Versi Online: http://journal.ubm.ac.id/index.php/jbase
DOI: http://dx.doi.org/10.30813/jbase.v4i2.3000
Hasil Penelitian

Journal of Business and Audit Information Systems
Vol 4 (No.2) : 16-22. 2021
p-ISSN: 2615-6431
e-ISSN: 2620-7907

ANALISA SENTIMEN DATA TEXT PREPROCESSING PADA DATA MINING DENGAN MENGGUNAKAN MACHINE LEARNING

DATA TEXT PRE-PROCESSING SENTIMENT ANALYSIS IN DATA MINING USING MACHINE LEARNING

Bhustomy Hakim[1)]

[1)]School of Computer Science and Technology, Harbin Institute of Technology

Diterima 12 Juli 2021 / Disetujui 20 Juli 2021

ABSTRACT

In this social media world, text is a data which produced easily by people daily. With the large amount of text data available on the internet, data mining such as sentiment analysis can be done for strategic needs. But to do data preprocessing on text still gets its own challenges. Like stopwords, stemming or normalization can be done in this pre-processing stage which will certainly affect the accuracy of the results of the data mining. Therefore, this study was conducted to see the influence of data preprocessing on text on the accuracy of data mining models with machine learning. The classifier used is Naïve Bayes for classifying sentiment reviews will be positive or negative. And the text dataset used was 50,000 reviews from the Internet Movie Database (IMDB) which was divided into 25,000 for training sets and 25,000 for testing sets. In each of those, there were 12,500 positive reviews and negative reviews. With these dataset, there are three different treatments, namely; Baseline where the dataset is left original which is not preprocessed anything, Stopwords where repeated words are considered as connecting words or clausal in the dataset will be deleted and leave the core sentence only, and Stemming where the text dataset will be normalized and cut to get the root of the sentence only. The three treatments were each implemented in the machine learning model for sentiment analysis. New reviews was also created to test the model results of all three different dataset treatments. Different results are surely also obtained from each of the three datasets. This proves that data preprocessing has an effect with the accuracy of the data mining model carried out. In this study, datasets with Baseline treatment were the most accurate.

Keywords: *Data Preprocessing, Sentiment Analysis, Data Mining, Stopwords, Stemming.*

ABSTRAK

Teks merupakan data perhari yang sangat mudah dihasilkan di zaman media sosial ini. Dengan banyaknya data teks yang tersedia di internet, *data mining* seperti analisa sentimen dapat dilakukan untuk kebutuhan strategis. Namun untuk melakukan data preprocessing pada teks masih mendapatkan tantangan tersendiri. Seperti perlakuan *stopwords, stemming* atau normalisasi dapat dilakukan dalam tahap preprocessing ini yang tentunya akan mempengaruhi akurasi dari hasil data mining tersebut. Oleh karena itu penelitian ini dilakukan untuk melihat pengaruh data preprocessing pada teks terhadap akurasi model data mining analisa sentimen dengan machine learning. Classifier yang digunakan adalah Naïve Bayes untuk pengklasifikasian sentiment ulasan akan positif atau negatif. Dan dataset teks yang digunakan adalah 50.000 ulasan di *Internet Movie Database* (IMDB) yang dibagi menjadi 25.000 untuk *training set* dan 25.000 untuk *testing set*. Di masing-masing itu, terdapat 12.500 ulasan positif dan ulasan negatif. Dengan dataset tersebut, terdapat tiga perlakuan berbeda yaitu; Baseline dimana dataset dibiarkan original tidak dilakukan preprocessing apa-apa, Stopwords dimana kata-kata yang berulang yang dianggap sebagai kata penghubung atau klausal di dataset akan dihapus dan menyisakan kalimat intinya saja, dan Stemming dimana dataset teks akan dinormalisasi dan dipotong untuk mendapatkan akar kalimatnya saja. Ketiga perlakuan itu masing-masing diimplemetasikan di *model machine learning* untuk analisa sentimen. Ulasan baru juga dibuat untuk menguji hasil model dari ketiga perlakuan dataset berbeda tersebut. Tentunya hasil yang berbeda juga didapatkan dari masing-masing ketiga dataset tersebut. Hal ini membuktikan bahwa *data preprocessing* berpengaruh dengan hasil akurasi dari model *data mining* yang dilakukan. Dalam penelitian ini, dataset dengan perlakuan Baseline menjadi yang paling tinggi akurasinya.

