PENERAPAN METODE SUPPORT VECTOR MACHINE UNTUK ANALISIS SENTIMEN PENGGUNA TWITTER

Zidna Alhaq [1] , Ali Mustopa [2], Sri Mulyatun [3], Joko Dwi Santoso [4]

[1) 2)] *Informatika Universitas Amikom Yogyakarta*
[3)] *Manajemen Informatika Universitas Amikom Yogyakarta*
[4)] *Teknik Komputer Universitas Amikom Yogyakarta*
email : zidna.a@students.amikom.ac.id [1], ali.m@amikom.ac.id [2], sri.m@amikom.ac.id [3], joko@amikom.ac.id [4]

Abstraksi

Twitter merupakan salah satu media sosial yang digunakan untuk menyampaikan pendapat dan mendiskusikan berbagai topik seputar. Salah satu topik yang sering dibahas adalah marketplace. Bukalapak merupakan salah satu marketplace terpopuler di Indonesia. Bukalapak memberikan penggunanya kemampuan untuk melakukan transaksi dengan cepat dan aman. Tanggapan yang diberikan oleh pengguna tersebut dapat berupa tanggapan positif, negatif dan netral. Oleh karena itu diperlukan suatu metode yang dapat digunakan untuk mengetahui pendapat pengguna Bukalapak di media sosial Twitter. Untuk mengatasi masalah ini, diperlukan suatu metode yang dapat mengkategorikan opini-opini tersebut. Support Vector Machines merupakan salah satu metode penggalian teks yang dapat mengkategorikan opini tersebut. Data yang diperoleh dari Twiiter akan diberi label dan dianalisis menggunakan metode SVM untuk mengklasifikasikan opini tersebut. Hasil klasifikasi menggunakan metode SVM diperoleh tingkat akurasi sebesar 93%.

***Kata Kunci* :**
Python, Text Mining, Support Vector Machine, Analisis Sentimen, Kecerdasan Buatan

Abstract

Twitter is one of the social media that is used to express an opinions and discuss various topics around. One topic that is often discussed is marketplace. Bukalapak is one of the most popular marketsplace in Indonesia. Bukalapak provides its users with the ability to make transactions quickly and securely. Responses given by these users can be positive, negative and neutral responses. Therefore a method that can be used to find out the opinion of Bukalapak users on Twitter social media is needed. To solve this problem, we need a method that can categorize these opinions. Support Vector Machines is one method of extracting text that can categorize these opinions. Data obtained from Twiiter will be labeled and analyzed using the SVM method to classify these opinions. The results of the classification using the SVM method obtained an accuracy rate of 93%.

***Keywords* :**
Python, Text Mining, Support Vector Machine, Sentiment Analysis, Artificial Intellegence

Pendahuluan

Sosial media adalah alat komunikasi atau wadah yang digunakan orang-orang untuk mengemukakan pendapat atau opini mereka untuk berbagai topik. Pengguna sosial media di Indonesia sangat besar, sehingga mendorong munculnya data tekstual yang tidak terbatas. Salah satu pemanfaatan data ini adalah mengetahui sentimen publik terhadap marketplace.

Twitter adalah situs jejaring sosial mikroblogging yang digemari oleh masyarakat Indonesia. Survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia menyebutkan bahwa Twitter merupakan media sosial yang menempati peringkat 4 dalam hal kunjungan oleh masyarakat Indonesia. Kelebihan media sosial Twitter adalah dibatasinya jumlah karakter menjadi maksimal 280 karakter, sehingga di twitter tidak ada fitur “read more” karena karakter sudah ditampilkan semuanya di layar. Tweet yang diposting oleh pengguna tersebut dapat di analisa dan diolah menjadi informasi yang bermanfaat dengan teknik analisa sentimen.

