JURNAL INFORMATIKA, Vol.7 No.1 April 2020, Halaman 85~90
ISSN: 2355-6579 | E-ISSN**:** 2528-2247

Penggunaan Algoritma Klasifikasi Terhadap Analisa Sentimen Pemindahan Ibukota Dengan Pelabelan Otomatis

Jananto Watori[1], Riska Aryanti[2], Agus Junaidi[3]

[1] Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
e-mail: [1]janantowatori@gmail.com

[2,3] Universitas Bina Sarana Informatika
e-mail: [2]riska.rts@bsi.ac.id, [3]agus.asj@bsi.ac.id

Abstrak

Perkembangan media yang begitu pesat, memunculkan banyak media online dari media berita sampai media sosial. Media sosial saja sudah begitu banyak, dari *Facebook*, *Twitter*, *Instagram*, *Tumblr*, *Linkedin* dan masih banyak lagi. Berdasarkan fakta yang ada dalam penerapannya sendiri untuk kehidupan sehari-hari sosial media sangat sering digunakan. Dampak positif internet dalam perkembangan *information technology* (IT) sebenarnya membawa banyak keuntungan, misalnya saja memudahkan dalam hal komunikasi, mencari dan mengakses informasi. Namun, terdapat dampak negatif dalam perkembangannya, yaitu contohnya dalam penyebaran berita *hoax* ataupun ujaran kebencian. Dengan menggunakan internet, dapat memperkuat atas suatu gagasan dan pendapat dalam suatu kelompok maupun individu pada situs web berita dan media sosial. Penelitian ini membahas tentang bagaimana melakukan analisa sentimen yang berasal dari *tweet* pengguna twitter tentang pemindahan ibukota Indonesia. Gagasan serta pendapat publik melalui *twitter* yang dalam jumlah besar, setidaknya dapat menganalisa secara global tentang sentimen pemindahan ibukota yang akan dilakukan di Indonesia. Penelitian ini menggunakan pelabelan otomatis menggunakan (*Valence Aware Dictionary and sEntiment Reasoner) Vader* dengan metode *Naïve Bayes* dan *Support Vector Machine*. Sehingga, dapat ditarik kesimpulan bahwa pelabelan pada setiap cuitan di twitter dapat dilakukan sehingga menghasilkan score pada dataset. Dan dari algoritma yang digunakan, algoritma *Support Vector Machine* menghasilkan nilai akurasi dan AUC yang paling baik yakni akurasi sebesar 76,40% dan AUC sebesar 0,771.

Kata Kunci*: Naïve Bayes*, *Support Vector Machine, Vader*

Abstract

Media development is so rapid, led to many online media from news media to social media. Social media have been so many, from Facebook, Twitter, Instagram, Tumblr, Linkedin and many more. Based on the facts in its own application to everyday life very often used social media. The positive impact of the internet in the development of information technology (IT) actually brings many benefits, for example it makes it easy in terms of communication, finding and accessing information. However, there is a negative impact on its development, ie for example in the dissemination of news hoax or hateful speech. By using the Internet, to strengthen on an ideas and opinions in a group or individually on a news website and social media. This study discusses how to analyze the sentiment that comes from twitter user tweet about moving the capital of Indonesia. The ideas and public opinion through Twitter in large numbers, at least can analyze globally about the sentiment of moving capital to be carried out in Indonesia. This study uses automatic labeling using (*Valence Aware Dictionary and sEntiment Reasoner)* Vader with the Naïve Bayes method and Support Vector Machine. So, it can be concluded that labeling on every tweet on Twitter can be done so as to produce a score on the dataset and from the algorithm used, the *Support Vector Machine* algorithm produces the best accuracy and AUC values, namely 76.40% accuracy and AUC of 0.771..

Keywords*: Naïve Bayes*, *Support Vector Machine, Vader*

Pendahuluan

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. Berdasarkan data pengguna internet dari APJII atau Asosiasi Penyelenggara Jasa Internet Indonesia, terdapat sebanyak 143,26 juta jiwa dari total 262 juta orang yang ada pada Indonesia. Serta dalam hal komposisi pengguna internet berdasar usia APJII menggolongkan dalam 4 golongan yang diantaranya terdapat golongan umur 13 sampai dengan 18 tahun sebanyak 16,68%, selanjutnya pada golongan umur 19 sampai dengan 34 tahun sebanyak 49,52%, yang ketiga untuk golongan umur 35 sampai dengan 54 tahun sebanyak 29,55%, dan yang terakhir yaitu pada golongan umur lebih dari 54 tahun sebanyak 4,24% (APJII 2017). Pemanfaatan internet saat ini yang menarik adalah untuk bidang gaya hidup dengan layanan paling sering terakses antara lain menempati urutan pertama sebanyak 89,35% untuk aplikasi *chatting*, dan sosial media 87,13 % dengan menempati urutan kedua. Berdasarkan fakta yang ada dalam penerapannya sendiri untuk kehidupan sehari-hari sosial media sangat sering tergunakan (APJII 2017).