Kata Kunci: Data Preprocessing, Analisa Sentiment, Data Mining, Stopwords, Stemming.

*)Korespondensi Penulis :
bhakim@bundamulia.ac.id

Journal of Business and Audit Information Systems
Vol 4 (No.2) : 16-22. 2021
p-ISSN: 2615-6431
e-ISSN: 2620-7907

PENDAHULUAN

A. Latar Belakang

Di era ini, orang cenderung menggunakan teknologi mereka untuk melakukan segalanya dalam hidup mereka. Lebih dari miliaran data telah diproduksi hampir setiap hari dan 80% dari mereka diwakili sebagai teks (Radford, 2019). Data teks dapat ditemukan dalam aktivitas virtual orang seperti ketika mereka melakukan percakapan, membaca berita, atau memberikan pendapat mereka tentang segala sesuatu di dunia maya terbuka terutama di media sosial atau platform tertentu untuk membahas topik tertentu.

Dari banyaknya data teks yang tersedia, para peneliti dapat melakukan data mining yang bermanfaat seperti machine learning untuk menjadi bahan analisa lanjutan; *sentiment analysis* untuk mengetahui kecondongan pandangan masyarakat terhadap sesuatu dengan klasifikasi naïve bayes (Rajput, 2019). Dan dalam melakukan data mining, tahap *data preprocessing* merupakan hal yang penting namun untuk melakukan *data preprocessing* terhadap teks masih terbilang cukup sulit (Landers, 2016). Selain karena teks merupakan data yang sangat raw dan bisa berbeda makna dari apa yang dimaksudkan penulis serta bisa tidak baku dan sesuai dengan tata bahasa dikarenakan pergeseran budaya.

Terdapat beberapa tipe *data preprocessing* pada teks antara lain; menghilangkan kata-kata yang merupakan konjunggasi atau penghubung (*stopwords*) seperti “or”, “and”, “if”, dan lain sebagainya, menormalisasi tiap kata menjadi kata dasarnya saja dan menghilangkan tanda-tanda baca yang tidak perlu seperti “isn’t”. Perlakuan *data preprocessing* ini akan merubah struktur kalimat yang sesungguhnya. Oleh karena itu, penelitian dilakukan untuk melihat apakah dengan melakukan data preprocessing pada teks akan berpengaruh terhadap keakurasian hasil *data mining* tersebut.

B. Identifikasi Masalah

Dari latar belakang masalah yang ada, maka dapat dirumuskan identifikasi masalah pada penelitian ini:

1. Apakah menghilangkan stopwords pada tahap *preprocessing* dapat meningkatkan akurasi model?
2. Apakah melakukan normalisasi pada suatu kalimat mempengaruhi akurasi model?

C. Tujuan dan Manfaat Penelitian

Penelitian ini dilakukan dengan tujuan untuk menganilisa dan mengetahui bahwa apakah *data preprocessing* pada teks bisa mempengaruhi dan atau meningkatkan hasil *data mining*, mengetahui detail proses *sentiment analysis* di *machine learning*, serta klasifikasi naïve bayes.

METODE PENELITIAN

Sentiment Analysis akan diterapkan dalam percobaan ini untuk mengekstrak pengetahuan tentang bagaimana ulasan positif atau negatif yang terlihat seperti dalam format teks. Ini adalah penambangan data *(data mining*) kontekstual teks sebagai pengidentifikasian informasi subjektif untuk memahami sentimen orang dari sesuatu sehingga mesin dapat mengklasifikasikan pesan masa depan dan memberikan pernyataan apakah sentimen yang mendasari diberi label (positif atau negatif) (Nakov, 2016).