Text Mining adalah sebuah teknik untuk menggali informasi dari sebuah teks. Informasi yang diambil melalui teknik text mining . Text mining merupakan bagian atau variasi dari data mining. Text mining berusaha menemukan pola dari sekumpulan data. Text mining juga dapat diartikan sebagai sebuah proses untuk menemukan suatu informasi atau tren baru yang sebelumnya tidak terungkap dengan memproses dan menganalisis data dalam jumlah besar [1].

Sentiment analysis atau opinion mining adalah cabang ilmu text mining yang mempelajari sentimen yang ada pada suatu teks opini. Prinsip dasar dari analisis sentiment adalah mengklasifikasikan sebuah teks apakah teks tersebut bernilai positif, negatif atau netral. Sentiment analysis atau opinion mining mengacu pada bidang yang luas dari pengolahan

bahasa alami, komputasi linguistic dan text mining yang bertujuan menganalisa pendapat, sentiment, evaluasi, sikap, penilaian, dan emosi seseorang [2].
Untuk melakukan proses jual beli saat ini tidak hanya menggunakan cara konvensional seperti pergi ke pasar tradisional, akan tetapi dapat dilakukan secara elektronik oleh konsumen, dari perusahaan ke perusahaan menggunakan Komputer sebagai media perantara transaksi tersebut. Bukalapak merupakan salah satu matketplace yang sangat popular di Indonesia. Bukalapak memberikan penggunanya untuk dapat melakukan transaksi secara cepat dan aman. Respon yang diberikan oleh pengguna bukalapak dapat berupa respon positif maupun negatif. Oleh karena itu diperlukan suatu metode yang dapat digunakan untuk mengetahui opini pengguna bukalapak pada sosial media twitter.
Untuk menyelesaikan masalah tersebut maka akan dilakukan penelitian mengenai text mining pada sosial media terhadap bukalapak. Dengan penelitian ini dapat menggali opini atau pendapat pengguna bukalapak yang di posting di sosial media apakah itu bersifat positif, negatif maupun netral.
Untuk mewujudkan hal tersebut penulis membuat penelitian dengan judul "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter" untuk melakukan klasifikasi opini menggunakan metode support vector machine ini mempertimbangkan pada penelitian-penelitian sebelumnya, menyatakan metode Support Vector Machine memiliki akurasi lebih besar dari metode Naïve Bayes Classifier [3].

Tinjauan Pustaka

Data Mining

Data Mining merupakan gabungan sejumlah disiplin ilmu computer, yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode-metode yang merupakan irisan dari artificial intelligence, machine learning, statistic, dan database system.
Secara umum, kegunaan data mining dapat dibagi menjadi dua: deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data. Sedangkan prediktif berarti data mining digunakan untuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi.

Analisa Sentimen

Analisis Sentimen (*sentiment analysis*) dikenal juga sebagai ekstraksi opini, penambangan opini, penambangan sentimen, dan analisis subjektivitas. Analisis sentimen merupakan studi komputasional dari opini-opini orang, appraisal dan emosi melalui entitas, event dan atribut yang dimiliki [2].
Secara istilah, analisis sentimen adalah deteksi sikap-sikap (*attitudes*) terhadap objek atau orang. Dari miliaran data cuwitan di twitter, dapat dilakukan analisis sentimen untuk menemukan berapa presentase sentimen positif dan berapa presentase sentimen negative terhadap seseorang, perusahaan, institusi, kelompok, atau sebuah situasi tertentu.

Text Mining

Text mining dapat didefinisikan secara luas sebagai suatu proses menggali informasi dimana seorang user berinteraksi dengan sekumpulan dokumen menggunakan tools analisis yang merupakan komponen-komponen dalam data mining yang salah satunya adalah kategorisasi. Text mining bisa dianggap subjek riset yang tergolong baru. Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian / pengelompokkan dan menganalisa *unstructured text* dalam jumlah besar [1].

Preprocessing

Preprocessing merupakan tahap yang dilakukan sebelum proses pengklasifikasian sentimen. Proses ini dilakukan untuk mengolah dataset agar siap digunakan untuk tahap selanjutnya. Tahapan ini diperlukan untuk mendapatkan data yang berkualitas dan meningkatkan proses penggalian informasi diantaranya dengan penanganan pada incomplete data, noisy text dan inconsistent data menurut buku *Data Mining: Concept and Techniquest* [4].