Dampak positif internet dalam perkembangan *information technology* (IT) sebenarnya membawa banyak keuntungan, misalnya saja memudahkan dalam hal komunikasi, mencari dan mengakses informasi. Namun, terdapat dampak negatif dalam perkembangannya, yaitu contohnya dalam penyebaran berita *hoax* ataupun ujaran kebencian. Dengan menggunakan internet, dapat memperkuat atas suatu gagasan dan pendapat dalam suatu kelompok maupun individu pada situs web berita dan media sosial (Evolvi 2018).

Menurut Brown, ujaran kebencian *online* berbeda dari ujaran kebencian offline karena internet ditandai oleh akses yang mudah, tanpa perlu adanya pendengar, anonimitas, dan spontanitas. Salah satu cabang riset yang kemudian berkembang dari situasi ledakan informasi di internet adalah analisa sentimen (Brown 2018). Analisis sentimen disebut juga opinion mining, adalah bidang ilmu yang meng-analisa pendapat, sentimen, evaluasi, penilaian, sikap dan emosi publik terhadap entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa, topik, dan atribut mereka(Zhang, Liu, and Francisco 2016). *Twitter* adalah sebuah media sosial yang meng-ijinkan penggunanya untuk mengirimkan pesan *realtime*. Disisi lain, *Data Mining* merupakan suatu metode menemukan pengetahuan yang bermanfaat dari volume *data* besar. Dengan demikian *Data Mining* dapat digunakan untuk klasifikasi analisa sentimen karena memiliki jumlah *data twitter* yang besar.

Terkait penelitian terhadap analisis sentimen sudah pernah dilakukan sebelumnya. Penelitian sebelumnya telah banyak dilakukan dengan menerapkan metode *Naïve Bayes Classifier* (*NBC)* dan *Support Vector Machine (SVM).* Terdapat judul "Analisis Sentimen *Hatespeech* Pada Twitter Dengan Metode *Naïve Bayes Classifier* dan *Support Vector Machine "*(Buntoro 2016b). Proses klasifikasi pada penelitian ini menggunakan metode klasifikasi *NBC* dan *SVM*. Data yang digunakan adalah *tweet* dalam bahasa Indonesia dengan tagar *HateSpeech* (*#HateSpeech*), dengan jumlah dataset sebanyak 522 tweet yang didistribusikan secara merata menjadi dua sentimen *HateSpeech* dan *GoodSpeech*. Hasil akurasi tertinggi didapatkan saat menggunakan metode klasifikasi SVM dengan tokenisasi *unigram*, *stopword* list Bahasa Indonesia dan *emoticons*, dengan nilai rata-rata akurasi mencapai 66,6%, nilai presisi 67,1%, nilai *recall* 66,7% nilai TP rate 66,7% dan nilai TN rate 75,8%(Buntoro 2016a).

Penelitian lainnya yaitu tentang "Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes" (Utami & Wahono 2015). Internet me-rupakan bagian penting dari kehidupan sehari-hari. Saat ini, tidak hanya dari anggota keluarga dan teman-teman, tetapi juga dari orang asing yang berlokasi diseluruh dunia yang mungkin telah mengunjungi restoran tertentu. Konsumen dapat memberikan pendapat mereka yang sudah tersedia secara online. Ulasan yang terlalu banyak akan memakan banyak waktu dan pada akhirnya akan menjadi bias. Klasifikasi sentimen bertujuan untuk mengatasi masalah ini dengan cara mengkla-sifikasikan ulasan pengguna ke

