

ANALISIS SENTIMEN MENGGUNAKAN LEXICALIZED-HIDDEN MARKOV MODEL (L-HMM)

Virani Kartika Satrioputri¹, Warih Maharani², Jondri³

¹Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Saat ini semakin banyak para pedagang yang menawarkan produknya melalui internet. Mereka biasanya meminta pelanggan untuk memberikan review atas produk yang dibelinya. Hal tersebut bertujuan untuk mempermudah calon pembeli lain menentukan produk apa yang akan dibelinya. Banyaknya review yang diberikan membuat para calon pembeli kesulitan dalam membaca satu persatu review yang ada. Oleh sebab itu, pada Tugas Akhir ini dibangun sebuah sistem yang dapat melakukan suatu mining dalam review yang ada tersebut.

Proses mining yang dilakukan disebut opinion mining atau analisis sentimen. Dimana hasil yang diharapkan adalah mengetahui apakah suatu kalimat tersebut merupakan kalimat opini positif atau negatif. Tugas Akhir ini berorientasi pada feature-based opinion dimana orientasi opini ditentukan, sebelum terlebih dahulu mengekstrak entitas produk yang dikomentari. Proses pertama yang dilakukan adalah mengekstrak fitur beserta opini yang mengikutinya, kemudian mengekstrak kalimat opini, dan terakhir menentukan orientasi opini (positif atau negatif). Metode yang digunakan dalam sistem ini adalah Lexicalized Hidden Markov Model (L-HMM) yang mengintegrasikan fitur linguistik, seperti part-of-speech, menjadi sebuah mesin pembelajaran otomatis. Sistem ini diharapkan mampu mengenali kosakata baru yang belum muncul pada data latih, yang akan diterapkan bersama dengan teknik smoothing. Dalam memperoleh data train beserta fitur, opini dan orientasinya dalam jumlah banyak, dilakukan teknik bootstrapping.

Hasil pengujian yang diberikan menunjukkan bahwa teknik bootstrapping dapat digunakan untuk memperkaya data latih yang digunakan. Selain itu, penggunaan L-HMM mampu mengekstraksi dengan baik fitur beserta orientasi opini dalam suatu kalimat.

Kata Kunci : data mining, sentiment analysis, feature-based opinion, part-of-speech, lexicalized hidden markov models (L-HMM)

Telkom
University

Abstract

Nowadays, there are merchants selling their product on the web. They usually ask their customer to give a review about product they have purchased. It will make another potential customer being easier to make a decision for buying a product they will bought. The amounts of the review will make the potential customer difficult for reading each of the review. Therefore, this thesis aims to mining an opinion on the review.

Process mining in this thesis is called opinion mining or sentiment analysis. Where the output was to determine whether a sentence is a sentence positive or negative opinions. This thesis oriented feature-based opinion in which subjective orientation is determined, prior to first extract the product entities commented. The first process is to extract features and their opinions are followed, then extract subjective sentences, and finally determine the orientation of opinion (positive or negative). The method used in this system is Lexicalized Hidden Markov Model (LHMM) that integrates linguistic features, such as part-of-speech, into an automated machine learning. The system is expected to be able to recognize the new vocabulary trainer has not shown up in the data, which will be implemented together with the smoothing technique. In obtaining the data train and its features, opinion and orientation in large quantities, do bootstrapping techniques.

The test results show that given bootstrapping techniques can be used to enrich the data used directly trainer. In addition, the use of L-HMM capable of extracting features and their orientation with good opinions in a sentence.

Keywords : data mining, sentiment analysis, feature-based opinion, part-of-speech, lexicalized hidden markov models (L-HMM)

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Penggunaan internet semakin hari berkembang semakin pesat. Saat ini hampir seluruh perusahaan memanfaatkan internet dalam mempromosikan produk yang mereka jual, bahkan tidak sedikit yang menggunakan internet dalam melakukan transaksi jual beli. Pemanfaatan media elektronik dalam perdagangan biasa disebut *e-commerce* (*electronic commerce*).

