STUDI KOMPREHENSIF ALGORITMA NAÏVE BAYES CLASSIFIER DENGAN SUPPORT VECTOR MACHINE PADA SENTIMENT ANALYSIS OPINI PEMBANGUNAN INFRASTRUKTUR DI MEDIA SOSIAL TWITTER

Dina Maulina[1], Ignasius Frans[2]

Informatika, Universitas AMIKOM Yogyakarta[1]
Manajemen Informatika, Universitas AMIKOM Yogyakarta[2]

Email : dina.m@amikom.ac.id [1], ignasius.dstn@students.amikom.ac.id [2]

ABSTRAK

Pada saat ini media sosial sebagai tempat untuk beropini bisa di pandang juga sebagai tempat penyimpanan data yang memiliki pola tertentu. Sentimen analisis akan menggunakan model teertentu untuk menghasilkan informasi yang lebih terproses agar data yang di sajikan bisa mensimpulkan pola yang berkembang di masyarakat. Penelitian ini akan menyimpulkan perbandingan hasil akurasi dari NBC dan SVM serta melihat pola opini yang berkembang di masyarakat terhadap pembangunan infrastruktur yang di jalankan oleh pemerintah indonesia. Tujuan dari penelitian ini adalah menganalisa perbandingan model pembelajaran dari dua algoritma yaitu *Support Vector Machine* (SVM) dan *Naïve Bayes Classifier* (NBC) yang di implementasikan kepada objek sentimen masyarakat Indonesia di media sosial *Twitter*. Selama ini telah di ketahui bahwa kebanyakan sentimen analisis yang telah di lakukan hanya menggunakan model pembelajaran tertentu. Diperlukan pembanding terhadap model lain guna menemukan nilai akurasi terbaik dari setiap model yang di pilih. Berdasarkan hasil pengujian didapatkan akurasi dari Naïve Bayes Classifier dan Support Vector Machine untuk klasifikasi sentiment analysis di dapat melalui proses pembelajaran terhadap model yang di buat dan menghasilkan nilai akurasi sebesar 85.6 % untuk NBC dan 86,21 % untuk SVM.

Kata Kunci : Sentimen Analisis, Naïve Bayes Classifier, Support Vector Machine, Pembangunan Infrastruktur, Klasifikasi.

ABSTRACT

At this time social media as a place for opinion can also be seen as a place to store data that has a certain pattern. Sentiment analysis will use certain models to produce more processed information so that the data presented can conclude the patterns that develop in society.This study will conclude the comparison of accuracy results from NBC and SVM and see the pattern of opinion that develops in the community towards infrastructure development carried out by the Indonesian government. The purpose of this study is to analyze the comparison of the learning models of the two algorithms, namely the Support Vector Machine (SVM) and the Naïve Bayes Classifier (NBC) which are implemented to the object of Indonesian public sentiment on social media Twitter. So far, it has been known that most of the sentiment analysis that has been carried out only uses certain learning models. Comparison with other models is needed in order to find the best accuracy value for each selected model. Based on the test results, the accuracy of the Naïve Bayes Classifier and Support

Vector Machine for sentiment analysis classification can be obtained through the learning process of the model created and produces an accuracy value of 85.6% for NBC and 86.21% for SVM.

Keywords: *Sentiment Analysis, Naïve Bayes Classifier, Support Vector Machine, Infrastructure Development, Classification.*

1. PENDAHULUAN

Berdasarkan kajian pustaka yang dilakukan oleh peneliti dari berbagai macam algoritma yang di gunakan di dalam sentimen analisis belum di temukan model paling tepat untuk menganalisa sentimen terhadap objek pembangunan infrastruktur di Indonesia. Maka dari itu peneliti melakukan komparasi terhadap dua buah algoritma klasifikasi yaitu support vector machine dan naïve bayes classifier untuk mendapatkan nilai akurasi terbaik dari analisa objek tersebut.

Metode yang digunakan dalam penelitian kali ini yaitu *Naïve Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM) dimana akan di lakukan perbandingan hasil terhadap kedua algoritma tersebut pada obyek yang akan di teliti. Hasil akhir yang di pusatkan pada penelitian kali ini adalah untuk memperoleh metode mana yang lebih efektif untuk mendapatkan hasil akurasi pengolahan data. Di harapkan dari penelitian ini dapat menghasilkan suatu informasi yang berkualitas untuk pengambilan keputusan dalam suatu skenario.

