
[5]Universitas Muhammadiyah Semarang (Indah Manfaati Nur)
email: indahmnur@unimus.ac.id

Abstract

Twitter is a microblogging site that allows its users to write about various opinions, comments, and news that discuss current issues. Many users post their opinion on a product or service they use. It can be used as a source of data to assess sentiment on Twitter. One method of automatic emotional grouping can be used, one of which is using Naïve Bayes. The purpose of this research is to build a system that is able to automatically classify the emotions of each tweet, and to determine the accuracy of the grouping. The process starts from preprocessing, there are several processes, namely tokenizing, stopword, stemming, word weighting, and normalization, which can then be processed using Naïve Bayes. Naïve Baye Process The creation of a sentiment analysis system using the Naïve Bayes method has proven that the algorithm can analyze sentiments automatically, with an accuracy rate of 95%. The results of this visualization can be used by the Government to determine policies to be taken in the future. After that perform accuracy calculations using confusion matrix.

Keywords: *Tweet, Naïve Bayes, government Maksimum*

1. PENDAHULUAN

Era globalisasi saat ini sangat memengaruhi pesatnya kemajuan teknologi informasi seperti dalam bidang ekonomi, kebudayaan, seni, pendidikan dan bahkan dunia politik. Seiring dengan kemajuan teknologi informasi tersebut, jumlah pengguna media social kini berkembang sangat pesat. Media sosial adalah sebuah media online, dengan para penggunanya bisa dengan mudah berpartisipasi, berbagi, dan menciptakan isi meliputi blog, jejaring sosial, wiki, forum dan dunia virtual. Blog, jejaring sosial dan wiki merupakan bentuk media sosial yang paling umum digunakan oleh masyarakat di seluruh dunia. Media sosial adalah sebuah wadah yang mampu menciptakan berbagai bentuk komunikasi dan pemberian berbagai macam informasi bagi semua kalangan masyarakat. Media Sosial yang saat ini ada terdiri dari berbagai

media online seperti situs media jejaring sosial, aplikasi media jejaring sosial, game, dan media online lainnya.

Banyak pemerintah daerah menggunakan media sosial sebagai satu layanan electronic government (*E-Government*) sebagai sarana dalam menyampaikan informasi kepada masyarakat.Mengingat media sosial sebagai alat baru untuk melengkapi layanan *E-Government* yang ada, perlu dipahami jenis layanan *E-Government* yang lebih sesuai dengan alat media sosial yang berbeda. Peran layanan *E-Government* dalam adopsi media sosial di pemerintah daerah sedikit dipahami dan diketahui. Penerapan *E-Government* menjanjikan perubahan paradigma yang tajam. Dengan penerapan *E-Government*, institusi publik akan lebih responsive dan transparan, mempromosikan kemitraan pemerintah lebih efisien, dan memberdayakan warga dengan membuat pengetahuan dan sumber daya lainnya lebih dapat diakses langsung.
Situs microblogging seperti *Twitte*r telah menjadi alat komunikasi yang sangat popular dikalangan pengguna internet di Indonesia. Pada tahun 2016 Indonesia mendapat peringkat ketiga negara dengan pengguna aktif *twitter* di dunia. (www.katadata.co.id,2016). Kegunaan *twitter* selain sebagai media untuk berbagi informasi dengan mem post berbagai macam *tweet*, twitter juga kerap sering digunakan untuk bersosialisasi antar pengguna dan mengungkapkan sentiment atau opini mereka terhadap suatu topik atau isu – isu yang sedang hangat diperbincangkan, tidak hanya opini yang positif tetapi juga yang negative.

2. KAJIAN LITERATUR DAN PENGEMBANGAN HIPOTESIS

Data Mining

Menurut Hermawati (2013) *Data Mining* berisi pencarian tren atau pola yang diinginkan dalam *database* besar untuk membantu pengambilan keputusan di waktu yang akan datang. Pola – pola ini dikenali oleh perangkat tertentu yang dapat memberikan suatu analisa data berguna dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin saja mengguankan perangkat pendukung keputusan yang lainnya.

Text Mining

Menurut Aditya B. R.(2015) *Text mining* adalah proses menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen dan tujuannya adalah mencari kata – kata yang dapat mewakili isi dokumen sehingga dapat dilakukan analisis keterhubungan antar dokumen tersebut.