Ekspansi yang cepat dari penggunaan *e-commerce*, menjadikan semakin banyak orang yang membeli produk secara *online*. Merupakan hal biasa jika pedagang online meminta pelanggan dalam membagi pengalaman mereka dalam menggunakan produk yang dijual. Banyaknya pelanggan yang memberikan *review* produk menyebabkan kesulitan dalam membaca *review* produk tersebut, terlebih lagi untuk item produk yang populer. Pelanggan akan kesulitan dalam memutuskan item produk mana yang akan dibelinya.

Kesulitan pelanggan dalam membaca *review* produk dalam jumlah banyak, menyebabkan perlunya dilakukan suatu analisis sentimen dalam melihat ekspresi dan opini pelanggan dalam tinjauan produk yang mereka buat, apakah positif atau negatif. *Sentiment analysis* atau *opinion mining* mengacu pada aplikasi dari *Natural Language Processing (NLP)*, *Computational Linguistic*, dan *text analytics* untuk mengidentifikasi dan mengekstrak informasi subjektif dari material sumber [3]. Dalam melihat sebuah *review*, pelanggan tidak hanya melihat opininya saja, namun juga melihat fitur yang dikomentari. Oleh sebab itu, Tugas Akhir ini mengacu pada fitur serta orientasi opini yang mengikutinya, atau biasa disebut *feature based opinion mining* dimana ekstraksi fitur akan dilakukan pada level kalimat. Fitur tersebut mengacu pada *component*, *feature*, dan *function* dari produk yang dikomentari. Ekstraksi fitur dalam suatu kalimat opini tersebut dapat mempermudah pelanggan dalam menemukan produk yang sesuai dengan keinginannya.

Dalam melakukan analisis sentimen ini, dibentuk suatu kerangka kerja baru yang secara alami merupakan integrasi antara model leksikal dengan *Hidden Markov Model (HMM)*. HMM kemudian akan diintegrasikan dengan model leksikal, *Lexicalized Hidden Markov Model (L-HMM)*. Berbeda dari metode lainnya yang hanya memanfaatkan metode statistik ataupun metode leksikal saja, L-HMM yang merupakan kombinasi keduanya mampu memperkaya model HMM, dimana kata-kata yang bersifat "luar biasa", yaitu kata-kata yang probabilitas distribusinya dalam kategori tertentu berbeda dari yang lain pada suatu analisis sentimen, akan dibuatkan state baru untuk memperkenalkan kata-kata tersebut dalam konteks.

Metode L-HMM yang digunakan, tidak hanya dapat mengenali suatu kalimat opini saja, namun juga dapat digunakan dalam menentukan orientasi opini yang dikenali, apakah opini positif atau opini negatif. Selain itu, metode L-HMM ini juga dapat mengenali pasangan opini beserta fiturnya sekaligus. Metode L-HMM ini kemudian akan dikombinasikan dengan teknik bigram *smoothing*, yaitu teknik yang

digunakan jika menghasilkan probabilitas 0. Teknik ini digunakan untuk membantu perhitungan peluang, sehingga hasil akan semakin baik dengan memaksimalkan kondisi peluang yang ada.

1.2 Perumusan Masalah

Permasalahan yang akan dibahas pada Tugas Akhir ini adalah :

1. Bagaimana mengintegrasikan model leksikal dengan metode statistik menggunakan *Lexicalized Hidden Markov Models* (L-HMM) dalam pengenalan kalimat opini?
2. Bagaimana menggali fitur produk beserta orientasi opini yang mendeskripsikan fitur tersebut dalam sebuah kalimat?
3. Bagaimana pengaruh penggunaan teknik *smoothing* dalam penentuan fitur produk dan orientasi opini?

1.3 Batasan masalah

Batasan masalah yang diberikan dalam pengerjaan Tugas Akhir ini adalah :

1. Dataset yang akan digunakan berasal dari *online review* Amazon.com
2. Teknik *smoothing* yang digunakan adalah teknik *smoothing* Jelinec-Mercer.

1.4 Tujuan

Tujuan Tugas Akhir ini adalah :

1. Mengintegrasikan model leksikal dengan metode statistik dalam *Lexicalized Hidden Markov Model* (L-HMM) dalam pengenalan kalimat opini.
2. Menggali fitur produk beserta orientasi opini yang mendeskripsikan fitur tersebut dalam sebuah kalimat.
3. Mengetahui pengaruh penggunaan teknik *smoothing* dalam penentuan fitur produk dan orientasi opini.