Aplikasi dasar analisis sentimen terletak pada pengumpulan pendapat orang. Pendapat semacam itu adalah pendahulu dari banyak keputusan bisnis. Mirip dengan stop-kata, domain analisis sentimen tergantung pada daftar kata-kata yang menggambarkan pengaruh penulis. (Rajput, 2019) menjelaskan bagaimana daftar baru digunakan untuk mengklasifikasikan pendapat pengguna sebagai negatif, netral atau

Journal of Business and Audit Information Systems
Vol 4 (No.2) : 16-22. 2021
p-ISSN: 2615-6431
e-ISSN: 2620-7907

positif. Penerapkan konsep analisis sentimen untuk topik daripada pendapat yang sebenarnya dan mengumpulkan bagaimana diskusi akan mengikuti sentimen suatu topik.

Dengan menggunakan *sentiment analysis*, dataset tersebut akan dipelajari dengan melakukan *training* dan *testing* (*supervised machine learning*) untuk mengklasifikasikan ulasan mana yang mewakili sentimen positif atau negatif. Polaritas dataset besar itu akan diturunkan dengan menggunakan teknik tingkat fitur untuk kalimat yang lebih kompleks. Area utama yang terfokus dalam percobaan ini adalah *feature extraction, classification*, dan *analysis method*. Perbandingan antara teks *preprocessing* yang dilakukan dan diimplementasikan di masing-masing dataset akan dijadikan bahan evaluasi pada penelitian ini.

A. *Feature Extraction*

Karena data ulasan film telah siap untuk training dan testing data dengan semua dokumen positif dan negatif yang terpisah, sehingga proses Pengumpulan data dilakukan. Proses *feature extraction* berikutnya dalam percobaan ini adalah text preprocessing yang sangat penting untuk dilakukan dalam analisis sentimen. Untuk membuat data rapi, semua ulasan dalam percobaan ini dikonversi dengan menghapus semua tag html seperti "
" dan tanda baca seperti "?", "!", ":", dan membuat semua teks lebih rendah kasus untuk mencegah kebingungan (Sarica and Luo, 2020).

Untuk membuat semua teks tersebut dapat dibaca oleh mesin, mereka telah diekstraksi ke dalam vektor fitur numerik bernama *Vectorization*. Jadi, semua data diterjemahkan (*tokenized*) ke dalam kamus biner yang disebut *bag of words* (Siddhartha, 2021).

B. Classification

Ide utama dari *classification* adalah untuk menemukan pengamatan dan *feature extraction* yang berguna, maka set kelas absolut harus diklasifikasikan dalam pengamatan. Selain itu, kemungkinan pengamatan bahwa berada di kelas akan memberitahu oleh pengklasifikasi probabilistik. Pengklasifikasi bangunan digunakan dalam banyak jenis algoritma *machine learning*, terutama yang diawasi. Ada begitu banyak jenis pengklasifikasi yang dapat digunakan untuk analisis sentimen seperti Naïve Bayes, *Logistic Regression*, n-Gramm, SVM, dan lain sebagainya.

Dalam percobaan ini, model klasifikasi yang telah digunakan adalah pengklasifikasi Naïve Bayes. Naïve Bayes *classifier* adalah pengklasifikasi generatif yang membangun model cara kelas menghasilkan beberapa input data (Jurafsky and Martin, 2020). Dan kemudian menemukan kelas yang paling mungkin dan menghasilkan mereka menjadi pengamatan yang akan diproses berikutnya.

Naïve Bayes *Classifier* mewakili dokumen teks sebagai bag of words tanpa memikirkan posisinya dari masing-masing kata, misalnya menggunakan kata pesanan dalam frasa seperti "I like the story in the movie" dan "I think the movie is great", itu akan disederhanakan kata "I" ada dua kali dalam seluruh dokumen mengutip kata "film" dua kali. Dan semua kata-kata akan dihitung berdasarkan jumlah katanya.

Pengklasifikasi probabilistik akan mengembalikan kelas yang memiliki probabilitas posterior maksimum di setiap dokumen yang diberikan. Notasi untuk memperkirakan kelas yang telah diklasifikasikan dengan benar adalah: c = argmax P (c | d). Ini berasal dari derivasi ide Bayesian Inference yang diterapkan pada klasifikasi teks pertama dalam sejarah pada tahun 1964.

Bayes' Rule disajikan untuk memberikan cara untuk memecah probabilitas bersyarat P (x | y) menjadi tiga probabilitas lainnya: P (x | y) = (P (y | x) P (x))/(P (y)), dan hasil argmax ditentukan (Jauhiainen, 2018).