pendapat positif atau negatif. Algoritma *Naïve Bayes* adalah tekhnik machine learning yang populer untuk klasifikasi teks, karena sangat sederhana, efisien dan memiliki performa yang baik pada banyak domain. Namun, *Naïve Bayes* memiliki kekurangan yaitu sangat sensitif pada fitur yang terlalu banyak, sehingga membuat akurasi menjadi rendah. Oleh karena itu, dalam penelitian ini menggunakan *Information Gain* untuk seleksi fitur dan metode *adaboost* untuk mengurangi bias agar dapat meningkatkan akurasi algoritma *Naïve Bayes*. Penelitian ini menghasilkan klasifikasi teks dalam bentuk positif dan negatif dari review restoran. Pengukuran *naïve bayes* berdasarkan akurasi sebelum dan sesudah penambahan metode seleksi fitur. Validasi dilakukan dengan menggunakan 10 *fold cross validation*. Sedangkan pengukuran akurasi diukur dengan *confusion matrix* dan kurva ROC. Hasil penelitian menunjukkan peningkatan akurasi *naïve bayes* dari 73.00% menjadi 81.50% dan nilai AUC dari 0.500 menjadi 0.887. Sehingga dapat disimpulkan bahwa integrasi metode *information gain* dan *adaboost* pada analisis sentimen review restoran ini mampu meningkatkan akurasi algoritma *naive bayes*(Utami & Wahono 2015).

Penelitian lainnya yaitu tentang "Analisis Sentiment Pada Sosial Media *Twitter* Menggunakan *Naive Bayes Classifier* Terhadap Kata Kunci "Kurikulum 2013""(Pamungkas, Setiyanto, and Dolphina 2015). Dalam penelitian ini menerapkan proses *n-gram* karakter untuk seleksi fitur serta menggunakan algoritma *NBC* untuk mengklasifikasi sentimen secara otomatis. Dalam penelitian ini menggunakan 3300 data *tweet* tentang sentimen kepada kata kunci "kurikulum 2013". Data tersebut diklasifikasi secara manual dan dibagi kedalam masing-masing 1000 data untuk sentimen positif, negatif dan netral. Hasil penelitian ini menghasilkan sebuah sistem yang dapat mengklasifikasi sentimen secara otomatis dengan hasil pengujian 3000 data latih dan 100 tweet data ujicoba mencapai 91 % (Pamungkas, Setiyanto, and Dolphina 2015).

Penelitian lainya yaitu dengan judul "Komparasi Algoritma Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Pengguna Busway" pada penelitian ini pelabelan dilakukan secara manual menganalisa sentimen menggunakan data busway dengan klasifikasi sentimen menggunakan Naive Bayes dan Algoritma Genetika(Aryanti et al. 2019). Dalam penelitian selanjutnya yaitu "*Election result prediction using Twitter Sentiment Analysis* (Ramteke et al. 2016)". Pada penelitian ini menggunakan 2 kali proses *labeling data*, tahap pertama menggunakan manual *labeling* dan tahap kedua menggunakan *VADER* untuk pelabelannya. Namun, setelah memberi label pada dataset, menghasilkan sekitar 30 ribu tweet. Ini karena ambang batas yang ditetapkan untuk *Valence Aware Dictionary and sEntiment Reasoner (VADER)* sangat tinggi, sekitar 80%. Akibatnya, tweet yang ambigu atau tweet yang sangat netral dihilangkan sehingga meningkatkan kualitas dataset training. Selanjutnya menggunakan dua algoritma, Multinomial *Naïve Bayes* dan *Support Vector Machines* dari *package nltk* dan *scikit-learn* pada *library python* untuk menentukan polaritas *tweet*. Hasilnya yaitu algoritma SVM dari scikit-learn memberikan akurasi terbaik untuk klasifikasi sebesar 0.99(Ramteke et al. 2016).

Pada penelitian ini akan membahas tentang bagaimana melakukan analisa sentimen yang berasal dari tweet pengguna twitter tentang pemindahan ibukota Indonesia. Gagasan serta pendapat publik melalui twitter yang dalam jumlah besar, setidaknya dapat menganalisa secara global tentang sentimen pemindahan ibukota yang akan dilakukan di Indonesia. Berdasarkan hasil dari beberapa penelitian diatas, analisa yang akan dilakukan yaitu dengan memanfaatkan analisa sentimen mengunakan metode *Supervised Machine Learning*.

