METODE HOLISTIC LEXICON-BASED UNTUK ANALISIS SENTIMEN PADA DOKUMEN BAHASA INDONESIA (STUDI KASUS : TWEETS MENGENAI ISU SOSIAL KOTA BANDUNG)

Immanuel Desmon Christianto Purba[1], Hetti Hidayati [2], Alfian Akbar Gozali[3]

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

Abstrak

Perkembangan media sosial mengalami peningkatan setiap tahunnya, khususnya Twitter. Berdasarkan rangking yang dibuat oleh situs eBizMBA, twitter berada pada urutan kedua dengan jumlah penggunanya menembus angka 250 juta pada periode November 2013. Dan menurut ranking yang dilakukan situs eMarketer, Indonesia berada pada urutan pertama tingkat penambahan jumlah akun twitter. Informasi yang dapat diambil dari twitter antara lain adalah berita baru, ide, dan bahkan opini. Contoh bentuk opini adalah pendapat seseorang mengenai permasalahan kemacetan ibu kota atau review seseorang mengenai produk tertentu. Opini yang dihasilkan ini dapat berguna untuk menentukan kebijakan atau keputusan bagi organisasi atau institusi ke depannya. Kebijakan ini dibuat berdasarkan sifat opini yang berupa sentimen positif atau negatif. Oleh karena itu, diperlukan analisis lebih lanjut untuk menentukan opini tersebut termasuk kelompok sentimen positif atau negatif.

Tugas akhir ini bertujuan untuk menentukan kelas sentimen dari setiap tweets yang diklasifikasikan menjadi dua jenis yaitu sentimenpositif dan negatif. Dataset yang digunakan pada penelitian tugas akhir ini adalah tweets mengenai isu sosial yang ada di Kota Bandung. Metode yang diterapkan pada penelitian ini adalah metode Holistic Lexicon-Based. Metode ini merupakan pengembangan dari metode Lexicon-Based.Metode ini dapat menangani permasalahan multi-opini untuk setiap data tweets. Data yang telah diambil dari twitter akan dilakukan proses preprocessing terlebih dahulu agar hasil klasifikasi lebih baik. Kemudian dilakukan proses klasifikasi ke dalam jenis sentimen dengan melihat setiap orientasi opinion words.

Berdasarkan hasil pengujian, dengan metode Holistic Lexicon-Baseddapat mengidentifikasi kalimat opini dari dataset twitter dan menentukan kelas sentimen dengan rata-rata akurasi 89%. Besarnya nilai akurasi ini dipengaruhi oleh jumlah kalimat opini yang dapat diidentifikasi dan kelengkapan kamus yang dipakai.

Kata Kunci : Opinion Mining, Sentiment Analysis, Sentiment Classification, Holistic Lexicon-Based, Tweets

Abstract

The evolution of social media is increasingevery year, especially Twitter. Based on the ranking that was created by eBizMBA site, twitter rankedin the second position with total of user exceeded 250 million on November 2013. And according to ranks that was made by eMarketer site, Indonesia ranks first on the growth of twitter account. The information that can be retrieved from twitter are hot news, ideas, even opinions. The example of opinion is someone's opinion about traffic jam in the capital city or someone's review about certain products. This opinion result can be useful for determine policy or decision for organizations or institutions in the future. This policy made based on opinion characteristics which is a positive or negative sentiments. Therefore, further analysis is needed to determine whether that opinion is positive or negative sentiment's class.

This final project aim to determine sentiment class from each tweets that classified as two kind of sentiment, positive and negative. The dataset that is used in this final project are tweets about social issues in Bandung city. The method that applied in this research is Holistic Lexicon-Based method. This method is the development of Lexicon-Based method. This method can handle multi-opinion problems for each tweets. The data that has been taken from twitter, will be processed in preprocessing first so that the result of classification is preferably. Then, the data will be classified into sentiment with considering each opinion words orientation.

Based on testing result, with Holistic Lexicon-Based methodthe opinion sentence can be identified from twitter dataset and determine sentiment class with average of accuracy 89%. This accuracy is affected by the number of opinion sentence that has been identified and the completeness of the dictionary.