Sementara klasifikasi teks Bayes naif standar dapat bekerja dengan baik untuk analisis sentiment. Beberapa perubahan kecil umumnya digunakan yang meningkatkan kinerja. Pertama, untuk klasifikasi sentimen dan sejumlah tugas klasifikasi teks lainnya, apakah sebuah kata terjadi atau tidak tampaknya lebih penting daripada frekuensinya. Varian ini disebut binari multinomial naïve Bayes atau biner NB.

Dengan demikian komputasi kelas kemungkinan diberikan dalam dokumen dengan memilih kelas yang mencapai produk tertinggi dari dua probabilitas, atau disebut probabilitas sebelumnya dari kelas dan kemungkinan probabilitas dokumen terjadi dalam metode ini.

Perhitungan Naïve Bayes akan dilakukan di ruang log, *underflow* dihindari untuk meningkatkan kecepatan. Dengan mempertimbangkan fitur dalam ruang log, kelas yang diprediksi telah dihitung dalam fungsi linear fitur input.

C. Analysis Method

Metode analisis yang terlatih akan dianalisis dengan menggunakan model evaluasi performa kinerja. Ada empat ukuran berbeda dalam metrik ini yang menentukan kinerja hasilnya (Flach, 2018):

True Positive (TP): Ulasan positif yang diklasifikasikan sebagai positif dengan benar

True Negative (TN): Ulasan negatif yang diklasifikasikan sebagai negatif dengan benar

False Positive (FP): Ulasan positif yang diklasifikasikan sebagai negatif secara salah.

False Negative (FN): Ulasan negatif yang diklasifikasikan sebagai positif secara salah.

Setelah itu, semua metrik tersebut akan dihitung sebagai skor *Precision*, Recall, dan F-measures yang mendefinisikan ke dalam rata-rata presisi dan recall tertimbang. Dan ini adalah rumusnya:

Precision (P) = TP/(TP+FP)

Recall (R) = TN/(TN+FN)

F-measure (F) = 2PR/(P+R)

HASIL DAN PEMBAHASAN

Dalam percobaan ini, terdapat 25.000 dataset untuk *training set* yang menjadi acuan dan 25.000 dataset *testing set* ulasan film yang terdiri dari masing-masing 12.500 dokumen positif dan negatif. (Sumber: https://ai.standford.edu/~amaas/data/sentiment/aclImdb_v1.tar.g). Dari data tersebut, masing-masing kata positif dan negatif akan dianalisis yang dapat diklasifikasikan sebagai ulasan yang baik atau ulasan buruk untuk mewakili pendapatnya tentang film. Sehingga dapat diimplementasikan dengan baik untuk memprediksi ulasan masa depan baru dan menghitung apakah model klasifikasinya akurat dan efisien.

A. Evaluasi Kriteria

Evaluasi kriteria dalam percobaan ini dibagi menjadi tiga perlakuan yang berbeda dari data teks. Kriteria pertama adalah satu set data tidak akan diproses, tetap bersih setelah *text preprocessing* yang disebut data “Baseline”. Kriteria kedua adalah seperangkat data yang semua *stopwords* nya telah dihapus. NLTK menghentikan kata-kata dari perpustakaan NLTK dalam python diimplementasikan dalam percobaan ini, sehingga tidak akan ada lagi *stopwords* seperti "if", "in", "it" "to", "but", dan lain sebagainya. Dan itu disebut data “Stop Words”. Data teks yang dinormalisasi adalah kriteria ketiga. Normalisasi mengubah semua bentuk kata yang diberikan menjadi satu bentuk yang baku. Ada begitu banyak jenis normalisasi teks, tetapi *stemming* yang digunakan dalam percobaan ini. *Stemming* menggunakan pendekatan *brute-force* untuk melakukan normalisasi teks. *Porter Stemmer* di *library* NLTK digunakan untuk melakukan stemming pada semua teks percobaan ini yang disebut data “Stemming”.

