KLASIFIKASI DOKUMEN BEROPINI ME NGGUNAKAN METODE NAÏVE BAYES DAN METODE CATEGORICAL PR OPORTIONAL DIFFERENCE

Melliana Merina[1], Warih Maharani[2], Imelda Ataina[3]

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

Abstrak

Analisis sentimen merupakan proses me ngidentifikasi opini di dalam sebuah dokumen apakah mengandung sentimen kemudian mengklas ifikasikannya ke dalam kelas positif atau negatif . Untuk tahapan preprocessi ng dilakukan proses – proses seperti : tokenisasi, stopwords, stemming dan NLP (Natural Language Processing) menggunakan SentiWordNet. Untuk mengurangi jumlah fitur yang sangat besar da lam proses mining-nya dapat menggunakan sebuah metode seleksi fitur yakni metode Categorical Proportional Difference . Klasifier menggunakan metode Naïve Bayes

Hasil pengujian menunjukkan bahwa tingkat perf ormansi metode Naïve Bayes sebagai klasifier semakin baik dilihat dari pada percobaan ke-7 dimana nilai akurasi kelas prediksi tertinggi sebesar 93.57%. Untuk metode Categor ical Proportional Difference sebagai metode seleksi fitur menghasilkan tingka t performansi cukup baik dimana pada percobaan ke-2 nilai akurasi tertinggi sebesar 91.43% dengan nilai threshold=1.

Kata Kunci : Analisis sentimen, NLP, Categorical Propor tional Difference, Naïve Bayes, Akurasi

Abstract

Sentiment analysis is a process of iden tifying opinion in a document which contains sentiments then classify into positive or nega tive classes. For prepro cessing performed process such as: tokenization, stopwords , stemming and NLP (Natural Language Processing) uses SentiWordNet. To reduce the feat ures which consist of very la rge number in the mining process we can use a feature selection method calle d Categorical Proportional Difference method. Classifier uses Naïve Bayes method The test results showed that the performan ce level of Naïve Bayes method as klasifier getting good views of the 7th trial in which the value of cl ass prediction accuracy of 93.57% the highest. To metod CPD as a met hod of feature selection produ ces quite good performance level in which the trial 2nd highest accuracy va lue of 91.43% with a th reshold value = 1.

Keywords : :Sentiment Analysis,NLP,Categorical Proportional Difference, Naïve Bayes, Accuracy

1. Pendahuluan

1.1 Latar Belakang Masalah

Dewasa ini perkembangan data mining sangat pesat hal ini tidak lepas dari perkembangan teknologi informasi yang pesat pula.Internet sebagai sarana untuk media-media online dalam memberikan informasi yang cepat dan akurat pasti memiliki jumlah data yang sangat besar di dalam database-nya. Bentuk data tersebut dapat dikategorikan ke dalam beberapa jenis antara lain: teks, gambar, audio maupun video. Untuk data teks sendiri, data dibagi ke dalam 2 kategori: fakta dan opini [6]. Fakta dalam istilah keilmuan merupakan suatu hasil pengamatan objektif dan dapat dilakukan verifikasi oleh siapapun sedangkan opini merupakan ekspresi subjektif yang menggambarkan sentimen, pendapat atau perasaan tentang sebuah entitas, kejadian atau sifat [13]. Untuk proses mining datanya dapat meliputi berbagai tahapan dan cara. Salah satu cara yang digunakan untuk menambang datanya adalah analisis sentimen. Analisis sentimen atau *opinion mining* merupakan bagian dari cabang keilmuan data mining yang berfungsi untuk mengidentifikasi opini publik terhadap suatu keadaan di lingkungannya dengn melihat opini terhadap suatu masalah sehingga dapat diidentifikasi kecendrungan suatu pasar [10,13].

