

KLASIFIKASI OPINI BERBASIS FITUR UNTUK REVIEW PRODUK PADA TWITTER MENGGUNAKAN HIGH ADJECTIVE COUNT ALGORITHM (HAC) DAN SUPPORT VECTOR MACHINE (SVM)

Mutiara Shabrina Utami¹, Z.k. Abdurahman Baizal², Yuliant Sibaroni³

¹Ilmu Komputasi, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Dengan semakin berkembangnya teknologi internet, maka informasi yang didapatkan semakin beragam. Beberapa informasi tekstual yang terdapat di dalam social media online seperti Twitter. Salah satu informasi yang terdapat di Twitter adalah tentang opini user terhadap suatu produk. Penggalian dan analisis dalam opini penting dilakukan karena bermanfaat terhadap identifikasi produk, perencanaan produk baru atau kecenderungan pasar bagi perusahaan, sehingga dapat diambil langkah-langkah strategis berikutnya. Namun, banyaknya opini yang tersebar membutuhkan sistem yang akan merangkum opini-opini terhadap suatu produk tersebut berdasarkan fitur-fiturnya. Proses pengklasifikasian opini tersebut dapat dilakukan dengan salah satu metode machine learning yaitu metode Support Vector Machine (SVM). SVM memiliki performansi yang lebih baik dibandingkan dengan metode machine learning lainnya pada pengklasifikasian data teks. Sedangkan untuk mengekstraksi fitur dari produk dapat dilakukan dengan mengimplementasikan High Adjective Count Algoritm. Dalam penelitian ini, penulis mengimplementasikan Support Vector Machine (SVM) dan High Adjective Count Algorithm pada data Twitter untuk menentukan kalsifikasi opini untuk setiap fitur pada satu produk. Hasil dari percobaan menunjukkan sistem dapat secara baik mengekstraksi fitur dengan akurasi ketepatan fitur yang diekstraksi sebesar 96.36% dan mengklasifikasikan opini dengan akurasi rata-rata sebesar 78,97%.

Kata Kunci : Klasifikasi Opini, Ekstraksi Fitur, Support Vector Machine (SVM), High Adjective Count Algorithm, Opinion mining

Abstract

With the rapid advancement of internet technology, the information obtained is more diverse. Some textual information contained in the online social media such as Twitter. One of the information contained in Twitter is about user opinions about a product. Datamining and analysis of the opinion is important to do because it is useful for the identification of products, new product planning or market trends for the company, so the next strategic steps can be taken. However, many widespread opinion require a system that will summarize the opinions of a product based on its features. Opinion classification process can be done with one machine learning method that is Support Vector Machine (SVM). SVM has a better performance compared to other machine learning methods on text data classification. Whereas for extracting features of the product can be done by implementing the High Adjective Count Algorithm. In this study, the authors implement Support Vector Machine (SVM) and High Adjective Count Algorithm on Twitter data to determine the opinion of calcification for each feature of the product. The results of the experiment show that the system can extract features very well with accuracy of 96.36% and classify the opinions with an average accuracy of 78.97%.

Keywords : Classification Opinion, Feature Extraction, Support Vector Machine (SVM), High Adjective Count Algorithm, Opinion mining

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Twitter adalah *social media* yang populer di mana pengguna membuat pesan status (disebut "tweet"). Sejak awal kemunculannya di tahun 2006, *Twitter* telah masuk ke dalam 10 besar halaman situs yang paling sering diakses oleh pengguna internet di seluruh dunia[2]. Sama seperti *social media* lainnya, pengguna *Twitter* biasanya menuliskan apa saja yang sedang mereka pikirkan, pandangan mengenai hal-hal politik dan agama, serta berbagai opini tentang produk/layanan yang mereka gunakan. Jumlah tweet berkembang dengan cepat dari waktu ke waktu. Pada tahun 2010, Indonesia menjadi negara ketiga terbesar penghasil tweet dengan porsi 12% persen atau sekitar 6 juta tweet per hari[14]. Volume tweet yang besar inilah yang kemudian membuat *Twitter* menjadi sumber opini dan sentimen yang cukup menjanjikan untuk keperluan marketing dan penelitian.