Journal of Business and Audit Information Systems
Vol 4 (No.2) : 16-22. 2021
p-ISSN: 2615-6431
e-ISSN: 2620-7907

Dan semua kriteria yang berbeda akan diperlakukan sama dengan metode machine learning yang sama, yaitu menggunakan Naïve Bayes *Classifier* sebagai klasifikasi, dan dianalisis dengan *metrics* evaluasi performa kerja.

B. Hasil Penelitian

Setelah percobaan ini berhasil dilakukan dengan python. Berikut ini adalah hasil dari keakuratan pengklasifikasi Naïve-Bayes dalam tiga jenis data yang berbeda (Baseline, Stop Words, dan Stem data) dengan dataset pelatihan 25.000 dan dataset pengujian 25.000.

Tabel 1. Hasil penelitian dari tiga jenis perlakuan data

Data text preprocessing	Akurasi
Baseline	0.83993
Stopwords	0.83632
Stemming	0.82364

Dengan melihat Tabel 1, pengklasifikasi Naïve Bayes bekerja dengan baik dalam data Baseline yang merupakan data yang bersih hanya tanpa pemrosesan teks dengan skor akurasi **0,83993**. Dan kemudian Stop Words (**0.83632**) berada di tempat kedua kemudian Stemming (**0.82364**). Dan kemudian untuk menguji apakah skor akurasinya cukup memuaskan, 10 ulasan baru telah ditambahkan dalam percobaan ini. Ada 5 ulasan positif yang berbeda dan 5 ulasan negatif yang berbeda. Setiap ulasan terdiri dari kata-kata polaritas positif dan kata-kata polaritas negatif. Tetapi beberapa kata membingungkan yang sulit untuk menentukan apakah polaritasnya positif atau negatif bahkan oleh manusia. Dan tabel di bawah ini mewakili hasilnya.

Tabel 2. Review baru yang dikerjakan di model machine learning

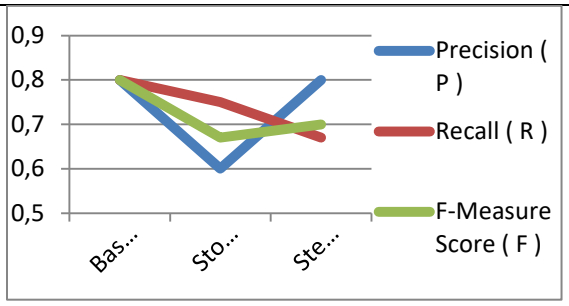
No	Review	B	SW	ST
1	It is an amazing movie. I like it very much. (**Positive**)	+	+	+
2	This is a dull movie. I would never recommend this movie ever to anyone. (**Negative**)	-	-	-
3	The cinematography is pretty great in this movie. The scenery is superbly awesome. (**Positive**)	+	+	+
4	The direction was terrible and the story was all over the place. (**Negative**)	-	-	+
5	The whole cast are awful to play the character. It such a waste to spend your time to watch it. (**Negative**)	-	-	-
6	The actress plays the role in depth, I would never expect the ending. Very bad for people who like romance. (**Positive**)	-	-	-
7	The movie reminds me of my childhood, the way how director plays with audience's feeling is well managed. (**Positive**)	+	+	+

Journal of Business and Audit Information Systems
Vol 4 (No.2) : 16-22. 2021
p-ISSN: 2615-6431
e-ISSN: 2620-7907

No	Review	B	SW	ST
8	I hate everything in this movie. (**Negative**)	+	+	+
9	I love the movie so much. All of action in this movie are fascinating.(**Positive**)	-	+	+
10	The story is terrible. The sense of humor makes me sick. (**Negative**)	+	+	+

Note: Kolom B untuk Baseline, SW untuk Stopwords, dan ST untuk Stemming. Dan apabila dikolom tersebut + maka menghasilkan klasifikasi bahwa review tersebut positif, apabila – maka klasifikasi negatif.