Beberapa penelitian yang telah di lakukan sebelumnya tentang sentimen analisis diantaranya adalah penelitian yang membahas tentang analisis sentimen pasar otomotif mobil pada *tweet* Twitter dengan metode *Naïve Bayes* dimana pada penelitian tersebut bertujuan untuk mengetahui merek mobil terlaris di Twitter. Hasil penelitian ini menunjukkan bahwa tingkat akurasi *Naïve Bayes* yaitu 93% (Rustiana, D., & Rahayu, N., 2017). Penelitian kedua yaitu sentimen analisis *tweet* berbahasa Indonesia dengan *Deep Belief Network* (DBN) dan hasil DBN dibandingkan dengan Naïve Bayes. Tujuan penelitian tersebut adalah untuk mengetahui hasil sentimen terhadap *tweet* berbahasa Indonesia di Twitter dan hasil akurasi DBN sebesar 93.31%, Naïve Bayes sebesar 79.10 (Zulfa, I., & Winarko, E., 2017).

Data Mining merupakan kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan , pola dan hubungan dalam dataset berukuran besar (Sugianto,2015). Dalam *data mining* kita juga mengenal *Text mining* yang mengacu pada proses pengambilan informasi berkualitas tinggi dari sebuah atau beberapa *sample text*. Text mining sendiri merupakan suatu proses penambangan data berupa teks yang di lakukan oleh komputer dimana data tersebut dapat memberikan informasi-informasi untuk dilakukan analisa keterhubungannya % (Rustiana, D., & Rahayu, N., 2017). Pembelajaran sebuah pola statistik bisa digunakan untuk menemukan informasi berkualitas tinggi melalui peramalan pola dan kecendrungan sarana pada sebuah *text*. Proses umum pada sebuah *text mining* biasanya meliputi pengelompokan *text*, *text clustering*, pengambilan intisari konsep/entitas, produksi taksonomi granular, *sentiment analysis*, penyimpulan dokumen, serta permodelan relasi entitas (pembelajan hubungan antara entitas berlabel).

Sentiment Analysis dalam *text mining* adalah bidang studi yang menganalisis pendapat seseorang, *sentiment* seseorang, evaluasi seseorang, sikap seseorang dan emosi seseorang ke dalam bahasa tertulis (Saputra, N. Bharata, T. Erna, A., 2015). Studi *Analysis Sentiment* sendiri dapat digunakan untuk membantu kita mengambil keputusan paling tepat guna menyelesaikan satu atau lebih skenario yang kita temukan. Dalam skenario kali ini peneliti juga membagi *sentiment analysis* menjadi tiga *class attribute*. Dimana ada label positif, label negatif, dan label netral pada setiap classnya. Berdasarkan latar belakang yang telah dikemukakan diatas maka dirumuskan masalah bagaimana SVM dan NBC dapat

mengklasifikasikan *tweet* dan memberikan nilai akurasi yang bisa di gunakan dalam perbandingan algoritma yang di pilih, sedangkan pembahasan masalah akan dibatasi pada *Tweet* yang diambil dan digunakan merupakan *tweet* yang berbahasa Indonesia. Penelitian mengambil objek berupa opini masyarakat tentang pembangunan infrastruktur yang telah dilakukan di Negara Indonesia. Penelitian dilakukan pada bulan Desember 2018 sampai Juni 2019 dengan data *tweet* berjumlah 437 *tweet* yang di ambil menggunakan metode *crawling* pada microbloging *Twitter*. *Tweet* yang di gunakan berupa *text* dan tidak mengandung gambar sedangkan Algoritma yang di gunakan untuk mengukur akurasi pada hasil pengolahan data kali ini adalah *Naïve Bayes Classifier* dan *Support Vector Machine*.

Tujuan yang di kedepankan dalam penelitian kali ini adalah mengetahui proses dan alur dari memperoleh data pada microbloging *Twitter* untuk *Sentiment Analysis* hingga peneliti dapat mengolah data tersebut guna mendapatkan nilai akurasi yang terbaik. Nilai akurasi di ukur melalui dua algoritma yaitu *Naïve Bayes Classifier* dan *Support Vector Machine*. Setelah mendapatkan nilai akurasi akan dilakukan perbandingan dan *analysis* terhadap nilai akurasi yang di peroleh.