Opinion Mining adalah proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi[6] serta mengklasifikasi data tersebut dalam kelas opini positif, negative, dan netral. Pengklasifikasian opini dapat diaplikasikan pada *document level*, *sentence level*, dan *feature level*[4]. Pengklasifikasian data tekstual pada *document level* dan *sentence level* memberikan hasil berupa sentimen positif dan negatif secara general. Sedangkan pengklasifikasian pada *feature level* memberikan hasil yang lebih spesifik yaitu berupa sentimen positif dan negatif untuk setiap aspek atau fitur dari data yang diteliti.

Hasil survey terhadap lebih dari 2000 orang dewasa di Amerika menyatakan bahwa 81% pengguna internet melakukan penelitian terhadap suatu produk secara online minimal sekali dan 20% lainnya melakukannya setiap hari[6]. Sementara itu,

pihak perusahaan mengetahui opini masyarakat mengenai produknya dengan menggunakan cara lama seperti menyewa jasa konsultan atau melakukan survey lapangan. Fenomena-fenomena tersebut membuktikan bahwa opini tentang suatu produk/layanan yang terdapat di internet sangat berpengaruh dalam menentukan keputusan bagi pihak produsen maupun konsumen. Namun, mengumpulkan serta menganalisis opini secara manual bukanlah hal yang mudah bahkan sering kali menghabiskan anggaran yang tak sedikit. Karena hal tersebut, dibutuhkan sistem yang dapat mengelola dan menganalisis dari review konsumen khususnya pada *social media* seperti *Twitter* secara otomatis, efisien, dan hemat.

Pada tugas akhir ini, penulis akan melakukan pengklasifikasian opini (positif atau negatif) dan berdasarkan fitur terhadap produk layanan salah satu Perusahaan Telekomunikasi di Indonesia pada dataset *Twitter* berbahasa Indonesia dengan menggunakan metode *Support Vector Machine (SVM)* dan *High Adjective Count Algorithm*. Sebelumnya, metode *Support Vector Machine (SVM)* telah banyak digunakan pada beberapa penelitian mengenai *sentiment analysis* [5, 6, 7]. Sedangkan *High Adjective Counting Algorithm* dianggap berhasil dalam mengekstraksi fitur penting di dalam data[8].

Pemanfaatan hasil dari tugas akhir ini ditujukan kepada dua subjek besar, yaitu pihak produsen (Perusahaan Telekomunikasi yang menyediakan layanan) dan pihak konsumen. Pihak produsen dapat mengetahui sentimen konsumen mengenai layanan-layanan yang dimilikinya, mengevaluasi layanan apa saja yang mendapat respon positif dan respon negatif, serta mengawasi trend perubahan sentimen ketika diberlakukan suatu kebijakan baru seperti berlakunya masa promosi, berlakunya tarif baru, dan lain sebagainya. Sedangkan untuk pihak konsumen, hal ini berguna untuk mengetahui review konsumen lain tentang layanan yang ditawarkan perusahaan, dan melihat trend perubahan sentimen mengenai layanan tersebut.

1.2 PERUMUSAN MASALAH

Berdasarkan latar belakang tersebut, maka dapat dirumuskan permasalahan sebagai berikut :

1. Bagaimanakah proses ekstraksi fitur menggunakan *High Adjective CountAlgorithm* (HAC) dari dataset Twitter.
2. Bagaimanakah pengimplementasian metode *Support Vector Machine* (SVM) dalam pengklasifikasian opini (positif atau negatif) pada review konsumen di *Twitter*.
3. Bagaimanakah pengujian dan analisa akurasi dari sistem yang dibangun dengan parameter *accuracy, precision, recall, dan F-score*.