Ada enam ulasan (Review No. 1, No. 2, No. 3, No. 5, No. 7, dan No. 8) yang diklasifikasikan dengan baik dengan menggunakan Naïve Bayes dengan berbagai jenis tipe *text preprocessing*. Tapi empat ulasan sulit untuk diklasifikasikan. Dalam Review No. 4 dan No. 10, bahkan ada kata "terrified" yang jelas memiliki polaritas negatif, tetapi mereka diklasifikasikan salah (No. 4 di Stemming saja). Begitu juga dengan Review No. 9. Meskipun kalimatnya sangat sederhana, kata "love" tidak ditentukan positif oleh mesin dengan Naïve Bayes *Classifier* ini. Dan dalam Review No. 6, kalimatnya sangat ambigu, itu sebabnya di tiga data berbeda diklasifikasikan salah. Berdasarkan hasil 10 ulasan (review) baru, maka metrics evaluasi performa kerja telah dihitung dan ditunjukkan pada grafik ini di bawah ini:



Gambar 1. Grafik hasil evaluasi performa kerja.

Mengacu grafik di atas, hasil yang baik masih dipegang oleh data Baseline yang mencapai 0,8 untuk Skor *F-Measure* Ini menafsirkan bahwa Naïve Bayes *Classifier* bekerja dengan baik jika data masih sepenuhnya bersih tanpa menghapus atau menormalkan kata-kata pada kalimat. Jadi, dalam data Baseline, setidaknya ada 20.000 data pengujian yang telah diklasifikasikan dengan benar, dan diperkirakan hanya 5000 di antaranya yang salah diklasifikasikan.

Data Stem berada di tengah dengan Skor *F-Measure* 0,73. Presisinya tinggi, tetapi *Recall* rendah. Hal ini karena dalam data Stem ini, Naïve Bayes tidak dapat mendefinisikan ulasan negatif karena orang menggunakan banyak kontraksi bahasa Inggris dalam kata kalimat negatif seperti "isn't", "doesn't", "aren't", "ain't" dan menormalisasi kata-kata itu menjadi kalimat aslinya.

Sementara itu data Stop Words adalah yang paing rendah tingkat akurasinya dengan Naïve Bayes *Classifier* dengan 0,67 untuk Skor *F-Measure* dalam percobaan ini. Karena percobaan menggunakan default bahasa Inggris modul NLTK Stop Words, hal ini menghilangkan begitu banyak kata yang merubah arti sehingga klasifikasi tidak bekerja dengan baik di Naïve Bayes ini.

KESIMPULAN

Naïve Bayes *Classifier* dapat bekerja dengan baik dalam kalimat kompleks yang masih penuh dan lengkap sebagai teks yang asli tanpa modifikasi (Baseline). Jika banyak kata yang telah dihapus dalam kalimat dengan *stopwords* atau *stemming*, hal itu akan merubah pesan dan arti yang terkandung dalam kalimatnya dan mengurangi keakuratannya pada model klasifikasi *machine learning*. Ini terjadi ketika default bahasa Inggris modul fungsi NLTK Stop Words menghapus begitu banyak *stopwords* yang ada pada kalimat. Serta Porter Stem memotong konstruksi negatif dalam kalimat negatif yang menyebabkan kalimat menjadi netral. Hal ini membuktikan bahwa text preprocessing juga berperan penting dalam menghasilkan keakurasian dan tingkat keefektifan model machine learning pada proses data mining.

Untuk penelitian selanjutnya, direkomendasikan untuk melakukkan pengklasifikasi lain seperti *Logistic Regression Classifier* atau n-Gramm *Classifier* sehingga dapat membantu kalimat kompleks yang *stopwordsnya* telah dihapus dan mengomparasi efesiensi dalam percobaan ini. Dengan demikian, perlu eksperimen lebih lanjut dengan pengklasifikasi yang berbeda dan cara perlakuan *text preprocessing* yang berbeda juga terhadap kalimat kompleks dalam dataset besar.

DAFTAR PUSTAKA

Flach, P. (2018). Performance Evaluation in Machine Learning: The Good, The Bad, The Ugly and The Way Forward.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lind´en, K. (2018). Automatic language identification in texts: A survey.

Jurafsky, D., Martin, J. H. (2020). Speech and Language Processing, Chapter: Naïve Bayes and Sentiment Classification.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter.

Radford A, Wu J, Child R, et al. (2019). Language models are unsupervised multitask learning.

Rajput, A. (2019). Natural Language Processing, Sentiment Analysis and Clinical Analytics.

Serhad Sarica and Jianxi Luo. (2020). Stopwords In Technical Language Processing.

Siddhartha, B. S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing