JISKa, Vol. 1, No. 1, MEI, 2016, Pp. 1 – 4
ISSN 2527-5836

PENGARUH STOPWORD TERHADAP PERFORMA KLASIFIKASI TWEET BERBAHASA INDONESIA

Ahmad Fathan Hidayatullah[(1)]
Jurusan Teknik Informatika, Universitas Islam Indonesia
Jl. Kaliurang km 14,5 Sleman, Yogyakarta
e-mail : fathan@uii.ac.id

Abstact

Data tweet has been used in research in the field of text mining. One of them is in the text classification. However, most of data is the dirty tweet and contains a lot of noise in it. Therefore, intial processed tweet is very important to do. One of an initial processing method for reducing noise in a tweet is stopword removal. Furthermore this research will do a comparasion between the results of the accuracy between intial processing involving the stopword removal process with intial processing without involving stopword removal. This is done to determine the significance stopword removal stages on indonesia language classification. On this research, conducted two intial processing in which one involving stopword removal and other process without stopword removal. Experimental results showed that the removal stopword on the pre-processing can improve classification performance as evidenced by an incrase accuracy.

Keywords: Stopword removal, pre-processing, text classification, tweet classification.

Abstrak

Data *tweet* telah banyak dimanfaatkan dalam penelitian di bidang *text mining*. Salah satu diantaranya adalah dalam klasifikasi teks. Namun, sebagian besar data *tweet* merupakan data yang masih kotor dan mengandung banyak *noise* di dalamnya. Oleh karena itu, pemrosesan awal terhadap *tweet* sangat penting untuk dilakukan. Salah satu metode pemrosesan awal yang dilakukan untuk mereduksi *noise* dalam *tweet* adalah *stopword removal*. Lebih lanjut penelitian ini akan melakukan perbandingan hasil akurasi antara pemrosesan awal yang melibatkan proses penghapusan *stopword* dengan permosesan awal yang tanpa melibatkan *stopword removal*. Hal ini dilakukan untuk mengetahui signifikansi tahapan *stopword removal* dalam klasifikasi teks berbahasa Indonesia. Dalam penelitian ini, dilakukan dua model pemrosesan awal dimana salah satu proses melibatkan *stopword removal* dan proses yang lainnya tanpa melakukan *stopword removal*. Hasil eksperimen menunjukkan bahwa melakukan penghapusan *stopword* dalam *pre-processing* mampu meningkatkan performa klasifikasi yang dibuktikan dengan adanya peningkatan akurasi.

Kata Kunci : *Stopword removal, pre-processing,* klasifikasi teks, klasifikasi tweet

1. PENDAHULUAN

Media jejaring sosial terus mengalami perkembangan dalam beberapa tahun terakhir. Media jejaring sosial sebagai aplikasi terpenting dari web 2.0 mengizinkan para penggunanya agar dapat membangun jaringan untuk saling mengenal satu sama lain, saling berbagi informasi, serta menggunakan layanan untuk berbagi foto, blog, wiki, dan sebagainya (Kumar dan Sebastian, 2012). Di kalangan pengguna internet saat ini, Twitter merupakan salah satu diantara sekian banyak media jejaring sosial yang cukup populer.Twitter memberikan fasilitas bagi para penggunanya untuk mem-*posting* segala hal terkait aktivitas, opini, dan segala hal yang terjadi kepada publik melalui pesan singkat yang dikenal dengan sebutan *tweet*.

Saat ini, data *tweet* telah banyak dimanfaatkan dalam penelitian di bidang *text mining*. Salah satu diantaranya adalah dalam klasifikasi teks. Namun, ada beberapa permasalahan yang harus ditangani terlebih dahulu sebelum proses klasifikasi teks dilakukan. Sebagian besar data *tweet* merupakan data yang masih kotor dan mengandung banyak *noise* di dalamnya. Hal tersebut disebabkan karena Twitter tidak memiliki aturan bagi para penggunanya dalam mengekspresikan tulisan mereka dalam *tweet*. Akibatnya para pengguna lebih cenderung

menggunakan bahasa sehari-hari yang tidak baku atau tidak sesuai dengan kaidah bahasa yang benar. Selain itu, Twitter hanya memberikan batasan maksimal 140 karakter kepada para penggunanya untuk menuliskan *tweet* sehingga menuntut pengguna Twitter untuk mempersingkat penulisan *tweet*.