Adapun beberapa batasan masalah dalam Tugas Akhir ini adalah :

1. Dataset yang digunakan berupa data Twitter yang berhubungan dengan layanan yang ditawarkan salah satu Perusahaan Telekomunikasi di Indonesia yang didapat dari crawling data pada Twitter API.
2. Dataset Twitter yang digunakan merupakan dataset berbahasa Indonesia.
3. Dataset Twitter yang digunakan tidak memiliki arti yang ambigu.
4. Setiap data Twitter yang digunakan memiliki hanya satu buah fitur.
5. Opini yang ada di dalam dataset hanya opini positif dan opini negatif.

1.3 TUJUAN PENELITIAN

Tujuan dari penelitian tugas akhir ini adalah :

1. Mengimplementasikan *High Adjective CountAlgorithm* dalam pengekstraksian fitur dari dataset *Twitter*.
2. Mengimplementasikan metode *Support Verctor Machine* (SVM) dalam pengklasifikasian opini (positif atau negatif) pada review konsumen di *Twitter*.

3. Menguji dan menganalisa akurasi dari sistem yang dibangun dengan parameter *accuracy*, *precision*, *recall*, dan *F-score*.

1.4 METODOLOGI PENELITIAN

Metodologi yang digunakan untuk menyelesaikan tugas akhir ini adalah :

1. Identifikasi Masalah

Pada tahap ini dilakukan identifikasi dan perumusan masalah yang akan dibahas pada tugas akhir ini, yaitu mengenai klasifikasi sentimen berdasarkan fitur dari suatu layanan perusahaan pada *social media*.

2. Studi Literatur

Pada tahap ini, penulis melakukan pencarian terhadap materi-materi yang berhubungan dengan klasifikasi sentiment berdasarkan fitur menggunakan metode HAC dan SVM. Materi-materi yang didapat kemudian dipelajari untuk mendukung penyelesaian tugas akhir ini.

3. Pengumpulan Data

Pada tahap ini penulis mengumpulkan dataset *Twitter* menggunakan NodeXL yang kemudian diberi label fitur dan label opini secara manual.

4. Perancangan Sistem

Perancangan sistem dibuat sebagai rancangan proses yang akan dilakukan dalam menyelesaikan masalah yang dibahas pada tugas akhir ini.

5. Implementasi dan Pembangunan Sistem

Implementasi dan pembangunan sistem dalam tugas akhir ini akan melalui proses preprocessing, pembobotan opini, dan klasifikasi opini. Implementasi untuk perancangan sistem yang dibuat untuk pengklasifikasian sentiment berdasarkan fitur menggunakan HAC dan *Support Vector Machine (SVM)*.

6. Pengujian dan Analisis Hasil

Pengujian dan analisis akan dilakukan pada hasil yang didapat dari proses penyelesaian masalah pada tugas akhir ini.

7. Perumusan Kesimpulan dan Penyusunan Laporan Tugas Akhir

Perumusan kesimpulan dibuat berdasarkan pengujian dan analisis hasil yang didapat. Setelah diambil kesimpulan, selanjutnya akan dilakukan penyusunan laporan tugas akhir.

1.5 SISTEMATIKA PENULISAN

BAB I Pendahuluan

Bab ini berisikan penjelasan tentang permasalahan yang diangkat dalam tugas akhir ini seperti latar belakang masalah, perumusan masalah, tujuan penelitian, batasan masalah, metode penelitian, dan sistematika penulisan.

BAB II Landasan Teori

Bab ini berisikan pembahasan tentang teori-teori serta metode yang digunakan untuk menyelesaikan masalah dalam tugas akhir ini.

BAB III Perancangan Sistem

Bab ini berisikan penjelasan metode dalam menyelesaikan masalah dan gambaran sistem yang akan digunakan dalam menyelesaikan masalah pada tugas akhir ini.

BAB IV Analisis dan Hasil Pengujian

Bab ini berisikan analisis dari hasil hasil pengujian yang didapatkan.

BAB V Kesimpulan dan Saran

Bab ini berisikan kesimpulan dari penelitian tugas akhir ini dan saran untuk pengerjaan kedepannya dengan masalah yang sama.