1.5 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang akan dilakukan pada penelitian Tugas Akhir ini terdiri dari 6 tahap. Tahapan-tahapan tersebut adalah sebagai berikut :

1. Studi literatur
Melakukan proses pencarian dan pengumpulan referensi yang berkaitan dengan analisis sentiment, *Lexicalized Hidden Markov Model* (L-HMM) dan teknik bootstrapping melalui buku, jurnal, dan referensi lainnya untuk kemudian dipelajari lebih dalam.
2. Pengumpulan data
Pada tahap ini akan dilakukan pengumpulan data yang berasal dari *online review*, yaitu Amazon.com
3. Analisis dan perancangan sistem

Melakukan analisa kebutuhan dan perancangan sistem yang akan dibangun sesuai dengan identifikasi kebutuhan. Perancangan sistem yang akan dibangun adalah sebagai berikut:

- a. *Pre-processing* : Melakukan penyisiran data, kemudian data dibersihkan sehingga menghasilkan suatu kumpulan tulisan yang akan ditinjau
 - b. *Evaluation Data* : Melakukan *tagging* berdasarkan *Part-of-Speech* (POS) oleh system ahli
 - c. *Training Corpus* : Melakukan *tagging* berdasarkan *Part-of-Speech* (POS) dan sebagian lagi dilakukan *tagging manual* untuk kemudian dilakukan *bootstrapping* data
 - d. *Bootsrapping data* : Melakukan pendekatan *bootstrapping* untuk mengekstrak data agar dapat melakukan pembelajaran mandiri
 - e. *Classifier* : Melakukan klasifikasi data berdasarkan metode *Lexicalized Hidden Markov Model* (L-HMM) kemudian melakukan *decoding* untuk mencari pola tag terbaik
 - f. *Post-processing* : Melakukan penghapusan kalimat yang bukan merupakan ekspresi dari pelanggan dan untuk entitas produk yang berbeda serta menentukan orientasi suatu kalimat opini, apakah merupakan kalimat positif atau kalimat negatif.
4. Implementasi dan pembangunan sistem
Melakukan implementasi dari hasil perancangan yang telah dibuat sebelumnya. Implementasi ini dilakukan dengan membangun sebuah sistem berupa program yang mampu memberikan label terhadap dataset yang diberikan.
 5. Pengujian sistem dan analisa hasil
Melakukan pengujian sistem yang telah dibuat, apakah sesuai dengan perancangan sistem yang diinginkan. Kemudian melakukan analisa terhadap hasil empirik yang diberikan oleh sistem.
 6. Pengambilan kesimpulan dan pembuatan laporan
Tahap ini merupakan tahap terakhir, dimana berdasarkan analisis dari hasil yang diberikan saat pengujian, ditarik kesimpulan dan juga dilakukan proses dokumentasi sebagai laporan untuk Tugas Akhir.

Telkom
University

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil berdasarkan pengujian dan analisis yang telah dilakukan pada Tugas Akhir ini adalah sebagai berikut :

1. Metode *Lexicalized Hidden Markov Models* (L-HMM) dapat digunakan dengan baik untuk melakukan ekstraksi fitur dan orientasi opini
2. Mengkombinasikan metode leksikal dengan metode statistik pada L-HMM dalam mengekstraksi fitur dan orientasi opini dilakukan dengan memberikan tag pada masing-masing kata dengan memperhatikan jumlah kemunculan pada masing-masing kata
3. Ekstraksi fitur dan orientasi opini menunjukkan hasil yang lebih baik jika menggunakan teknik *smoothing* daripada yang tidak menggunakan teknik *smoothing* dengan nilai *precision* 74.77% untuk ekstraksi fitur dan 79.70% untuk orientasi opini
4. Proses bootstrapping yang dilakukan berpengaruh dengan baik dalam memperbaiki nilai *precision* yang diberikan, dimana perbedaan maksimal nilai *precision* untuk ekstraksi fitur mencapai 25.08% dan maksimal nilai *precision* untuk orientasi opini mencapai 27.08%
5. Dari ketiga parameter koefisien interpolasi yang diberikan, yaitu λ , β , dan α , dapat dilihat bahwa semakin besar nilai parameter tersebut, maka semakin besar pula nilai *precision*, *recall*, dan *f-measure* yang diberikan
6. Dari ketiga parameter koefisien interpolasi, nilai λ memiliki tingkat pengaruh yang paling tinggi, dibanding dengan α dan β , yang berarti bahwa teknik *smoothing* akan sangat berpengaruh terhadap urutan nilai POS tag yang ada