Part Of Speech Tagging (POS Tagging)

Part-of-speech (POS) Tagging atau biasa juga disebut dengan *grammatical tagging* adalah proses untuk melabeli kata berdasarkan *part-of-speech* didalam sebuah kalimat [5]. Beberapa metode yang digunakan dalam *POS-Tagging* ini adalah metode *statistic* yang terdiri dari Hidden Markov Model, Maximum Entropy, dan Conditional Random Field.

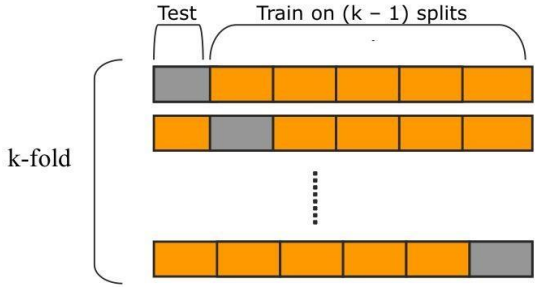
Support Vector Machine

Support Vector Machine (SVM) diperkenalkan oleh Vapnik pada tahun 1992 sebagai suatu teknik klasifikasi yang efisien untuk masalah nonlinear. *Support Vector Machine* (SVM) juga dikenal sebagai teknik pembelajaran mesin (machine learning) paling mutakhir setelah pembelajaran mesin sebelumnya yang dikenal sebagai *Neural Network* (NN). Baik SVM maupun NN tersebut telah berhasil digunakan dalam pengenalan pola. Pembelajaran dilakukan dengan menggunakan pasangan data input dan data output berupa sasaran yang diinginkan. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada input space. SVM berusaha menemukan fungsi pemisah (*hyperplane*) dengan memaksimalkan jarak antar kelas. Dengan cara ini, SVM dapat menjamin kemampuan

generalisasi yang tinggi untuk data-data yang akan dating [6].

Validasi dan Evaluasi

Proses validasi dilakukan menggunakan *10-fold cross validation* dimana data awal akan dipartisi menjadi 10 segment fold yaitu K1, K2, K3, …..,Kn dengan ukuran yang sama. Proses training dan testing data dilakukan sebanyak n kali dimana pada setiap iterasi ken, segment Kn akan menjadi data *testing* sedangkan yang lain menjadi data *training*. Proses tersebut akan dilakukan sebanyak n kali dengan segment dan akan menjadi data testing tepat satu kali dan akan menjadi data *training* [7].



Gambar 1. Ilustrasi *K-fold Cross Validation*

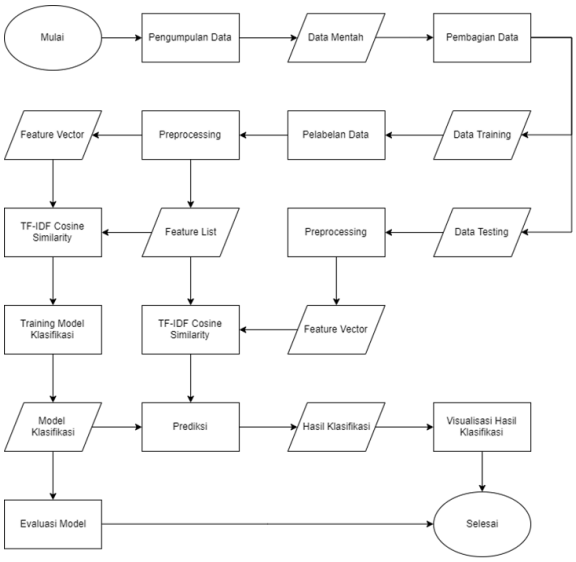
Pengujian sistem menggunakan metode *confusion matrix* untuk pengukuran *precision, Recall, dan Accuracy*. Tabel 1 menunjukkan table *confusion matrix* yang akan digunakan untuk perhitungan.