Media Sosial

Menurut Setyani (2013) Media sosial mempunyai banyak bentuk, diantaranya yang paling popular yaitu *microblogging* (*twitter*), facebook, dan blog. *Twitter* adalah suatu situs web yang merupakan layanan dari *microblog*, yaitu suatu bentuk blog yang membatasi ukuran setiap posnya, yang memberikan fasilitas bagi pengguna untuk dapat menuliskan pesan dalam twtiiter *update* hanya berisi 140 karakter. *Twitter* merupakan salah satu jejaring sosial yang paling mudah digunakan, karena hanya memerlukan waktu yang singkat tetapi informasi yang disampaikan dapat langsung menyebar secara luas.

Twitter API

Twitter adalah sebuah media sosial dan layanan microblogging yang mengijinkan penggunanya untuk mengirimkan pesan realtime. Pesan ini populer dengan sebutan *tweet*. *Tweet* adalah sebuah pesan pendek dengan panjang karakter yang dibatasi hanya sampai 140 karakter. Dikarenakan keterbatasan karakter yang bisa dituliskan, sebuah tweet seringkali mengandung singkatan, bahasa selang maupun kesalahan pengejaan(Agarwal et al.,2014).

Twitter diciptakan oleh Jack Dorsey ditahun 2006 dan pertama meluncur didunia maya saat Juli 2006. Dengan alamat http://www.twitter.com yang masih digunakan hingga saat ini. *Twitter* memiliki Application Programming Interface sedemikian hingga developer dapat mengembangkan aplikasi sesuai dengan kebutuhannya masing-masing. Dokumentasi mengenai *twitter* API dapat dilihat pada http://www.developer.twitter.com.

Analisis Sentimen

Sentimen analisis atau bisa disebut juga opinion mining, adalah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi orang-orang terhadap entitas seperti produk,layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya.

Opinionatau pendapat adalah pusat dari semua aktifitas manusia karena merupakan pemberi pengaruh utama perilaku kita. Analisis sentimen dan Opinion miningterutama berfokus pada opini yang mengekspresikan atau menyiratkan sentimen positif atau negatif. Pada awal tahun 2000, analisis sentimen sudah mulai berkembang menjadi salah satu penelitian aktif dalam natural language processing(NLP).

Information Retrieval

Information Retrieval merupakan sekumpulan algoritma dan teknologiuntuk melakukan pemrosesan, penyimpanan dan menemukan kembali informasi (terstruktur) pada suatu koleksi data yang besar (Manning, Raghafan dan Schutze,2009). Berikut proses-proses information retrieval :

a. Tokenizing

Merupakan pemotongan kata berdasarkan tiap kata yang menyusunnya menjadi potongan tunggal. Sehingga hasil dari proses ini merupakan kata tunggal yang dimasukkan ke dalam *database* untuk keperluan pembobotan.

b. Stopword Removal

Merupakan tahap menghilangkan kata yang tidak sesuai dengan topik dokumen, jika ada kata tersebut tidak mempengaruhi akurasi dalam klasifikasi sentimen dokumen. Kata yang akan dihilangkan dihimpun dalam databse kata stopword. Jika dalam dokumen *tweet* ada yang sesuai dengan kata dalam stopword maka kata tersebut akan dihilangkan dan diganti dengan karakter spasi.

c. Stemming

Merupakan suatu proses untuk mengubah kata – kata yang terdapat dalam suatu dokumen ke dalam kata – kata akarnya dengan menggunakan aturan – aturan tertentu. Proses stemming bahasa Indonesia dilakukan dengan menghilangkan sufiks, prefix, dan konfiks pada dokumen.

d. Pembobot Kata

Setelah melakukan Preprocessing Text dihasilkan berbantu token yang terpisah dari kata yang lain dan sudah dalam bentuk dasar.

Metode TF-IDF merupakan metode pembobotan dalam bentuk sebuah metode yang merupakan integrasi antara term frequency (TF) dan Inverse document frequency (IDF). (Yan dan Liu,1999). Berikut rumus yang digunakan untuk mencari bobot kata dengan metode term frequency (TF) dan Inverse document frequency (IDF) :

$$idf = log\left(\frac{D}{df}\right)$$

Keterangan :
D = Jumlah semua dokumen dalam koleksi
df = Jumlah dokumen yang mengandung term t

e. Penggabungan Kata Berdasarkan Sinonim

Menurut Kamus Besar Bahasa Indonesia (KBBI) sinonim adalah bentuk bahasa yang maknanya mirip atau sama dengan bahasa lain. Proses sinonim akan dilakukan ketika ada kata berbeda namun memiliki makna yang sama, untuk meminimalkan jumlah kata yang terdapat pada sistem, tanpa menghilangkan jumlah frekuensi. (Rarasati,2015).

Teks Preprocesing

Preprocessing merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, *preprocessing* data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah yang diproses oleh sistem. *Preprocessing* sangat penting dalam pembuatan *analisis sentimen*, terutama untuk media sosial yang sebagian besar berisi kata – kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki *noise* yang besar.

Naïve Bayes Classifier (NBC)

Naïve Bayes Classifier adalah salah satu metode yang popular digunakan untuk keperluan *data mining* karena kemudahan penggunaannya (Hall, 2006)

Teorema Bayes merupakan teorema yang mengacu pada konsep probabilitas bersyarat.Secara umum Teorema Bayes dapat dinotasikan pada persamaan 2.4 berikut:

$$\mathrm{P(A\backslash B)} = \frac{P(A\backslash B)P(A)}{P(B)}$$

Keterangan :
B : Data dengan class yang belum diketahui
A : Hipotesis data merupakan suatu class spesifik
P(A|B) : Probabilitas hipotesis A berdasar kondisi B (posteriori probabilitas)
P(A) : Probabilitas hipotesis A (prior probabilitas)
P(B|A) : Probabilitas B berdasar kondisi pada hipotesis A.
P(B) : Probabilitas B

Evaluasi

Evaluasi performasi dilakukan untuk menguji hasil dari klasifikasi dengan mengukur nilai performasi dari sistem yang telah dibuat.

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Possitive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Keterangan :

1. TP (*True Positive*) : kelas yang diprediksi positif dan diprediksi oleh system klasifikasi kelas positif.
2. TN (*True Negative*) : kelas yang diprediksi positif dan diprediksi oleh system klasifikasi kelas negatif.
3. FP (*False Positive*) : kelas yang di prediksi negative dan di prediksi oleh system klasifikasi kelas positif.
4. FN (*False Negative*) : kelas yang di prediksi positif dan di prediksi

oleh system klasifikasi kelas positif.

Confussion Matrix bermanfaat untuk menganalisis kualitas classifier dalam mengenali tuple-tuple dari kelas yang ada. TP dan TN menyatakan pada classifier mengenali tuple dengan benar, artinya tuple positif dikenali sebagai positif dan tuple negatif dikenali sebagai negatif. Sedangkan, FP dan FN menyatakan bahwa classifier salah dalam mengenali tuple, tuple negatif dikenali sebagai positif dan tuple negative dikenali sebagai positif. Ada beberapa dalam formula perhitungan performa klasifikasi salah satunya yaitu nilai akurasi biasa ditampilkan dalam presentase. Akurasi adalah nilai ketepatan dimana pengguna memprediksi suatu kata sesuai dengan jawaban suatu sistem. Berikut perhitungan nilai akurasi :