Analisis sentimen atau *opinion mining* mulai popular setelah *paper* B.Pang dan L.Lee pada tahun 2002 dipublikasikan. Ide dibalik penelitian yang dilakukan analisis sentimen adalah proses menyajikan informasi dengan membangun sebuah sistem yang dapat mengklasifikasikan dokumen teks ke dalam dua kategori, yakni positif dan negatif yang sesuai dengan keseluruhan sentimen yang dinyatakan di dalam setiap dokumen tersebut [9]. Metode klasifikasi dalam analisis sentimen menggunakan metode-metode klasifikasi yang biasa digunakan untuk kategorisasi teks antara lain metode *supervised learning* seperti *Naïve Bayes (*NB*), Support Vector Machine (*SVM*), Rule Based Approach* maupun *Maximum Entrophy*. Untuk metode Naïve Bayes dan SVM sendiri metode ini terbukti menghasilkan nilai akurasi lebih baik jika dibandingkan dengan metode berbasis teks lainnya [9]. Dalam proses untuk melabelkan kelas ke dalam sentimen positif atau negatif dari sebuah korpus dokumen teks, jumlah data yang dimiliki sebuah korpus teks biasanya terdiri dari ratusan bahkan ribuan data. Hal ini dapat menghasilkan jumlah fitur data yang sangat besar pula. Besarnya jumlah fitur data yang digunakan dalam proses klasifikasi dapat mempengaruhi tingkat performansi klasifikasi itu sendiri karena dari sekian banyak fitur yang digunakan kenyataannya banyak sekali fitur yang sebenarnya tidak penting dalam proses klasifikasi namum ikut ke dalam proses klasifikasinya. Dalam analisis sentimen dokumen-dokumen yang akan diklasifikasikan harus mengandung unsur subjektif (opini) di dalam teksnya sehingga ketika diklasifikasikan, dokumen teks yang mengandung unsur subjektif saja yang bisa ditentukan apakah teks tersebut mengandung sentimen [1]. Untuk mengatasi permasalahan tersebut maka digunakan sebuah metode seleksi fitur yang dapat menyeleksi fitur-fitur yang dirasa penting saja untuk dimasukkan ke dalam proses klasifikasi

serta dapat mengurangi jumlah fitur yang sangat besar tersebut. Selain itu, metode seleksi fitur juga memiliki keuntungan dalam meningkatkan efisiensi dan akurasi sebuah mesin pembelajaran pada *supervised* klasifikasi teks yang dibangun dengan mengekstrak sebuah dokumen kemudian mendapatkan sebuah set fitur yang relevan [11].

Penelitian yang dilakukan di dalam Tugas Akhir ini menggunakan metode seleksi fitur *Categorical Proportional Difference* (CPD). Metode CPD adalah mengukur sejauh mana kontribusi kata untuk membedakan kategori tertentu dari kategori lainnya dalam sebuah korpus teks [11]. Dengan kata lain CPD untuk sebuah kata dalam kategori tertentu dalam korpus teks adalah rasio yang memperhatikan jumlah dokumen dari kategori tertentu di mana kata itu muncul dan jumlah dokumen dari kategori lainnya di mana kata itu juga terjadi.Untuk menyeleksi fitur-fitur hasil ekstraksi, CPD menyeleksi berdasarkan jumlah kemunculan tiap fitur pada setiap dokumen teksnya. Fitur-fitur yang sering muncul hampir di seluruh dokumen teks dianggap sebagai fitur yang tidak penting (semakin kecil kegunaannya dalam proses klasifikasi), hal ini ditandai dengan besar nilai cpd fiturnya yang mendekati -1 dan sebuah fitur dianggap penting ditandai dengan besar nilai cpd fiturnya yang mendekati atau sama dengan 1. Menggunakan beberapa nilai *threshold* (ambang batas) yang berbeda untuk menghapus/membuang fitur-fitur yang mempunyai nilai cpd lebih kecil atau sama dengan nilai *threshold*-nya [9]. Dalam Tugas Akhir ini digunakan nilai *threshold* antara [0-1] dimana nilai *threshold*=1 merupakan nilai cpd maksimal yang dimiliki oleh fitur sedangkan nilai *threshold*=0 dianggap sebagai nilai cpd terendah yang dimiliki fitursehingga fitur dengan nilai cpd kurang dari nilai *threshold* akan dihapus.

