IMPLEMENTASI DAN ANALISIS KLASIFIKASI SENTIMEN PADA TWEET BERBAHASA INDONESIA DENGAN MULTINOMIAL NAIVE BAYES

Jaka Arya Pradana[1], Warih Maharani[2], Kemas Rahmat Saleh Wiharja[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Twitter merupakan jejaring sosial yang sedang marak di dunia, termasuk juga di Indonesia. Penggunanya di Indonesia mencapai 9,9 juta dengan tingkat pengaksesan 22% dari populasi netizen. Ini menunjukan bahwa Twitter merupakan salah satu media yang patut diperhitungkan untuk menganalisis sentimen masyarakat Indonesia.

Tugas akhir ini akan menggunakan multinomial naive bayes untuk mengelompokkan tweet berdasarkan kelas sentimennya, yaitu positif, negatif, atau netral. Data collection yang digunakan adalah data collection berbahasa Indonesia, sebab mayoritas masyarakat Indonesia mentweet dengan bahasa Indonesia. Ada pun klasifikasi yang dilakukan menggunakan unigram, bigram, trigram, dan pos tag sebagai atribut.

Setelah dilakukan pengujian dengan beberapa skenario, menunjukkan bahwa multinomial naive bayes dapat digunakan untuk mengelompokkan tweet berbahasa Indonesia berdasarkan kelas sentimen. Ini terlihat dari accuracy sistem yang mencapai 95,539%.

Kata Kunci : analisis sentimen, twitter, multinomial naive bayes

Abstract

Twitter is a social network wich is glowing in the world, including Indonesia. Twitter users in Indonesia reached 9.9 million with accessing rate 22% of all Indonesian netizen population. This show that Twitter is one of media to be reckoned with to analize Indonesian public sentiment.

This undergraduate thesis wil use multinomial naive bayes to classified tweet by its sentiment, that are positive, negative or neutral. The data collection that used was the data collection with Indonesian language, because most of Indonesian people twitting in Indonesian language. And for the classifivation were performed using unigrams, bigrams, trigrams, and pos tag as the attributes.

After testing with several scenarios, showed that the multinomial naive bayes is relevant to be used for classifiying Indonesian language tweets by the sentiment. This is showed by the system accuracy is achieved 95.539%.

Keywords : sentiment analysis, twitter, multinomial naive bayes

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Twitter merupakan jejaring sosial yang sedang marak pada masa kini. Pada bulan Maret 2011, jumlah pengguna twitter di seluruh dunia setidaknya ada sekitar 200 juta pengguna [8] dengan 65 juta *tweet* (status atau pesan) perhari [9]. Melalui twitter, para pengguna mentweetkan berbagai hal, termasuk pandangan dan ekspresi kesukaan maupun ketidak sukaan mereka terhadap berbagai hal, seperti produk komersial, kebijakan publik, isu sosial, politik, ekonomi, dan sebagainya.

Indonesia merupakan salah satu negara dengan pengguna twitter terbanyak di dunia dengan jumlah pengguna sekitar 9,9 juta dan tingkat pengaksesan sebesar 22% dari populasi netizen [10]. Ini artinya twitter merupakan salah satu sosial media yang patut diperhitungkan untuk menganalisis sentimen masyarakat Indonesia

Penelitian tentang analisis sentimen pada twitter kini marak dikembangkan [1, 2 & 4]. Namun sampai saat ini belum ada penelitian analisis sentimen untuk *tweet* berbahasa Indonesia. Oleh karena itu penulis berinisiatif untuk menelitinya.

Sentimen yang terkandung dalam suatu *tweet* terkadang bersifat eksplisit, namun terkadang juga bersifat implisit [3], artinya tidak ada korelasi yang benar-benar bersifat pasti antara masing-masing unsur penyusun suatu *tweet* (atribut) dengan kelas sentimennya. Maka dari itu, penulis memilih untuk melakukan klasifikasi dengan metode *Naive Bayes*, karena *Naive Bayes* merupakan pendekatan probabilistik, sehingga cocok untuk klasifikasi yang tidak dapat diprediksi secara pasti, meskipun atribut-atributnya identik dengan kelas tertentu [6].