Metode Penelitian

Pada penelitian ini penulis mencoba menggunakan pelabelan otomatis menggunakan *Vader* dengan metode *Naïve Bayes* dan *Support Vector Machine* untuk melihat hasil akurasi yang terbaik dari algoritma yang digunakan. Tahap-tahap yang penulis lakukan pada penelitian ini:

1. Pengumpulan Data

Penelitian dilakukan dengan mengambil dataset dari media sosial twitter pada periode Januari 2019 sampai November 2019. Selanjutnya data tersebut diberikan label menggunakan vader sebagai pelabelannya.

2. *Preprocessing Data*

Kemudian sebelum data di klasifikasikan, data tersebut akan diproses melalui tahapan *pre-processing* yaitu.

Transform Cases untuk mengubah huruf kapital yang masih ada pada text akan diubah menjadi huruf kecil semua. Hal ini dilakukan agar ketikan dilakukan proses ke dalam model klasifikasi terdapat keseragaman huruf dan tidak terjadi kesalahan dalam proses tokenize.

Tokenize merupakan proses untuk memisah-misahkan kata. Proses memotong setiap kata dalam teks dan mengubah huruf dalam dokumen menjadi huruf kecil. Hanya huruf yang diterima, sedangkan karakter khusus atau tanda baca akan dihilangkan.

Filter token by length adalah proses yang ada pada data *preparation* untuk menghilangkan sejumlah kata (setelah proses *tokenize*) dengan panjang karakter tertentu. Pada penelitian ini panjang minimum karakter yang digunakan adalah 4 karakter dan panjang maksimum 25 karakter. Artinya kata yang panjangnya kurang dari 4 karakter dan lebih dari 25 karakter akan dihilangkan. Untuk mendapatkan hasil seperti ini maka dilakukan setting pada parameter dari operator ini

Stopwords removal adalah proses menghilangkan kata-kata yang sering muncul namun tidak memiliki pengaruh apapun dalam ekstraksi sentimen suatu *review*. Kata yang termasuk seperti kata penunjuk waktu, kata tanya.

Stemming akan mengubah setiap kata yang berimbuhan akan menjadi kata dasar.

3. Metode Algoritma

a) Naïve Bayes Classifier

Algoritma *Naïve Bayes* dibangun di atas teorema Bayes, dinamai menurut Pendeta Thomas Bayes. Karya Bayes dijelaskan dalam "*Essay Towards Solving a Problem dalam Doctrine of Chances*" (1763), yang diterbitkan secara anumerta dalam Transaksi Filosofis dari Royal Society of London oleh Richard Price(Kotu and Deshpande 2014). *Naïve Bayes* merupakan peng-klasifikasian dengan metode probabilitas dan statistik yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes (Wicaksono 2018).

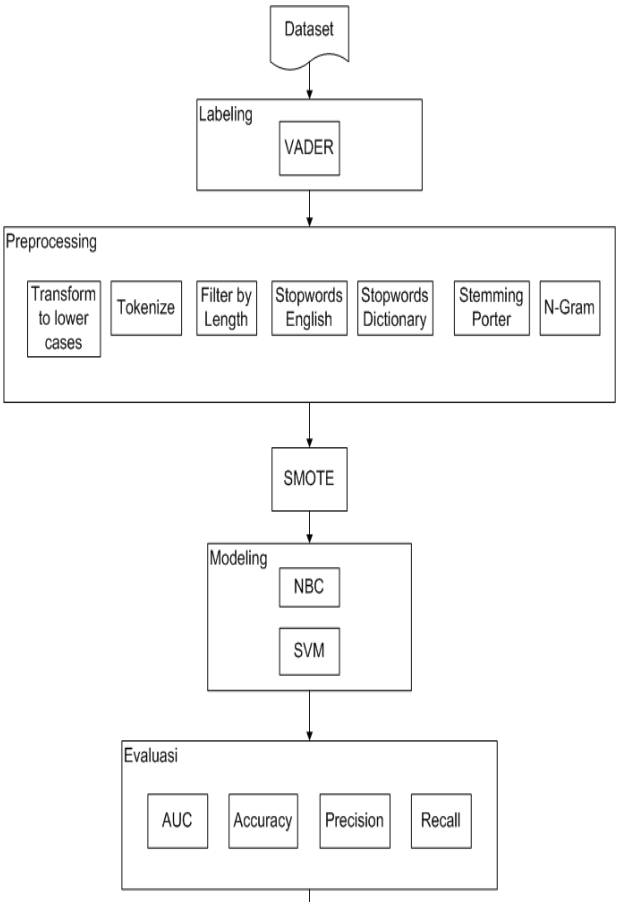
b) Support Vector Machine Classifier (SVM)

Gagasan mendasar di balik algoritma SVM adalah untuk mengidentifikasi hyperplane (atau serangkaian hyperplanes) yang dapat memberikan pemisahan terbaik antara instancedata pelatihan. Di antara semua kemungkinan *hyperplanes* separasi, algoritma SVM mencoba mengidentifikasi yang memiliki jarak terbesar ke instance pelatihan terdekat dari kelas apa pun, karena hal itu akan tercermin ke dalam kesalahan generalisasi yang lebih rendah dari c*lassifier* (Ignatow & Mihalcea 2018).