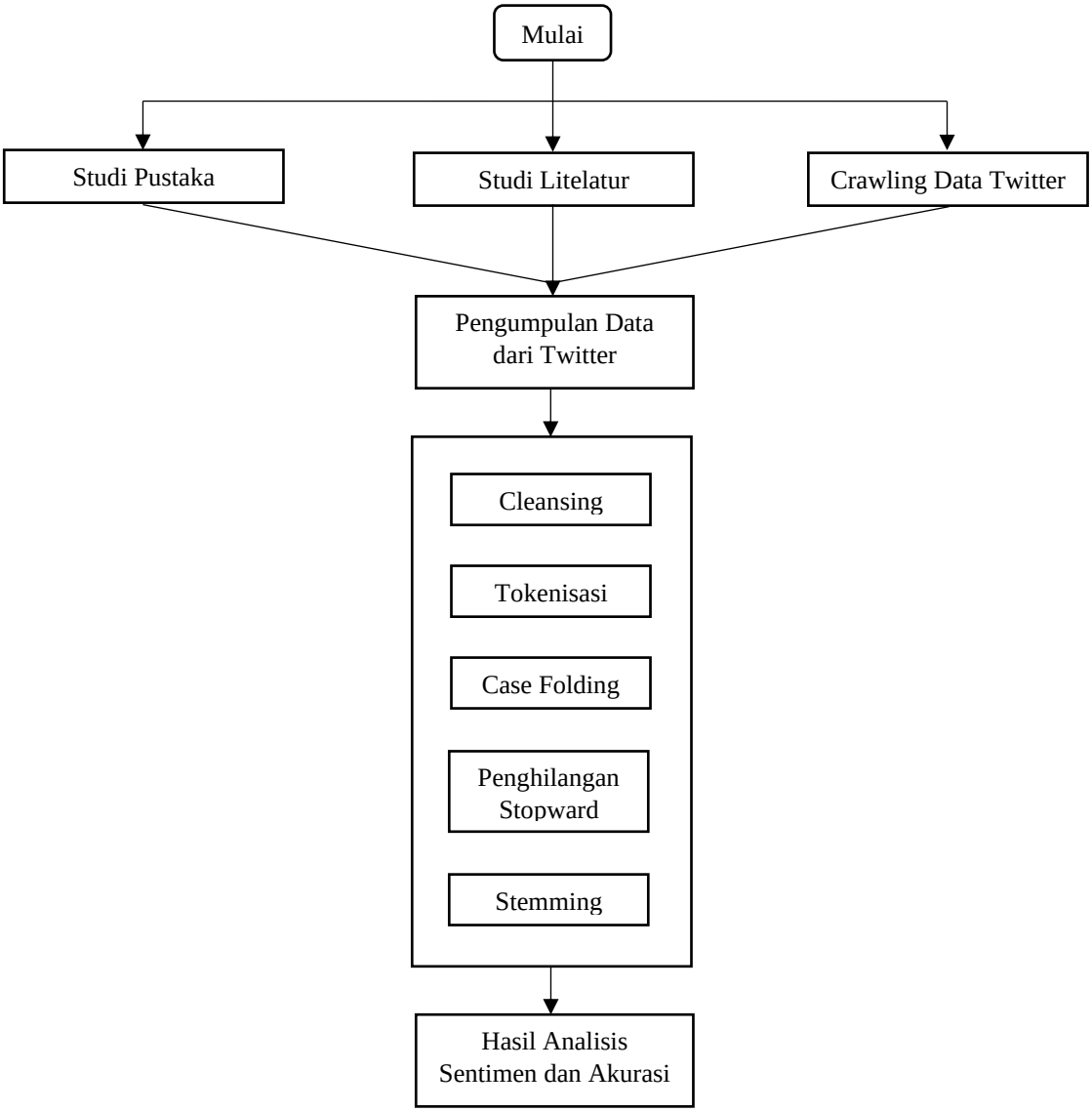
$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN}$$

3. METODE PENELITIAN

Pada penelitian ini, data yang digunakan ialah *tweet* berbahasa Indonesia yang terdapat di *Twitter*. *Tweet* yang digunakan ialah *tweet-tweet* yang mengandung opini positif dan negatif tentang pemerintah daerah.

Pencarian data dilakukan dengan menggunakan hastag #pemerintahdaerah dengan pengambilan sample data sebanyak 500 sampel diambil dengan menggunakan *twetter API* yaitu memilih kalimat-kalimat *tweet* yang berbahasa Indonesia dan tidak mengandung gambar. *Tweet* yang telah dipilih kemudian disimpan ke file teks. Kemudian file teks tersebut digunakan sebagai inputan pada system untuk diolah lebih lanjut. Data yang didapatkan dari hasil scapping data menggunakan twitter API dibagi menjadi dua yaitu data Training dan data testing dengan perbandingan 80% : 20%.

Analisis sentiment ini dimulai proses input data *tweet* dengan cara *crawling*. *Input* yang dimasukkan sistem adalah dokumen yang berupa *tweet* dari akun *Twitter* yang berupa opini. Data *tweet* tersebut didapat dengan memanfaatkan *fitur API (Application Interface)* yang telah disediakan oleh *Twitter.*



4. HASIL PENELITIAN

Proses Preprocessing Dokumen

a. Case Folding

Fungsi case_folding ini akan mengubah huruf kapital menjadi huruf kecil, fungsi *case_folding* yang dibuat.