Pendapat ini diperkuat oleh hasil penelitian yang telah dilakukan oleh Alec Go, dkk [2] dan Albert Bifet [1] yang menunjukan bahwa *Multinomial Naive Bayes* merupakan metode dengan pencapaian akurasi yang termasuk paling tinggi untuk *tweet* berbahasa Inggris. Boleh jadi hal tersebut juga lah yang mendorong Alexander Pak menggunakan *Multinomial Naive Bayes* untuk melakukan klasifikasi sentimen untuk *tweet* berbahasa Inggris, Cina, dan Perancis [4].

Seperti penelitian sebelumnya [4], penulis akan menjadikan *N-Gram* (*unigram*, *bigram*, dan *trigram*) dan *POS Tag* sebagai atribut.

1.2 Perumusan Masalah

Permasalahan yang akan dibahas dalam Tugas Akhir ini adalah bagaimana melakukan klasifikasi sentimen terhadap *tweet* berbahasa Indonesia dengan menerapkan *Multinomial Naive Bayes*, serta melakukan analisa atas hasil klasifikasi tersebut dengan menghitung *accuracy* terhadap data testing.

1.3 Batasan Masalah

Adapun batasan masalah dalam tugas akhir ini adalah :

A. Menggunakan *data collection* untuk *tweet* berbahasa Indonesia, yaitu JAP Corpus.
B. *POS tagging* akan dilakukan dengan menggunakan *tools* yang sudah tersedia, yaitu IPOS tagger [7]. Kerja sistem IPOS Tagger tersebut bukan merupakan bagian dari sistem yang akan dibahas ataupun dianalisis.
C. Persoalan kata yang tidak sesuai dengan ejaan yang benar (bahasa alay) akan diminimalkan dan tidak menjadi bagian yang akan ditangani/dinormalisasi.

1.4 Tujuan

Tujuan yang ingin dicapai pada Tugas Akhir ini adalah :

A. Melakukan klasifikasi sentimen pada *tweet* berbahasa Indonesia dengan menerapkan *Multinomial Naive Bayes.*
B. Menganalisis akurasi penerapan *Multinomial Naive Bayes* untuk klasifikasi sentimen pada *tweet* berbahasa Indonesia dengan menghitung *accuracy* klasifikasi hasil keluaran sistem terhadap *data testing* pada JAP corpus.

1.5 Metode Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan masalah dalam Tugas Akhir ini adalah :

A. Studi Literatur

Mempelajari landasan teori dan literatur dari jurnal, paper, artikel, dan buku yang berkaitan dengan *Data Mining*, *Sentiment Analysis*, *Naive Bayes* serta hal-hal lain yang juga berkaitan.

B. Pengumpulan JAP corpus

Mengumpulkan *data collection* untuk *tweet* berbahasa Indonesia dengan menggunakan API Twitter, yaitu Archivist. *Tweet* positif dan negatif akan digolongkan berdasarkan *emoticon*, sedangkan *tweet* netral akan dikumpulkan dari *tweet* yang berasal dari akun situs berita [4]. *Data collection* ini selanjutnya akan disebut JAP Corpus.

C. Perancangan dan Implementasi Sistem

Merancang dan melakukan implementasi sistem dengan membangun perangkat lunak yang sesuai dengan perancangan yang telah dilakukan. Sistem akan dibangun menggunakan IDE NetBeans dengan bahasa Java.

D. Pengujian Sistem dan Analisa Hasil

Melakukan *training* dengan menggunakan *data training* dari JAP Corpus, kemudian melakukan *testing* dengan menggunakan *data testing* dari JAP Corpus. Lalu melakukan analisa terhadap hasil uji dengan menghitung *accuracy* hasil klasifikasi keluaran sistem dengan *data testing* pada JAP Corpus.