Tabel 1. *Confussion Matrix*

	Predicted Positive	***Predicted Negative***
Actual Positive	*Number of True Positives Intance (TP)*	*Number of False Negatives Intance (TP)*
Actual Negative Instance	*Number of False Positive Instances (FP)*	*Number of True Negatives Instances (FP)*

Metode Penelitian

Pada penelitian ini, sistem yang dibuat mampu untuk menganalisis sentimen yang terdapat pada dokumen yang diberikan kadalam sistem. Analisis pada penelitian ini diambil dari kumpulan tweet yang disimpan dalam sebuah dokumen. Kumpulan tweet yang digunakan untuk analisis tersebut nantinya juga akan digunakan sebagai data training dengan sebuah label dan data testing tanpa label.



Gambar 2. Diagram Alir Metode Penelitian

Adapun penjelasan terkait dari Gambar 2 :

1. Pengumpulan data tweet menggunakan Twitter Scraper yaitu salah satu library yang dimiliki oleh bahasa pemrogramman Python. Data Twitter diambil dengan rentang waktu satu bulan dan disimpan dalam bentuk .csv. Data tweet yang telah dikumpulkan akan diseleksi mana yang cocok untuk dijadikan data training maupun data prediksi karena tidak semua data yang telah dikumpulkan dapat digunakan.
2. Pembagian data merupakan proses pembagian tweettweet yang sudah dikumpulkan. Data training berbeda dengan data prediksi. Setelah data dibagi, untuk proses training dilakukan proses pelabelan sentimen secara manual. Sedangkan untuk data prediksi, tidak dilakukan pelabelan sentimen.
3. Pada data testing ataupun data training dilakukan preprocessing. Proses dari preprocessing meliputi Tweet cleaning, tokenizing, Normalisasi Kata, stopword remove, POS Tagging, POS Filtering, dan stemming. Setelah preprocessing pada proses training akan meghasilkan feature list yang merupakan sebuah daftar kata-kata berlabel positif,negatif, atau netral dari data training. Feature List ini yang akan digunakan dalam proses prediksi.
4. Setelah menghasilkan feature vector yang merupakan kata-kata. Masing-masing kata itu akan dirubah ke dalam bentuk binary. Dalam penelitian ini penulis menggunakan metode Cosine Similarity antara bobot TF-IDF suatu kata yang terdapat tweet dengan feature list. Hasil dari metode ini adalah sebuah bobot yang mengindikasikan kedekatan tweet tersebut dengan feature list yang diberikan. Bobot ini nantinya akan diklasifikasikan oleh Support Vector Machine yang ada pada library Scikit-Learn.

5. Setelah mendapatkan bobot dari hasil perhitungan Cosine Similarity akan dilakukan proses training yang nantinya akan menghasilkan model klasifikasi yang diperlukan dalam proses prediksi. Model ini juga nantinya akan dievaluasi tingkat akurasi, precission,dan recall dengan metode cross validation.
6. Setelah model terbentuk maka akan digunakan untuk proses predksi sentimen dengan tweet yang didapatkan langsung menggunakan twitterscraper. Kemudian tweet tersebut akan melalui proses preprocessing dan kemudian di hitung bobotnya menggunakan TF-IDF Cosine Similarity. Kemudian bobot yang ada akan dimasukkan ke model prediksi yang telah dibuat. Lalu hasil dari prediksi tersebut akan ditampilkan persentase dan data tweetnya.

Hasil dan Pembahasan

Analisis Sentimen dengan Algoritma *Support Vector Machine*

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah sebuah tweet yang diambil dari Twitter dengan hastag Telkomsel. Pengambilan data menggunakan Twitter Scraper dan disimpan dalam bentuk .csv. Dalam tugas akhir ini, peneliti mengambil tweet-tweet yang berhubungan atau di tujukan kepada marketplace Bukalapak.