Sebelum	Sesudah
Pemerintah kota bekasi mencatat realisasi pendapatan daerah per agustus ini sudah masuk sebesar rp triliun	pemerintah kota bekasi mencatat realisasi pendapatan daerah per agustus ini sudah masuk sebesar rp triliun

b. Cleansing

berguna untuk membersihkan data tweet seperti angka, tanda baca, link, hastag, mention, dan menghasilkan kata yang akan diolah pada proses selanjutnya.

Sebelum	Sesudah
pemerintah bakal glontorkan dana rp796,3 triliun untuk pemulihan ekonomi daerah 2021 https://t.co/60znsclp4g	Pemerintah bakal glontorkan dana rp triliun untuk pemulihan ekonomi daerah

c. Tokenizing

berguna untuk memisahkan setiap kata yang dihubungkan dengan karakter spasi menjadi setiap kata yang dihimpun pada array.

Sebelum	Sesudah
Pemerintah kota bekasi mencatat realisasi pendapatan daerah per agustus ini sudah masuk sebesar rp triliun	pemerintah kota bekasi mencatat realisasi pendapatan daerah per agustus ini sudah masuk sebesar rp triliun

d. Stopword Removal

digunakan untuk menghilangkan kata yang tidak berpengaruh dalam proses sentimen.

Sebelum	Sesudah
Perkasaupdate pemerintah kota memprioritaskan rapid tes masal kepada asn kota di dinstansi pelayanan	Pemerintah Memprioritaskan Instansi Pelayanan publik

publik

Pelabelan Kelas Sentimen

Proses pelabelan dilakukan secara otomatis dengan cara menghitung nilai pelabelan sentiment menggunakan kamus Lexicon dan manual. Pelabelan di bagi menjadi 2 yaitu Sentimen positif dan sentiment negative dengan cara melakukan skoring. Jika suatu kalimat memiliki skor > 0 dan = 0 akan diklasifikasikan dalam kelas positif sedangkan jika kalimat memiliki skor < 0 diklasifikasikan dalam kelas negative. Hasil pelabelan data ulasan dapat dilihat pada table 5 sebagai berikut :

Kelas Sentimen	Skor	Ulasan
Positif	2	Pemerintah Daerah (Pemda) DIY terus berupaya meningkatkan perekonomian di tengah pandemi seperti saat ini.
Negatif	0	

Jumlah Ulasan Pada Kelas Sentimen

Kelas Sentimen	Jumlah Ulasan
Positif	402
Negatif	98
Total	500

Pembuatan Data Training dan Data Testing

Klasifikasi	Jumlah	Data Training (80%)	Data Testing (20%)
Positif	402	321.6	80.4
Negatif	98	78.4	19.6
Total	500	400	100

Berdasarkan tabel diatas dengan perbandingan data training dan data testing sebesar 80% : 20%. Dari data ulasan berbahasa Indonesia sebanyak 500, sebanyak 402 digunakan sebagai data training dan 98 sebagai data testing. Dengan masing – masing data training klasifikasi positif sebanyak 322 dan negative sebanyak 78. Sedangkan

data testing sebanyak 80 klasifikasi positif dan 20 klasifikasi negative.

Klasifikasi Metode Naive Bayes

Klasifikasi sentimen ini dilakukan secara otomatis dengan mengimplememtasikan algoritma *Naïve Bayes Classification*. Proses ini diimplementasikan pada fungsi klasifikasi_sentimen() dengan membandingkan bobot setiap kata pada data *testing* dengan kata pada data *training*, jika kata tersebut tidak ditemukan dalam data *training* maka bobotnya dinilai 1. Hasilnya setiap dokumen training ini dijumlah bobot kata probabilitas positif dan probabilitas negatifnya. Selanjutnya bobot dokumen dibandingkan, jika bobot dokumen probabilitas positif lebih besar maka hasil sentimen adalah positif, dan jika bobot probabilitas negatif lebih besar maka hasil sentimen adalah negatif.