2. LANDASAN TEORI

Penelitian yang digunakan sebagai tinjauan pustaka dalam penelitian ini diantaranya adalah:

Penelitian yang dilakukan oleh Siti Mujilahwati pada tahun 2016 dapat disimpulkan bahwa ujicoba klasifikasi dengan menggunakan metode *Naïve Bayes* bisa mendapatkan hasil nilai akurasi sebesar 93,11% dari hasil *preprocessing* yang telah di lakukan. Klasifikasi di atas di anggap masih kurang baik bn karena masih adanya data *mention* terhadap *customer support* (Mujilahwati,S., 2016)

Penelitian pengklasifikasian *Sentiment Analysis* sendiri juga pernah di lakukan pada implementasikan dalam pengklasifikasian SMS, menghasilkan suatu nilai akurasi dengan menggunakan metode *Naïve Bayes* yaitu 95% sedangkan *Support Vectro Machine* sebesar 76% dan C4.5 sebesar 95,5%. Dari hasil ini bisa di lihat bahwa yang paling tinggi nilai akurasinya adalah C4.5 dengan nilai akurasi sebesar 95,50% (Sari, R., 2017).

Penelititan oleh Elly Indrayuni pada tahun 2018 *[7]* memberikan hasil ekperimen pada penggunaan metode Naïve Bayes sebesar 84,50% untuk nilai akurasinya sedangkan SVM mendapatkan 90.00% nilai akurasi untuk klasifikasi sentimen review film.

Selain beberapa penelitian sebelumnya menggunakan metode tersebut sebagai solusi dalam penelitian yang terdahulu oleh karena itu pada penelitian ini menyusulkan metode super vector mesin sebagai bagian salah satu solusi.

3. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini antara lain adalah;

a. Metode Pengumpulan Data

   Dalam metode ini penulis mencoba mendapatkan data dari microbloging *Twitter* dengan cara crawling data menggunakan API *Twitter* dari akun-akun acak berdasarkan *keyword* ‘pembangunan infrastruktur’.

b. Metode Pengolahan Data

   terdiri dari beberapa bagian tahapan yang memiliki masing-masing metode untuk *Analysis Sentiment*. Hal pertama yang di lakukan adalah melakukan *crawling* data dari *Twitter* menggunakan API yang telah disediakan oleh *Twitter* sendiri. Selanjutnya data tersebut di simpan dalam bentuk *file document*, lalu data yang telah di proses di bagi menjadi dua tipe data yaitu *data training* dan *data testing*. Berikut tahapan selanjutnya dalam data pengolahan *data training* dan *data testing* :

1. Pelabelan Manual
2. Tokenizing
3. Stopword Removal
4. Stemming

c. Metode Analisis

Penelitian ini menganalisis setiap data yang di peroleh melalui cara kerja algoritma *Naïve Bayes Classifier* dan *Support Vector Machine* dalam mengklasifikasikan setiap data *tweet* yang di peroleh dari *crawling data* itu sendiri

Alur penelitian ini dimulai dari Pengambilan data mentah berupa *tweet* dari media sosial *twitter* dilakukan dengan menggunakan pemrograman R dari aplikasi R- studio. Kemudian Data mentah dibagi menjadi dua bagian data, yang pertama adalah 80 % *data training* (350 data) yang di ambil dari keseluruhan data mentah dan 20 % (87 data) untuk *data testing*. Setelah pembagian data mentah,penelitian di lanjutkan dengan pelabelan secara manual terhadap *data tweet* yang telah di *crawling* menjadi *sentiment* positif dan *sentiment* negatif. Proses selanjutnya adalah *preprocessing* data. Di dalam *preprocessing* sendiri terdiri dari beberapa subproses berupa *tokenizer*, *stopword removal*, dan *stemming* guna menghilangkan kata kata yang tidak perlu. Pada *data testing* juga akan di lakukan tahapan *preprocessing* guna menghilangkan setiap kata yang tidak perlu dalam kalimat *tweet* yang telah di peroleh. Langkah selanjutnya adalah melakukan penghiutangan klasifikasi menggunakan NBC dan SVM pada *data training* dan *data testing* yang telah melalui proses *preprocessing*. Membuat *confusion matriks* dari masing-masing model guna menemukan nilain akurasi yang diinginkan. Terakhir adalah membandingkan nilai akurasi untuk menemukan model terbaik berdasarkan nilai akurasi.