B. Pelabelan Tweet

Setelah dataset dikumpulkan, data tersebut akan diberikan sebuah label positif, negatif, dan netral. Dalam penelitian ini, peneliti pertama kali memilih secara random data yang akan digunakan didalam penelitian ini.

Untuk mendukung validitas penelitian, peneliti meminta bantuan dari ahli untuk melabeli 300 tweet. Ahli merupakan seorang yang biasa menangani tingkah laku dan mental. Disini ahli adalah seorang psikolog yang bekerja sebagai HRD di sebuah perusahaan.

C. Preprocessing

Sebelum dilakukan analisa sentimen dari data tweet yang telah diambil, perlu dilakukan proses pengolahan data supaya siap digunakan untuk analisa sentimen.

Tweet mentah yang didapat dari twitter akan diproses terlebih dahulu sebelum nantinya digunakan untuk analisis. Proses tersebut dimulai dari tweet cleaning hingga nantinya menghasilkan term. Term inilah yang nantinya akan diberikan pembobotan.

1. Tweet Cleaning yaitu data tweet akan dilakukan pembersihan. Pembersihan tweet meliputi : menghilangkan URL, menghilangkan emoticon, menghilangkan symbol, menghilangkan @username, menghilangkan hashtag, proses penyeragaman bentuk huruf menjadi huruf kecil.
2. Tokenize yaitu langkah membagi kalimat tweet menjadi kata-kata yang menyusun kalimat tersebut. Tahap ini digunakan untuk memudahkan analisa kata-kata tersebut lebih lanjut.
3. Normalisasi kata yaitu pengubahan kata singkatan maupun kata tidak baku menjadi kata baku.
4. Remove stopword yaitu menghilangkan katakata yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Contoh stopword dalam bahasa Indonesia adalah "yang", "dan", "di", "dari", dll.
5. POS tagging yaitu memberikan label tag terhadap masing-masing kata yang telah ditokenize. Tag ini berfungsi untuk mengetahui peranan masing-masing kata didalam kalimat tweet yang akan diproses.
6. POS filtering yaitu menyeleksi fitur yang akan dimasukan sebagai list fitur yang akan digunakan sebagai data training. Tahap ini disebut juga tahap feature selection dengan menggunakan metode filtering.
7. Stemming yaitu proses pencarian kata dasar dari suatu kata. Kata yang memiliki imbuhan seperti ber-, per-, di-, ke-, mem-, men-, -an, -kan, dan lain-lain akan dihilangkan.

D. Pembobotan TF-IDF

Dalam tahapan ini, kata-kata yang tersusun dalam masing-masing tweet akan diberikan bobot dengan cara mengalikan nilai Term Frequency (TF) dengan Inverse Document Frequency (IDF). Pengukuran bobot dilakukan dengan mencari kemunculan masing-masing kata yang terdapat dalam tweet pada feature list positif, negatif, dan netral. Tweet tersebut akan digunakan sebagai data latih dan juga data uji (data latih dan data uji memiliki daftar tweet yang sama untuk contoh implementasi ini).

Menurut Salton dalam bukunya nilai IDF didapatkan dengan persamaan [8]:

$$IDF = \log(\frac{D}{dfi})$$

D adalah jumlah dokumen serta dfi adalah jumlah kemunculan term terhadap D. Sedangkan untuk perhitungan bobot *TF-IDF*(W) untuk setiap dokumen terhadap kata kunci yaitu:

$$w_{ij} = tf_{ij}\ x\ log(\frac{D}{dfi})$$

E. Cosine Similarity

Cosine Similarity merupakan metode yang digunakan untuk mengukur kemiripan suatu tweet terhadap feature list yang akan diberikan. Dalam penelitian ini, terdapat 3 dokumen feature list, masingmasing yaitu positif, negatif, dan netral. Nilai yang dihasilkan dari Cosine Similarity adalah nilai diantara 0 sampai 1. Semakin mirip suatu tweet dengan feature list yang dibandingkan, angkanya