Dan setelah dilakukan pengklasifikasian menggunakan algoritma *NBC* dan *SVM*. Maka tahap berikutnya melakukan pengujian dengan data training dan data testing dari dataset yang telah di ambil, dan setelahnya dilihat perbandingan performance yang terhadap algoritma yang digunakan.

Hasil dan Pembahasan

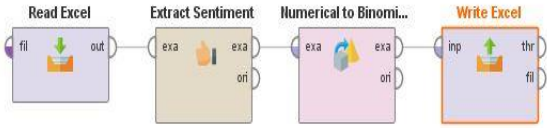
Ada beberapa langkah yang dilakukan pada penelitian ini seperti terlihat pada gambar 1.



Gambar 1. Metodologi Penelitian

1. Pengumpulan Data

Penelitian dilakukan dengan mengambil dataset dari media sosial *twitter* pada periode Januari 2019 sampai November 2019. Data tersebut dimasukkan dalam *excel* untuk mempermudah proses pengolahan data pada *Tools RapidMiner* untuk dilakukan pelabelan menggunakan *vader* seperti pada gambar 2.



Gambar 2 Proses pelabelan

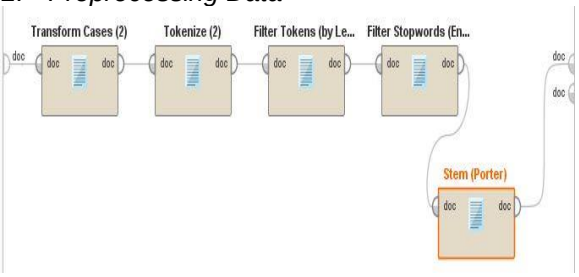
Pada tabel 1 merupakan hasil proses pelabelan menggunakan vader.

Tabel 1. Hasil pelabelan dengan vader

Vader	***Tweet***
Negative	*Do ngarepin folk dong demo pack .. Members of the House that much amount of what all otaknye already parroting to jokowi ?? Make a motion in the House of Representatives rejects transfer of capital dong .. .. @RamliRizal @karniilyas*
Positive	*Oops ... not if another bro ni ... yes weve really speechless .... Lawong our leaders as lackeys for them .... check it out regime very busy divert this issue with the issue of transfer of capital by way will sell state assets. #WeStandWithUighur*
Positive	*CAPITAL MOVE So that there is reason to selling state land to brokers? Call Society Coalition Capital Displacement Ambitious https://idtoday.co/2019/12/koalisi-masyarakat-sebut-pemindahan-ibu.html... #UyghurHumanitarianCrisis @Idtodaydotco*
Negative	*In addition to inaugurating the toll in Borneo, @jokowi President also met with about 30 indigenous leaders to discuss East Kalimantan Indonesia's capital redeployment plan #KalimantanPunyaTol*

Sumber: Hasil Penelitian(2020)

2. *Preprocessing Data*



Gambar 3 Proses Preprocessing Data

Setelah dataset terkumpul, maka selanjutnya adalah proses untuk memulai pengolahan data, yaitu proses *pre-processing*. Data tersebut tidak bisa langsung dimasukan dalam pengolahan untuk sentimen analisis, maka dilanjutkan dengan tahapan *pre-processing* seperti pada gambar 3.