E. Penyusunan Laporan Tugas Akhir.

Membuat laporan sesuai dari langkah-langkah pertama sampai akhir dalam melakukan penelitian tugas akhir ini.

BAB V
PENUTUPAN

5.1 Simpulan

1. Klasifikasi sentimen untuk *tweet* berbahasa Indonesia dapat dilakukan dengan metode *Multinomial Naive Bayes*

2. *Accuracy* klasisfikasi sentimen untuk *tweet* berbahasa Indonesia akan baik bila setiap *tweet* memiliki minimal satu atribut yang ada pada data *training*. Bila ada *tweet* yang tidak memiliki atribut yang tersedia dalam data *training*, hal ini akan menjatuhkan akurasi dari sistem. Maka sebaiknya pemilihan atribut yang akan dipakai disesuaikan dengan karakter data *training*-nya.

3. Semakin banyak atribut yang digunakan akan semakin baik. Hal ini terlihat dari akurasi paling tinggi dicapai saat menggunakan unigram, bigram, trigram, dan POS Tag sebagai atribut.

4. Penggunaan POS Tag secara tunggal tidak baik, hal ini dapat dilihat dari buruknya akurasi klasifikasi yang hanya menggunakan POS Tag sebagai atribut. Hal ini dikarenakan jenis-jenis POS Tag sangat terbatas.

5.2 Saran

1. Membuat data set standar untuk klasifikasi tweet berbahasa Indonesia.

2. Membuat Tagger yang khusus untuk tweet berbahasa Indonesia, karena tweet sering kali memiliki struktur kalimat yang jauh berbeda dengan kalimat bahasa Indonesia yang baik dan benar

DAFTAR PUSTAKA

[1]	Bifet, Albert., dan Frank, Eibe. 2010. *Sentiment Knowledge Discovery in Twitter Streaming Data.* Hamilton: Wakaito University.
[2]	Go A., Huang L., Bhayani R. 2009. *Twitter Sentiment Analysis*. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
[3]	Liu, Bing. 2010. *Sentiment Analysis and Subjectivity*. To appear in Handbook of Natural Language Processing, Second Edition (editors: N. Indurkhya and F. J. Damerau).
[4]	Pak, Alexander., dan Paroubek, Patrick. 2010. *Microblogging for Micro Sentiment Analysis.* Paris: Limsi.
[5]	Suyanto. 2007. *Artificial Intelligence Searching, Reasoning, Planning, and Learning*. Bandung: Penerbit Informatika.
[6]	Tan, Pang-Ning., Steinbach, Michael., dan Kumar., Vipin. 2006. *Introduction to Data Mining*. Boston: Addison Wesley.
[7]	Wicaksono, Alfan Farizki., dan Purwarianti, Ayu. 2010. *HMM Based Part-of-Speech Tagger for Bahasa Indonesia*. In Proceeding of the Fourth International MALINDO Workshop. Jakarta, Indonesia.
[8]	Shiels Maggie. Twitter co-founder Jack Dorsey Rejoins Company. 2011. http://www.bbc.co.uk/news/business-12889048 dibuka pada tanggal 5 Mei 2011
[9]	Garret, Seans. Big Goals, Big Game, Big Records. 2010. http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html dibuka pada 5 Mei 2011
[10]	Tim Pedomannews.com. Aneh, SBY Bukan Pemakai Twitter, Bangga 9,9 Juta Tweeps di Indonesia. http://www.pedomannews.com/bisnis-a-keuangan/berita-bisnis-a-keuangan/investasi/2992-investasi dibuka pada 5 Mei 2011
[11]	Horn Christoper. 2010. *Analysis and Classification of Twitter Message*. Graz University of Technology.