akan mendekati 1. Sebaliknya, semakin tidak miripnya suatu tweet, angkanya akan mendekati 0. Adapun rumus yang digunakan untuk menghitung cosine similarity adalah sebagai berikut:

$$\cos(a,b) = \frac{a \,.\, b}{|a||b|}$$

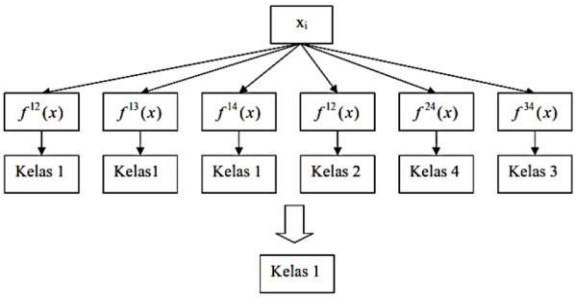
$$\cos(Query, FeatureList) = \frac{\sum(TFIDF_{term} * IDF_{term})}{\sum|TFIDF_{term}| * \sum|IDF_{term}|}$$

F. Multi-Class Support Vector Machine

Support Vector Machine dapat melakukan klasifikasi lebih dari dua kelas. Sebagai contoh dalam menentukan sentimen suatu tweet. Ada tiga sentimen, yaitu positif, negatif, dan netral. Beberapa pendekatan/strategi yang umum digunakan dalam menangani klasifikasi terhadap multi-class yaitu :

1. One Againts One

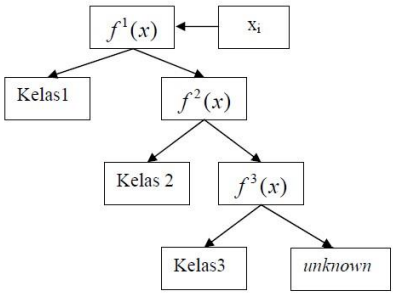
Strategi "one against one", juga dikenal sebagai "pairwise coupling", "all pair" atau "round robin", yaitu membuat satu classifier (penggolong) terhadap masingmasing pasangan kelas. Jadi banyaknya jumlah classifier dapat diketahui dari dimana N adalah jumlah kelas. Lalu, classifier dilatih untuk membedakan sampel satu kelas dari sampel kelas lain. Untuk mengantisipasi klasifikasi terhadap pola yang tidak diketahui, akan dilakukan dengan vote maksimum, dimana masing-masing classifier menghasilkan satu vote dan nantinya vote tersebut akan di akumulasikan untuk memilih kelas.



Gambar 3. Ilustrasi *One Againts One*

2. One Againts All

Strategi "one against all" terdiri dari satu classifier per kelas, yang dilatih untuk membedakan sampel satu kelas dari sampel semua kelas yang tersisa. Biasanya, klasifikasi pola yang tidak diketahui dilakukan sesuai dengan output maksimum di antara semua SVM.



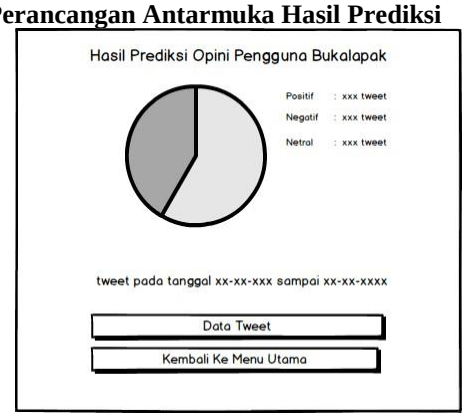
Gambar 4. Ilustrasi One Againts All

Perancangan Antarmuka Pengguna

A. Perancangan Antarmuka Halaman Utama Prediksi

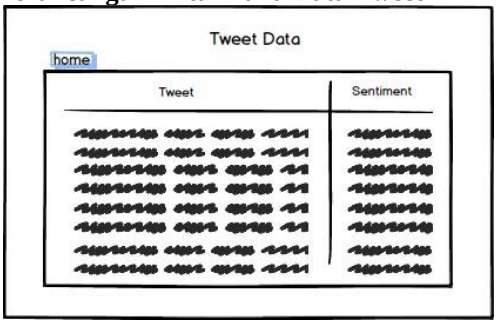
Gambar 5. Rancangan Halaman Utama Prediksi