Keywords : Opinion Mining, Sentiment Analysis, Sentiment Classification, Holistic Lexicon-Based, Tweets

1. Pendahuluan

1.1 Latar Belakang

Twitter adalah sebuah media sosial yang memberikan kemudahan bagi penggunanya untuk membagi dan mengakses informasi. Perkembangan jumlah pengguna *twitter* mengalami peningkatan setiap tahunnya. Berdasarkan *ranking* yang dibuat oleh situs *eBizMBA*, *twitter* berada pada urutan kedua setelah *Facebook* di dalam *Top 15 Most Popular Social Networking Sites*. Jumlah penggunanya menembus angka 250 juta pada periode November 2013 [19]. Dan menurut *ranking* yang dilakukan situs *eMarketer*, Indonesia berada pada urutan pertama tingkat penambahan jumlah akun *twitter*. Data ini diperoleh dalam setahun dengan peningkatan sebesar 44,20% [20]. Bertambahnya jumlah pengguna berdasarkan data tersebut, membuat informasi yang tersebar pun semakin meningkat. Informasi yang tersebar di *twitter* antara lain adalah berita terbaru, ide, dan bahkan opini.

Menurut *Webster's New Collegiate Dictionary*, opini merupakan suatu pandangan , keputusan atau sebuah taksiran yang terbentuk di dalam pikiran mengenai suatu persoalan tertentu. Opini banyak berpengaruh di dalam kehidupan sosial. Pengaruhnya di antara lain adalah di dalam penjualan produk, perubahan kebijakan di dalam sistem pemerintahan, dan bahkan suara rakyat pada pemilihan umum [1]. Dengan kumpulan opini, organisasi atau individu dapat mengetahui keputusan yang tepat dalam menangani sesuatu hal. Namun, opini bukan hanya berisi hal – hal yang positif, banyak informasi dari opini yang mengandung hal negatif. Untuk itu, diperlukan sebuah metode untuk mengambil opini dan mengelompokkannya berdasarkan sentimen yang sesuai.

Sentiment Analysis adalah sebuah teknik perkembangan dari *Natural Language Processing* dengan area riset dari klasifikasi level dokumen [2] sampai pembelajaran polaritas kata dan frase [3]. Teknik ini memiliki tujuan untuk mencari opini dari seseorang tentang sesuatu hal yang spesifik. Berbagai penelitian telah dilakukan untuk menganalisis opini, khususnya pada media sosial *twitter*, diantaranya adalah "Twitter as a Corpus for *Sentiment* Analysis and Opinion Mining" oleh Alexander Pak dan Pattrick Paroubek. Penelitian tersebut memiliki tujuan untuk membangun *sentiment classifier* yang dapat menentukan sentimen positif, negatif, dan netral pada sebuah dokumen [6].

Pada tugas akhir ini, penelitian berfokus pada *sentiment analysis* terhadap isu sosial Kota Bandung. Penelitian ini diperlukan karena *sentiment analysis* masih didominasi dengan studi kasus *review* produk. Selain itu isu sosial memiliki struktur opini yang berbeda dengan produk atau sebuah layanan [4]. Dan kebijakan baru dari walikota yang terpilih menciptakan isu sosial yang baru juga. Penelitian ini bertujuan untuk menemukan kelas *sentiment* yang tepat dari setiap opini yang disebar pada *twitter*. Di dalam sebuah *tweets* bisa terdapat satu atau lebih kata yang menggambarkan suatu opini. Namun, kata – kata tersebut terkadang memiliki hubungan dengan kata lainnya. Di dalam sebuah isu pada satu data *tweets* dapat mengandung beberapa kata opini dengan sentimen yang berbeda-beda.

Berdasarkan masalah tersebut, dibutuhkan sebuah metode yang dapat menangani permasalahan *multi*-opini. Masalah tersebut dideskripsikan dengan metode *Holistic Lexicon-Based*. Metode ini dapat mengidentifikasi sentimen dari setiap *opinion words* yang terdapat pada data *tweet* dan dapat menangani permasalahan *multi*-opini di dalam suatu data. Metode ini merupakan peningkatan dari metode *Lexicon-Based* yang tidak bisa menangani permasalahan *multi*-opini. Di dalam penanganan masalah *multi*-opini, metode ini mengumpulkan seluruh sentimen dari kata opini berdasarkan jarak antara kata opini dengan fiturnya. Sehingga akhirnya dapat dipakai untuk menentukan kelas opini dari setiap data [5].

1.2 Perumusan Masalah

Berdasarkan pada latar belakang di atas, permasalahan yang akan diuraikan dan diteliti adalah :

1. Bagaimana cara mengidentifikasi data opini dari *tweets* dan menentukan sentimen kata opini dari setiap data menggunakan *opinion lexicon*?
2. Bagaimana cara menerapkan metode *Holistic Lexicon-Based* pada *sentiment analysis* terhadap data *tweets?*
3. Bagaimana akurasi klasifikasi sentimen berdasarkan metode *Holistic Lexicon-Based* pada *sentiment analysis* terhadap data *tweets*?