Pada Penelitian kali ini telah di lakukan pengambilan data sebanyak 437 data tweet. Data tersebut di rincikan sebagai dataset awal yang akan di gunakan untuk melakukan klasifikasi sentiment analysis.

4. HASIL DAN PEMBAHASAN

Confusion Matriks Pada Naïve Bayes Classifier

Hasil yang di peroleh dari pengolahan data di atas yang melalui proses preprocessing dihitung dengan perhitungan akurasi menggunakan *confusion matriks*. Akurasi yang di peroleh adalah sebesar 85.0574713 % dengan perhitungan akurasi sebagai berikut :

$$Akurasi = \frac{TN + TP}{TN + FN + FP + TP} * 100\%$$
$$= \frac{1 + 73}{1 + 2 + 73 + 11} * 100\%$$
$$= 85.0574713\%$$

Perhitungan akurasi di atas berdasarkan nilai *confusion matriks* yang di peroleh dari data uji dengan nilai *true negatif* sebanyak 1, *true positif* sebanyak 73, *false negatif* sebanyak 2, dan *false positif* sebanyak 11.

Confusion Matrik Pada Support Vector Machine

Hasil pengolahan data yang telah melalui proses *preprocessing* maka dari pada itu di masukkan kedalam permodelan menggunakan algoritma *Support Vector Machine* adalah sebesar 86.2068966 % dengan perincian perhitungan nilai akurasi seperti
berikut seperti berikut:

$$Akurasi = \frac{TN + TP}{TN + FN + FP + TP} * 100\%$$
$$= \frac{0 + 75}{0 + 12 + 0 + 75} * 100\%$$
$$= 86.2068966\%$$

Pada *Support Vector Machine* nilai *confusion matrik* yang di dapat dari pengolahan

preprocessing pada data uji adalah sebanyak 0 pada nilai *true negatif*, 75 pada nilai *true positif*, 12 pada nilai *false negatif*, dan 0 pada nilai *false positif*.

Pada proses pembelajaran dan analisis dengan menggunakan metode Naïve Bayes Classifier dan Support vector Machine menggunakan dataset yang sama berupa hasil crawling twitter untuk kedua model pembelajaran. Pada penelitian ini dataset yang digunakan memiliki objek berupa pembangunan infrastrutur di Indonesia sehingga keyword yang dipakai adalah "pembangunan infrastrutktur". Dokumen hasil dari crawling tweet pada twitter berjumlah 437 tweet. Sebanyak 87 tweet dimasukkan kedalam data testing dan 350 tweet dimasukkan kedalam data training. Peneliti juga membagi class pada sentiment menjadi 2 kategori yaitu positif dan negatif. Berikut hasil pengujian dengan menggunakan metode NBC dan

Accuracy 85.06%	True positif	True negatif	Class Precision
Pred. Positif	73	11	86.90%
Pred. Negatif	2	1	33.33%
Class Recall	97.33%	8.33%	

Gambar 1. Akurasi NBC

Tabel 1 yang memuat hasil perbandingan nilai akurasi yang di dapat dari proses klasifikasi *sentiment analysis*

Tabel 1 Perbandingan Nilai Akurasi

No	Metode	Nilai Akurasi
1	Naïve bayes Classifier	85.06 %
2	Support Vector Machine	86.21 %

Dari hasil perhitungan nilai akurasi yang di tunjukan perbandingannya pada tabel di atas maka *support vector machine* lebih unggul 1.15 % dari pada *naïve bayes classifier* walaupun nilai perbandingan yang di hasilkan belum signifikan karena data yang di olah tidak banyak. Dari pengujian di atas dapat dilihat bahwa *support vector machine* lebih baik unjuk kerjanya dari pada *naïve bayes classifier* dalam hal klasifikasi *sentiment analysis*.