Oleh karena itu, pemrosesan awal terhadap *tweet* sangat penting untuk dilakukan. Salah satu metode pemrosesan awal yang dilakukan untuk mereduksi *noise* dalam *tweet* adalah *stopword removal*. *Stopword removal* dilakukan dengan membuang kata-kata yang cukup umum dan sering muncul namun tidak mempunyai pengaruh yang signifikan terhadap makna suatu teks atau kalimat. Lebih lanjut penelitian ini akan melakukan perbandingan hasil akurasi antara pemrosesan awal yang melibatkan proses penghapusan *stopward* dengan permosesan awal yang tanpa melibatkan *stopword removal*. Hal ini dilakukan untuk mengetahui signifikansi tahapan *stopword removal* dalam klasifikasi teks berbahasa Indonesia. Apabila tahap penghilangan *stopword* dalam *tweet* dapat meningkatkan akurasi maka proses tersebut dapat dipertahankan dalam pemrosesan awal. Namun sebaliknya, apabila *stopword removal* tidak perlu dilakukan dalam pemrosesan awal jika performa akurasi semakin menurun.

Penulisan makalah ini terdiri dari lima bagian. Bagian pertama merupakan pendahuluan yang memuat latar belakang dari penelitian ini. Penelitian-penelitian sebelumnya yang terkait dan mendukung penelitian ini dijelaskan pada bagian kedua. Pada bagian ketiga, dibahas metode yang digunakan dalam penelitian ini. Bagian keempat memaparkan hasil eksperimen dan pembahasan dari hasil yang diperoleh. Bagian kelima merupakan bagian terakhir yang berisi kesimpulan dari penelitian.

Proses eliminasi *stopword* dalam *information retrieval* akan membantu dalam mereduksi *feature space* sehingga akan membantu dalam mengurangi waktu dan kompleksitas ruang (Jayashree et al., 2014). Namun demikian, *stopword removal* tidak memberikan kontribusi yang signifikan terhadap peningkatan kinerja klasifikasi. Saif et al. (2014) mengungkapkan bahwa telah terjadi perdebatan mengenai efektivitas penggunaan *stopword removal* untuk klasifikasi sentimen pada Twitter. Dari hasil penelitian, diketahui bahwa proses *stopword removal* menggunakan data *pre-compiled stopword* yang dilakukan memberikan dampak negatif terhadap hasil kinerja klasifikasi sentimen. Sebaliknya, Srividhya dan Anitha (2010) menyatakan bahwa penghapusan *stopword* dapat meningkatkan performa kinerja klasifikasi. Ghag dan Shah (2015) juga melakukan analisis mengenai dampak dari *stopword removal* dalam analisis sentimen. Dari hasil penelitian, diketahui bahwa terjadi peningkatan akurasi setelah penghapusan *stopword* dilakukan.

2. METODE PENELITIAN

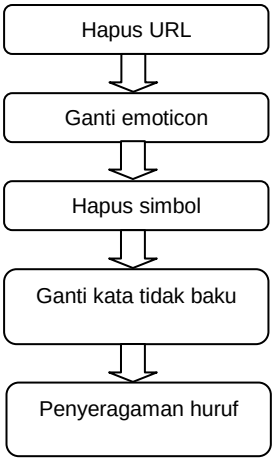
2.1 Data Penelitian

Penelitian ini menggunakan data Twitter yang digunakan oleh Hidayatullah (2016). Data *tweet* diperoleh secara berkala menggunakan Twitter API v1.1. Sebanyak 2000 data *tweet* yang didapat kemudian diberi label secara manual ke dalam dua polaritas sentimen yaitu positif dan negatif. Dari proses pelabelan secara manual, diketahui terdapat 1074 *tweet* dengan sentimen positif dan 926 *tweet* dengan sentimen negatif.

2.2 *Pre-processing*

Pre-processing dalam *text mining* bertujuan untuk mempersiapkan data sebelum diproses pada langkah selanjutnya. Dikarenakan penelitian ini akan membandingkan pengaruh dari *stopword removal*, maka dilakukan dua tahap *pre-processing* yang berbeda. *Pre-processing* yang pertama dilakukan melalui tahap *stopword removal* sedangkan yang kedua tanpa melalui tahap *stopword removal*. Adapun pada *pre-processing* yang melibatkan proses *stopword removal,* proses penghapusan *stopword* dilakukan pada tahap paling akhir. Selain *stopword removal*, penelitian ini melakukan beberapa tahapan lainnya untuk membersihkan data *tweet* diantaranya menghapus URL, mengganti *emoticon* menjadi kata-kata yang mewakilinya, menghapus simbol dan tanda baca, mengganti kata tidak baku menjadi kata baku, serta menyeragamkan huruf ke

dalam bentuk huruf kecil. Urutan tahapan *pre-processing* yang dilakukan tanpa proses pengapusan *stopword* dapat dilihat pada Gambar 1.



Gambar 1. Tahapan pre-processing tanpa *stopword*

2.3 Pemilihan dan Ekstraksi Fitur

1. *Unigram Feature*

Unigram feature merupakan salah satu bagian dari metode *n-gram*. *Unigram* dilakukan dengan membagi suatu dokumen ke dalam bentuk kata per kata. Dalam hal ini, *unigram* tidak memperhatikan konteks dan tidak memperhatikan adanya keterkaitan antara satu kata dengan kata yang lain (Manning et al., 2009).