BAB V

KESIMPULAN DAN SARAN

5.1 KESIMPULAN

Dari hasil analisis dan pengujian pada bab sebelumnya dalam tugas akhir ini, maka didapatkan kesimpulan :

1. Metode *Support Vector Machine* (SVM) memberikan hasil yang baik dengan rata-rata akurasi untuk adalah 78.9745% terhadap pengklasifikasian kelas opini. Sedangkan untuk parameter precision, recall, dan F-score memberikan nilai rata-rata 79%, 66%, dan 71.37% untuk kelas opini positif dan nilai rata-rata 70%, 83%, dan 75.5% untuk kelas opini negatif.
2. Untuk algoritma *High Adjective Count* (HAC) memberikan hasil akurasi yang baik sebesar 96.36 % terhadap fitur yang diekstraksi secara benar setelah diberikan penanganan untuk fitur implisit dan penyaringan kata benda oleh *threshold*.

5.2 SARAN

Berdasarkan percobaan yang dilakukan di atas, sebaiknya perlu diperhatikan beberapa poin sebagai berikut:

1. Proses pengambilan data tes dapat dilakukan secara online dan otomatis, sehingga tidak ada lagi proses persiapan data secara manual.
2. Mengeksplorasi lebih banyak dan lebih baik lagi kombinasi parameter pada proses klasifikasi yang memungkinkan dihasilkannya akurasi lebih baik daripada penelitian Tugas Akhir ini.
3. Mencari pola untuk kata *slang* (kata gaul; tidak baku) untuk meminimalkan penggunaan kamus data.

DAFTAR PUSTAKA

- [1] A. Go, L. Huang, and R. Bhayani. (2009). *Twitter sentiment classification using distant supervision*, in CS224N Project Report, Stanford
- [2] Alexa Top 500 Global Sites, (2012), <http://alexa.com/topsites> diakses tanggal 23 Desember 2012
- [3] Anto Satriyo, Arief Budi, and Dwi Handoko.(2003). *Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika*, Kuliah Umum IlmuKomputer.Com
- [4] Bing Liu. (2010). *Sentiment Analysis and Subjectivity*, in Handbook of Natural Language Processing.
- [5] Bing Liu. (2010). *Sentiment Analysis: A Multi-Faceted Problem*, IEEE Intelligent Systems.
- [6] Bo Pang and Lillian Lee. (2008). *Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, vol. Volume 2, no. Issue 1-2, pp. 1-135.*
- [7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. (2002). *Thumbs up? Sentiment Classification using Machine Learning*, in Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol. Volume 10, pp. 79–86, Morristown, NJ, USA.
- [8] Eirinaki, Magdalini (2011). *Feature-Based Opinion Mining and Ranking*. United States.
- [9] Hsu, Chih-Wei, et al. (2004) *A Practical Guide to Support Vector Classification*. Departement of Computer Science and Information Engineering, National Taiwan University.
- [10] Lee D.L. (1997). *Document Ranking and the Vector-Space Model*. IEEE March- April 1997
- [11] Liddy, E.D. *In Encyclopedia of Library and Information Science*, 2nd Ed. Marcel Decker, Inc

- [12] Muhammad Yusuf Nur and Diaz D santika. (2011). *Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine*, dalam Konferensi Nasional Sistem dan Informatika 2011, Bali.
- [13] Pak Alexander and Patrick Paroubek. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, in Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 13201326.
- [14] Semiocast Press Release, (2010), <http://semiocast.com/> diakses tanggal 26 April 2011
- [15] Sembiring, Krisantus. 2007. *Penerapan Teknik Support Vector Machine untuk Pendeteksian Intrusi pada Jaringan*. Institut Teknologi Bandung.
- [16] Tan Pang-Ning, Steinbach, Michael. 2005. *Introduction to Data Mining*. University of Minnesota.
- [17] Wicaksono, Alfian Farizki (2010) *HMM Based Part-of-Speech Tagger for Bahasa Indonesia*. On Proceedings of 4 th International MALINDO (Malay and Indonesian Language) Workshop.