1.3 Batasan Masalah

Adapun batasan masalah untuk tugas akhir ini adalah :

1. Data yang dianalisis adalah *tweets* berbahasa Indonesia yang didapat dari situs jejaring sosial http://www.twitter.com
2. Klasifikasi opini hanya untuk sentimen opini positif dan negatif.
3. Data akan dianalisis secara *offline*, yaitu sistem tidak terhubung ke dalam jaringan internet.
4. Suatu *tweet* dinyatakan data opini jika di dalamnya terdapat minimal satu *opinion word*.
5. Metode ekstraksi kata opini dan klasifikasi yang digunakan adalah *Holistic Lecicon-Based*.
6. Penentuan *query* pencarian adalah *tweets* yang memiliki *mention* ke :
   - @infobdg
   - @infobandung
   - @pemkotbandung
   - @ridwankamil
   - @distarcipbdg
   - @pdamtirtawening
   - @dinsos_bdg
   - @dishub_kotabdg
   - @dbmpkotabdg
   - @pdkebersihan
   - @diskamtam

- @pjudbmpbdg
- #suarabdg.

1.4 Tujuan

Tujuan yang ingin dicapai dalam pengerjaan Tugas Akhir ini adalah sebagai berikut :

1. Menganalisis dan mengidentifikasi kata opini dari dataset *tweets* menggunakan *opinion lexicon*.
2. Menerapkan metode *Holistic Lexicon-Based* pada *sentiment analysis* terhadap data *tweets.*
3. Mengukur akurasi klasifikasi sentimen berdasarkan metode *Holistic Lexicon-Based* pada *sentiment analysis* terhadap data *tweets.*

1.5 Metodologi Penyelesaian Masalah

Penyelesaian masalah dilakukan dalam beberapa tahap, secara garis besar sebagai berikut :

1. Studi Literatur

Mempelajari literatur – literatur yang relevan untuk mempelajari dan memahami permasalah *sentiment analysis,* khususnya konsep dan langkah – langkah metode *Holistic Lexicon-Based*.

2. Pengumpulan Data

Pengumpulan data *tweets* berdasarkan *query* pencarian mengenai isu sosial Kota Bandung. *Tweets* yang diambil mengandung *mention* kepada user @infobdg, @infobandung, @pemkotbandung, @ridwankamil, @distarcipbdg, @pdamtirtawening, @dinsos_bdg, @dishub_kotabdg, @dbmpkotabdg, @pdkebersihan, @diskamtam, @pjudbmpbdg dan *hashtag* #suarabdg. Data yang diambil hanya *tweets* berbahasa Indonesia.

3. Perancangan Sistem

Pada tahap ini dilakukan perancangan sistem untuk analisis sentimen dan klasifikasi *tweets* ke dalam jenis sentimen dengan menggunakan metode *Holistic Lexicon-Based.*

4. Implementasi

Pada tahap ini dilakukan implementasi berdasarkan rancangan sistem yang telah dibuat sebelumnya menjadi suatu sistem klasifikasi opini berdasarkan isu sosial.

5. Analisis dan Pengujian

Pada tahap ini dilakukan analisis dan pengujian terhadap metode *Holistic Lexicon-Based* terhadap data tweets mengenai isu sosial Kota Bandung

6. Pengambilan Kesimpulan dan Pembuatan Laporan

Pada tahap ini dilakukan pengambilan kesimpulan berdasarkan analisis dan pengujian yang telah dilakukan serta pembuatan laporan tugas akhir untuk dokumentasi setiap proses kegiatan penelitian.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan pada bahasan sebelumnya, maka dapat diambil beberapa kesimpulan sebagai berikut :

1. Metode *Holistic Lexicon-Based* dapat diterapkan pada *sentiment analysis* terhadap data *tweets.*
2. Modifikasi dan pengembangan kamus mampu meningkatkan akurasi prediksi orientasi kalimat opini menjadi rata-rata 88.6% yang semula hanya berada pada rata-rata akurasi 75.3%.
3. Penanganan multi-opini pada metode *Holistic Lexicon-Based* mampu meningkatkan akurasi prediksi orientasi kalimat opini dengan peningkatan sebesar 2.64%.
4. Analisis sentimen masih sulit dilakukan pada kalimat sindiran (orientasi sebenarnya dengan orientasi berdasarkan skor kata opini berbeda).