Tabel 2. Perbandingan Teks Sebelum dan Sesudah Dilakukan Proses *Transform Cases*

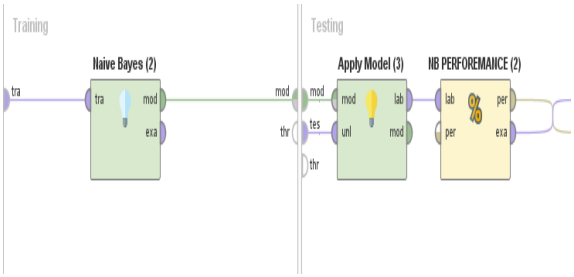
Proses	**Teks**
Teks sebelum dilakukan proses *Transform Cases*	*Thank stay in Indonesia; Jakarta pollution, capital of stay moving. People in contact with the smoke of forest Singapur want to move the capital where?*
Teks sesudah dilakukan proses *Transform Cases*	*thank stay in indonesia; jakarta pollution, capital of stay moving. people in contact with the smoke of forest singapur want to move the capital where?*

Sumber: Hasil Penelitian(2020)

3. Proses Validasi Algoritma

Pada proses ini menggunakan beberapa operator, sebelumnya menggunakan operator *cross validation* dengan k-10 *fold cross validation*, seperti

pada gambar 1. Yang di dalamnya terdapat berbagai operasi di antaranya dapat dilihat pada Gambar 4.



Gambar 4 Operator Cross Validation

4. Evaluasi

Setelah proses yang sudah dilakukan maka dapat digambarkan dalam bentuk tabel matrik evaluasi sebagai berikut :

Tabel 3. Matriks Hasil Evaluasi

Model Algoritma	NBC	SVM
Akurasi	78,39%	76,40%
Precision	57,01%	62,66%
Recall	54,88%	13,56%
AUC	0,534	0,771

Sumber: Hasil Penelitian(2020)

Dari tabel 3 dapat dilihat bahwa akurasi pada algoritma lebih tinggi dibanding algoritma svm, namun tingkat auc dari algoritma nbc 0,534. Sedangkan pada algoritma svm nilai akurasinya 76,40% dan tingkat AUC sebesar 0,771.

Kesimpulan

Setelah melakukan penelitian ini, dapat ditarik kesimpulan pelabelan pada setiap cuitan di *twitter* dapat dilakukan dengan menggunakan *vader* sehingga menghasilkan sentimen pada datanya. Dan dari algoritma yang digunakan, terbukti bahwa algoritma *SVM* menghasilkan nilai akurasi yang paling baik yakni mencapai 76,40% dengan tingkat *AUC* mencapai 0,771 sehingga termasuk *fair classification* dibandingkan dengan algoritma *Naive Bayes*.

Referensi

APJII. 2017. “Penetrasi Dan Perilaku Pengguna Internet Indonesia.”

Aryanti, Riska et al. 2019. “Komparasi Algoritma Naive Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Pengguna Busway.” V(1): 135–38.

Brown, Alexander. 2018. “What Is so Special about Online (as Compared to Offline) Hate Speech?” *Ethnicities* 18(3): 297–326.

Buntoro, Ghulam Asrofi. 2016a. “Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naive Bayes Classifier Dan Support Vector Machine.” *Dinamika Informatika* 5(September).

———. 2016b. “Analisis Sentimen Hatespeech Pada Twitter Dengan Metode Naïve Bayes Classifier Dan Support Vector Machine.” *Jurnal Dinamika Informatika* 5(2): 1–12.

Dini Utami, Lila, and Romi Satria Wahono. 2015. “Integrasi Metode Information Gain Untuk Seleksi Fitur Dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes.” *Journal of Intelligent Systems* 1(2): 120–26.

Evolvi, Giulia. 2018. “Hate in a Tweet: Exploring Internet-Based Islamophobic Discourses.” *Religions* 9(10).

Ignatow, Gabe, and Rada Mihalcea. 2018. *An Introduction to Text Minig*. United States of A: SAGE Publications, Inc.

Kotu, Vijay, and Bala Deshpande. 2014. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann.

Pamungkas, Dyarsa Singgih, Noor Ageng Setiyanto, and Erlin Dolphina. 2015. “Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci ‘Kurikulum 2013’’.’” 14(4): 299–314.

Ramteke, Jyoti, Samarth Shah, Darshan Godhia, and Aadil Shaikh. 2016. “Election Result Prediction Using Twitter Sentiment Analysis.” In *2016 International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 1–5.

Wicaksono, Soetam Rizky. 2018. *Studi Kasus Sistem Berbasis Pengetahuan: Membahas Metode ID3, Naïve Bayes Dan Certainty Factor*. Seribu Bintang.

Zhang, Lei, Bing Liu, and San Francisco. 2016. “Sentiment Analysis and Opinion Mining.” *Encyclopedia of Machine Learning and Data Mining*.