B. Perancangan Antarmuka Hasil Prediksi



Gambar 6. Rancangan Halaman Hasil Prediksi

C. Perancangan Antarmuka Data Tweet



Gambar 7. Rancangan Halaman Data Tweet

Kesimpulan dan Saran

Kesimpulan

Berdasarkan penelitian yang telah dilakukan, kesimpulan yang didapatkan adalah :

1. Pengklasifikasian dilakukan dengan melewati tahap Tweet cleaning, tokenizing, Normalisasi Kata, stopword remove, POS Tagging, POS Filtering, dan stemming. Setelah tahap preprocessing, data akan melalui proses TF-IDF Cosine Similarity yang akan menghasilkan nilai bobot yang digunakan untuk proses klasifikasi dengan metode Support Vector machine.
2. Klasifikasi sentimen dengan algoritma Support Vector Machine menghasilkan nilai akurasi tertinggi sebesar 93% jika diuji menggunakan data uji yang didapatkan pada saat pengumpulan data.
3. Evaluasi terhadap algoritma Support Vector Machine dengan K-Fold Cross Validation didapatkan akurasi tertinggi sebesar 93 % dan rata-rata sebesar 83.5 %.

Saran

Penelitian ini masih terdapat beberapa keterbatasan dan kekurangan yang dapat menjadi acuan bagi penelitian dan pengembangan selanjutnya. Adapun saran yang didapat dari penelitian ini adalah:

1. POS Tagging pada penelitian ini berasal dari korpus yang terbentuk dari kalimat-kalimat baku, sedangkan di dalam tweet banyak ditemukan kata yang tidak baku. Sehingga dibutuhkan korpus yang berasal dari pola-pola kalimat yang tidak baku.
2. Penelitian ini masih meneliti sebatas teks pada tweet saja. Pengembangan penelitian bisa dilakukan dengan menambahkan variabel-variabel lainnya seperti emoji/emotikon dan gambar.
3. Perlu penambahan kamus untuk kata tidak baku agar semakin banyak kata tidak baku yang tersaring dan dokumen menjadi semakin bersih.

Demikian Saran yang dapat penulis berikan, semoga saran tersebut dapat dijadikan sebagai bahan masukan. Sehingga dapat bermanfaat khususnya bagi penulis dan umumnya bagi masyarakat luas.

Daftar Pustaka

[1] R. Feldman dan J. Sanger , "The text mining handbook: advanced approaches in analyzing unstructured data," Cambridge university press, 2007.

[2] B. Liu, "Sentiment Analysis and Subjectivity," Handbook of Natural Language Processing, vol. 2, pp. 627-666, 2010.

[3] Y. Marchel dan J. Nasri, "Perbandingan Tingkat Akurasi Support Vector Machine Dengan Naive Bayes Pada Studi Kasus Okupansi Lahan Berdasarkan Kondisi Cuaca," eProceedings of Engineering,, vol. 20, 2017.

[4] J. Han, M. Kamber dan J. Pei, Data mining: concepts and techniques, Elsevier, 2011.

[5] R. E. Setyaningsih, "Part of Speech Tagger Untuk Bahasa Indonesia Dengan Menggunakan Modifikasi Brill," *Dinamika Teknologi*, vol. 9, no. 1, p. 37, 2017.

[6] Suyanto. Data Mining Untuk Klasifikasi dan Klasterisasi Data, Bandung: Informatika Bandung, 2017.

[7] J. Friedman, T. Hastie dan R. Tibshirani, The elements of statistical learning (Vol. 1), Springer, Berlin: Springger Series in Statistics, 2001.

[8] G. Salton, Automatic text processing: the transformation, analysis, and retrieval of, Addison-Wesley , 1989.