2. *Term Frequency*

Term frequency bertujuan untuk menghitung kemunculan suatu *term* dalam suatu *corpus* berdasarkan bobot suatu *term* pada dokumen tertentu. Dalam suatu dokumen, apabila *term* tertentu memiliki kemuncuan yang tinggi, maka akan semakin tinggi bobot dokumen untuk *term* tersebut, dan sebaliknya**.**

2.4 Klasifikasi

Penelitian ini menggunakan *Holdout* sebagai metode untuk membagi antara data *training* dengan data *testing*. *Metode Holdout* menggunakan sebanyak setengah atau dua per tiga dari data keseluruhan untuk keperluan proses *training* dan sisanya digunakan untuk *testing* (Witten, et al., 2011). Dalam penelitian ini, ditentukan sebanyak dua per tiga data sebagai data *training* sebagai data model dan sepertiga sisanya sebagai data *testing*. Proses klasifikasi dokumen dilakukan menggunakan tiga buah algoritma yaitu SVM, *Naïve Bayes*, dan K-NN.

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil eksperimen yang diperlihatkan oleh Tabel 1, diketahui bahwa proses *stopword removal* yang dilakukan dalam tahap pemrosesan awal mampu meningkatkan hasil akurasi. Namun demikian, selisih hasil akurasi antara pemrosesan awal yang melibatkan pembuangan *stopword* tidak terlalu tinggi.

Tabel 1. Akurasi Hasil Klasifikasi

Algoritma	Akurasi	
	Stopword	**Tanpa *Stopword***
SVM	88,59%	87,09%
Naïve Bayes	86,94%	84,32%
K-NN	80,93%	78,23%

Dari ketiga algoritma yang digunakan, SVM memiliki hasil akurasi paling tinggi. Sedangkan hasil akurasi dengan *Naive Bayes* dan K-NN menempati posisi kedua dan ketiga. Pada *pre-processing* dengan melibatkan proses penghapusan *stopword*, diperoleh akurasi sebesar 88,59% untuk algoritma SVM. Sebaliknya, hasil akurasi dengan SVM tanpa melalui proses *stopword removal* diperoleh sebesar 87,09%.
Pada urutan kedua dengan algoritma *Naive Bayes*, diperoleh akurasi sebesar 86,94% dengan menghapus *stopword* dan 84,32 % dengan tanpa melakukan penghapusan *stopword*. K-NN menghasilkan akurasi paling rendah dengan 80,93% jika melalui tahap *stopword removal*. Apabila tanpa melalui *stopword removal* akurasi dengan K-NN didapatkan sebesar 78,23%.

4. KESIMPULAN

Penelitian ini telah melakukan analisis pengaruh *stopword removal* terhadap performa klasifikasi teks berbahasa Indonesia. Dalam penelitian ini, dilakukan dua model pemrosesan awal dimana salah satu proses melibatkan *stopword removal* dan proses yang lainnya tanpa melakukan *stopword removal*. Hasil eksperimen menunjukkan bahwa selisih akurasi yang terjadi antara dua model pemrosesan awal yang diusulkan tidak terlalu tinggi. Namun demikian, dilakukannya penghapusan *stopword* dalam *pre-processing* mampu meningkatkan performa klasifikasi yang dibuktikan dengan adanya peningkatan akurasi.

DAFTAR PUSTAKA

Ghag, K. V., & Shah, K. 2015. *Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification*, Computer, Communication and Control (IC4), 2015 International Conference, IEEE, pp. 1-6.

Jayashree, R., Murthy, K.S. and Anami, B.S. 2014. *Effect of Stopword Removal on The Performance of Naïve Bayesian Methods for Text Classification in The Kannada Language*. International Journal of Artificial Intelligence and Soft Computing, 4(2-3), pp.264-282.

Kumar, A., & Sebastian, T.M. 2012. *Sentimen Analysis on Twitter*, International Journal of Computer Science Issues (IJCSI), 9(3), pp.1694-0814.

Manning, C., Raghavan, P., & Schutze, H. 2009. *Introduction to Information Retrieval*, Cambridge University Press.

Saif, H., Fernández, M., He, Y., & Alani, H. 2014. *On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter*. Proceedings of The Ninth International Conference on Language Resources and Evaluation (LREC 2014), pp. 810-817.

Srividhya, V. & Anitha, R. 2010. *Evaluating Preprocessing Techniques in Text Categorization*. International Journal of Computer Science and Application, 47(11), pp. 49-51.

Witten, I. H., Frank, E., & Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*, Elsevier.