5. SIMPULAN DAN SARAN

Berdasarakan uraian sebelumnya maka dapat di simpulkan bahwa Akurasi dari Naïve Bayes Classifier dan Support Vector Machine untuk klasifikasi sentiment analysis di dapat melalui proses pembelajaran terhadap model yang di buat dan menghasilkan nilai akurasi sebesar 85.6 % untuk NBC dan 86,21 % untuk SVM. Hasil akurasi SVM memiliki unjuk kerja 1.15 % lebih baik dari pada NBC untuk pengklasifikasian sentiment analysis. Dari hasil penelitian yang di lakukan terhadap objek *dataset* berupa *tweet* yang bertemakan pembangunan infrastruktur Indonesia dapat disimpulkan bahwa SVM lebih baik nilai akurasinya dalam melakukan pengklasifikasian *Sentiment Analysis* di bandingkan dengan NBC

6. UCAPAN TERIMA KASIH

Penulis haturkan terimakasih kepada segenap pihak yang telah berperan serta dalam terwujudnya penulisan publikasi penelitian ini. Kami haturkan trimakasih kepada LPPM Universitas AMIKOM Yogykarta yang memfasiliitasi tercapainya publikasi ini. Terima kasih kami haturkan juga kepada Jurnal Inspiration yang telah menampung publikasi dari penelitian kami ini.

DAFTAR PUSTAKA

Han, J., & Kamber, M. (2006). "*Data Mining Concepts and Techniques"*. Buku. San Fransisco: Elsevire.7 & 641

Indrayuni, E. (2018). "Komparasi Algoritma *Naïve Bayes* dan *Support Vector Machine* Untuk Analisa *Review* Film". Jurnal. PILAR Nusa Mandiri Vol.14, No.2. September 2108.

Larose, D.T. (2005). "Discovering Knowledge in Data". Canada: Wiley-Interscience",. 2. Buku.

Lestari, A. Perdana, R. Fauzi, M. (2017). "Analisis Sentimen Tentang Opini Pilkada DKI 2017 Pada Dokumen Twitter Berbahasa Indonesia Menggunakan Naïve Bayess Classfier dan Pembobotan Emoji". Jurnal

Maulina, D. Sagara, R. (2018). Klasifikasi Artikel Hoax Menggunakan *Support Vector Machine Linear* Dengan Pembobotan *Term Frequency-Inverse Document Frequency*. Jurnal. Ilmu Komputer Univesitas AMIKOM Yogyakarta.

Mujilahwati, S. (2016). Pre-processing *Text Mining* Pada *Twitter*. Jurnal. Teknik Informatika, Fakultas Teknik, Universitas Islam Lamongan.

Pengembangan Teknologi Informasi dan Ilmu Komputer Vol.1, *No*.12. Desember (2017). 1718-1724.

Pudjajana, A. M., Manongga, D. (2018). "Sentimen Analisis Tweet Pornografi Kaum *Homoseksual* Indonesia Di Twitter Dengan Naïve Bayes". Jurnal. SIMETRIS, Vol.9 No.1, April 2018.

Sugianto, C.A. (2015). "Analisis *Komparasi* Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan". Jurnal. *Teknil* Informatika. Politeknik TEDC Bandung.

Rustiana, D., and Rahayu, N., 2017. "Analisa Sentimen Pasar Otomotif Mobil: Tweet Twitter menggunakan Naïve Bayes", *Jurnal Simetris* 8. 1, 113

Susanto, C. P., Setiyawan, E. I. (2015). "Algoritma *Support Vector Machine* Untuk Mendeteksi SMS Spam Berbahasa Indonesia". Jurnal. Seminar Nasional "Inovasi dalam desain dan Teknologi:, 109-116.

Sari, R. (2017). "Komparasi Algoritma *Support Vector* Machine, *Naïve Bayes Classifier*, dan C4.5 Untuk Klasifikasi SMS". Jurnal. *Indonesian Journal on Computer and Information Technology, Vol.2, No.2, November 2017.*

Saputra, N. Bharata, T. Erna, A. (2015). "Analisis Sentimen Data Presiden Jokowi Dengan Pre*Processing* Normalisasi dan *Stemming* Menggunakan *Metode Naïve Bayes* dan SVM". Jurnal Dinamika Informatika. Vol.5, No.1, November 2015.

Zulfa, I., and Winarko, E., (2017). "Sentimen Analisis Tweet Berbahasa Indonesia dengan Deef *Belief* Network". *IJCSS* 11. 2, 187-198.