5.2 Saran

Saran yang ingin disampaikan untuk pengembangan lebih lanjut Tugas Akhir ini adalah sebagai berikut :

1. Pengembangan POS Tagging Bahasa Indonesia dengan *corpus* yang berkaitan dengan studi kasus lain.
2. Pengembangan sistem yang dapat mengambil *dataset* dan menghasilkan *sentiment* secara *real-time* atau *online*.
3. Pengembangan kamus lebih lanjut, baik itu kamus pembakuan kata maupun kamus orientasi kata opini dapat meningkatkan akurasi prediksi kalimat opini.
4. Pengembangan kamus translasi Bahasa Sunda ke dalam Bahasa Indonesia lebih lanjut dengan studi kasus yang berkaitan dengan dataset *tweets.*
5. Analisis sentimen terhadap kalimat yang mengandung sindiran.

Daftar Pustaka

[1] Clement Yu, & Weiyi Meng. (2011). *Collaborative Research Handling Negations and Temporal Aspects for Opinion Retrieval*.

[2] Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

[3] Esuli, A., & Sebastiani, F. (2006). *SentiWordNet: A publicly available lexical resource for opinion mining*. In Proceedings of LREC.

[4] Karamibekr, M., & Ghorbani, Ali A. (2012). *Sentiment Analysis of Social Issues*. 2012 International Conference on Social Informatics.

[5] Ding, X., Liu, B., & Yu, Philip S. (2008). *A Holistic Lexicon-Based Approach to Opinion Mining*. WSDM 2008.

[6] Pak, A., & Paroubek, P. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508, F-91405 Orsay Cedex, France.

[7] Liu, B. *Opinion Mining*. Department of Computer Science, University of Illinois at Chicago, 851 S. Morgan Street, Chicago, IL 60607-0753.

[8] Fisher, R.A. (1936). *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, 7, 179–188.

[9] Twitter, Inc. (2013), "*About Twitter.*" https://about.twitter.com/ (diakses tanggal 5 November 2013).

[10] Taboada, M., Brooke, J., Tofiloski, M., Voll., K., & Stede, M. (2011). *Lexicon-Based Methods for Sentiment Analysis*. Association for Computational Linguistics.

[11] Catatan Kecil (2013), "*Sentiment Analysis Menggunakan Pendeketan Lexicon-Based*."http://adiyasan.wordpress.com/2013/02/08/*sentiment*-analysis-menggunakan-pendekatan-lexicon-based/ (diakses tanggal 31 Oktober 2013).

[12] Liddy, E. D. *In Encyclopedia of Library and Information Science, 2nd Ed*. Marcel Decker, Inc.

[13] Manurung, Ruli. Adriani, Mirna. (2008). *A survey of bahasa Indonesia NLP research conducted at the University of Indonesia*.Second MALINDO Workshop. Selangor, Malaysia: 12-13 June 2008.

[14] Mashape (2012), "[Featured API Series] Chatterbox, *Sentiment* Analysis for Social Media."http://blog.mashape.com/post/44116882888/featured-api-chatterbox-*sentiment*-analysis-fo-54402 (diakses tanggal 7 November 2013).

[15] Olson, D. and Delen, D. 2008. *Advanced Data Mining Techniques*, 1st Ed. Springer. March 2008.

[16] Rozi, Imam Fahrur., Pramono, Sholeh Hadi., & Dahlan, Erfan Achmad. (2012). *Implementasi Opinion Mining (Analisis Sentimen) untuk Ekstraksi Data Opini Publik pada Perguruan Tinggi* . Jurnal EECCIS Vol.6, No. 1, Juni 2012.

[17] Wicaksono, Alfan Farizki & Puwarianti, Ayu (2010). *HMM Based Part-of-Speech Tagger for Bahasa Indonesia.* Proceeding of the Fourth International MALINDO Workshop (MALINDO2010). Agustus 2010. Jakarta, Indonesia.

[18] Nurfalah, Adiyasa. *Analisis Sentimen Pada Opini Berbahasa Indonesia Menggunakan Pendekatan Lexicon-Based*. Laporan Tugas Matakuliah Data Mining Lanjut, Fakultas Pascasarjana Institut Teknologi Telkom, Bandung, Indonesia, 2011.

[19] eBizMBA (2013), *"Top 15 Most Popular Social Networking Sites"* http://www.ebizmba.com/articles/social-networking-websites (diakses tanggal 5 November 2013).

[20] eMarketer (2013), *"Emerging Markets Drive Twitter User Growth Worldwide"* http://www.emarketer.com/Article/Emerging-Markets-Drive-Twitter-User-Growth-Worldwide/1010874 (diakses tanggal 5 November 2013).