5.2 Saran

Saran yang dapat dipertimbangkan dalam pengembangan sistem lebih lanjut adalah sebagai berikut :

1. Pengembangan sistem agar dapat mengenali seberapa besar tingkat orientasi opini, apakah termasuk opini “lemah” atau “kuat” dengan menambahkan tag “intensifier” agar kata-kata tersebut dapat dikenali.

6. DAFTAR PUSTAKA

- [1] Agrawalr, T, I., & A, S. (1993). Mining Association Rules between Sets of Items in Large Databases. *ACM Sigmoid Conference*, (pp. 207-216). Washington DC.
- [2] Bilmes Jeff. 2002. What HMMs Can Do. UWEE Technical Report Number UWEETR-2002-0003.
- [3] Fu Guohong dan Kang-Kwong Luke. Chines Named Entity Recognition using Lexicalized HMMs. *SIGKDD Explorations Volume 7*, Issue 1.
- [4] Hu Mingqing dan Bing Liu. 2004. Mining and Summarizing Customer Reviews. *KDD'04*, August 22–25, 2004, Seattle, Washington, USA.
- [5] Hu Mingqing, dan Bing Liu. 2004. Mining Opinion Features in Customer Reviews. University of Illinois, Chicago.
- [6] Jin, W., & Ho, H. H. A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining. U.S.A.
- [7] Jin Wei, Hung Hay Ho, dan Rohini K. Srihari. OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction. *KDD'09*, June 28–July 1, 2009, Paris, France.
- [8] Leung Cane W. K. dan Chan Stephen C. E. 2009. Sentiment Analysis of Product Reviews. IGI Global.
- [9] Lee Sang-Zoo, Jun-ichi Tsuji, dan Hae-Chang Rim. Lexicalized Hidden Markov Models for Part-of-Speech Tagging.
- [10] Liddy, E. D. *In Encyclopedia of Library and Information Science, Second Edition*. Marcel Decker, Inc.
- [11] Liu, B. (2008). Opinion Mining. *Invited contribution to Encyclopedia of Database Systems*. Chicago.
- [12] Liu Bing. 2010. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing, Second Edition*, (editors: N. Indurkha and F. J. Damerau).
- [13] Nur Muhammad Yusuf dan Diaz D Santika. Analisis Sentimen pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine. *Konferensi Nasional Sistem dan Informatika 2011*.
- [14] Pang Bo dan Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008)* 1–135.
- [15] Pla Ferran dan Antonio Molina. 2004. Improving Part-of-Speech Tagging Using Lexicalized HMMs. Cambridge University Press.
- [16] Qi Luole dan Li Chen. 2010. A Linear-Chain CRF-Based Learning Approach for Web Opinion Mining. Hong kong Baptist University, China.
- [17] Qi Luole dan Li Chen. Comparison of Model-Based Learning Methods for Feature-Level Opinion Mining. 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

- [18] Rabiner Lawrence R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, VOL. 77, No. 2, February 1989.
- [19] Rentoumi Vassiliki, George Giannakopoulos, dan Vangelis Karkaletsis. Sentiment Analysis of Figurative Language using a Word Sense Disambiguation Approach. International Conference RANLP 2009 - Borovets, Bulgaria, pages 370–375.
- [20] Taboada Maite, dkk.2011. Lexicon-Based Methods for Sentiment Analysis. Association for Computational Linguistic.
- [21] Taras, Z., & Carroll, J. (2008). Automatic Seed Word Selection for Unsupervised Sentiment. *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester.
- [22] Turney Peter D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL), Philadelphia, July 2002. Pp. 417-424
- [23] Wicaksono Alfani Farizki dan Ayu Purwarianti. HMM Based Part-of-Speech Tagger for Bahasa Indonesia. Institut Teknologi Bandung, Bandung.
- [24] Zhuang Li, Feng Jing, dan Xiao-yan Zhu. Movie Review Mining and Summarization.