Dalam proses evaluasi untuk mengetahui hasil akurasi klasifikasi digunakan metode *Confution Matrix* pada setiap kelas. Berikut merupakan hasil confusion matrix data ulasan pemerintah daerah:

		Aktual		
	Kelas	Negatif	Positif	Total
Prediksi	Negatif	17	29	46
	Positif	0	54	54
	Total	17	83	

Pada hasil perhitungan dengan *Confution Matrix* didapatkan hasil prediksi ulasan yang masuk dalam kelas negative adalah sebanyak 17 dengan 17 ulasan yang telah terklasifikasi dengan benar dan tidak terdapat kesalahan prediksi yang masuk dalam ulasan positif. Sedangkan dari 83 ulasan yang masuk dalam kelas positif terdapat 54 ulasan yang terklasifikasi dengan benar dam terdapat kesalahan prediksi sebanyak 29 yang masuk dalam ulasan negative. Dari hasil *confution matrix* tersebut diperoleh tingkat akurasi sebesar 71% yang artinya dari 100 data ulasan yang diuji terdapat 71 data ulasan yang benar diklasifikasikan dengan *Naïve Bayes Classifier*.

5. SIMPULAN

Berdasarkan hasil analisis sentiment yang telah dilakukan, diperoleh kesimpulan :

1. Berdasarkan 500 data tweet yang telah diperoleh dari aplikasi twitter sebanyak 402 kata ulasan yang masuk ke dalam sentiment positif dan 98 ulasan masuk ke dalam sentiment negative. Sedangkan kata yang paling banyak muncul diantaranya adalah kata “pemerintah”, “daerah”, lalu diikuti kata “buzzer”, “dana”, “sumber”,”daya”, dll.
2. Hasil dari menggunakan metode Naïve Bayes Classifier dalam mengklasifikasikan data ulasan mengenai pemerintah daerah dengan perbandingan data training dan data testing masing – masing sebanyak 80% : 20% diperoleh hasil klasifikasi sentiment dengan tingkat akurasi sebesar 71%.

6. REFERENSI

Aditya, B. R. (2015). Penggunaan Web Crawler untuk Menghimpun Tweet dengan Metode Pre-Processing Text Mining. Jurnal Infotel Vol. 7 No. 2 .

Dragut, E, Fang, F., Sistla, P., Yu, S. & Meng, W. 2009. *Stop Word and Related Problems in Web Interface Integration*. Diakses dari http://www.vldb.org/pvldb/2/vldb09-384.pdf. Diakses pada 20 November 2016.

Faradhillah, N. Y., Kusumawardani, R. P., & Hafidz, I. (2016). Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter pada Akun Resmi Pemerintah Kota Surabaya Berbasis Pembelajaran Mesin. Seminar Nasional Sistem Informasi Indonesia .

Hadna, N. M., Santosa, P. I., & Winarno, W. W. (2016). Studi Literatur tentang Perbandingan Metode untuk Proses Analisis Sentimen di Twitter. Seminar Teknologi Informasi dan komunikasi (SENTIKA) .

Liu,B. (2012)., Sentimen Analysis and Opinion Mining., Morgan & Claypool Publishers. Diakses dari https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinonMining.pdf. Diakses pada 18 September 2016.

Mujilahwati, S. (2016). Pre-processing Text Mining pada Twitter. Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA) .

Nurhuda, F., Sihwi, S. W., & Doewes, A. (2013). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter menggunakan Metode Naive Bayes Classifier. Jurnal ITSMART

Pang, B dan Lee, L. (2008). *Opinion Mining and Sentimen Analysis, Foundation and Trends In Information Retrieval*, vol. Volume 2, no. Issue 1-2,pp. 1-135.

Syadid, Faqi (2019). Analisis Sentimen Komentar Nitizen Terhadap Calon Presiden Indonesia 2019 Dari Twitter Menggunakan Algoritma Term Frequency-Invers Document Frequency (TF-IDF) dan Metode Multi Layer Perceptron (MLP) Neural Network.

Wati, R. (2016). Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan menggunakan Naive Bayes. Jurnal Evolusi .

www.apjii.or.id. (2017). Penetrasi dan Perilaku Pengguna Internet Indonesia

www.apjii.or.id. (2018). Penetrasi dan Perilaku Pengguna Internet Indonesia