Penggunaan metode klasifikasi dalam Tugas Akhir ini adalah menggunakan metode *Naïve Bayes*. Metode *Naïve Bayes* adalah metode klasifikasi teks yang menerapkan Teorema Bayes dengan asumsi keindependenan atribut dalam mengklasifikasikan sebuah dokumen teks[8]. Metode ini sudah banyak digunakan baik untuk penelitian di kategorisasi teks maupun penelitian mengenai analisis sentimen dan terbukti efektif, sederhana, cepat dan menghasilkan akurasi yang tinggi [3].

1.2 Rumusan Masalah

Adapun rumusan masalah dalam Tugas Akhir ini adalah:

1. Bagaimana cara mengimplementasikan sistem yang mampu melakukan proses klasifikasikan dokumen teks.
2. Bagaimana caramengimplementasikan sistem yang mampu menyeleksi fitur-fitur yang dianggap penting atau tidak penting untuk di-inputkan ke dalam proses klasifikasi sehingga dapat mengurangi besarnya dimensi fitur yang dihasilkan selama proses ekstraksi fiturnya.

3. Bagaimana cara menentukan tingkat performansi sistem yang telah dibangun dalam mengklasifikasikan dokumen teks.

Adapun batasan masalah pada Tugas Akhir ini adalah:

1. Dataset yang digunakan adalah dokumen *movie review* Bo Pang/Lee 2002.
2. Dataset yang diproses adalah data dalam bahasa inggris.
3. Tidak menangani opini implisit.
4. Tidak menangani negasi suatu data.
5. Klasifikasi hanya terhadap dua kategori yakni kategori positif dan kategori negatif.

1.3 Tujuan

Tujuan dari dilakukannya Tugas Akhir ini adalah :

1. Menerapkan metode klasifikasi *Naïve Bayes* (NB) untuk mengklasifikasikan dokumen teks ke dalam kelas positif dan negatif.
2. Menerapkan metode seleksi fitur *Categorical Proportional Difference* (CPD) untuk menyeleksi fitur-fitur yang dianggap penting atau tidak penting untuk digunakan ke dalam proses klasifikasinya.
3. Menghitung dan menganalisis hasilproses klasifikasi yang dilakukan sistem dengan menggunakan pparameter uji performansi klasifikasi teks seperti *precision, recall, F-Measure* dan *Accuracy*.

1.4 Metodologi Penyelesaian Masalah

Adapun metodologi penyelesaian masalah yang digunakan dalam Tugas Akhir ini adalah:

1. Studi Literatur

   Melakukan pencarian serta mempelajari informasi, referensi dan pembelajaran yang berhubungan dengan Data Mining, Analisis Sentimen, *Naïve Bayes Classifier*dan metode *Categorical Proportional Difference*untuk seleksi fitur.
2. Pengumpulan dan Pengolahan Data

   Mengumpulkan dan memahami data-data yang akan digunakan untuk mendukung penyelesaian masalah. Data yang digunakan adalah data *movie review* yang digunakan dalam penelitian Pang/Lee/Vaithyanathan 2002.
3. Analisis dan Perancangan

   Menganalisis dan merancang permasalahan masalah yang akan diselesaikan. Analisis kebutuhan sistem dan perancangan perangkat lunak *objek-oriented* menggunakan *flowchart*.
4. Implementasi

Pada tahap ini dilakukan implementasi terhadap hasil analisis dan perancangan sistem yang telah dibangun. Tahapan yang dilakukan adalah tahap *preprocessing* seperti: tokenisasi, *stopword*, *Natural Language Processing* (NLP) dan *stemming*. Untuk metode seleksi fitur digunakan metode *Categorical Proportional Difference* (CPD) dan menggunakan metode *Naïve Bayes* sebagai klasifier dokumen teksnya.

5. Pengujian dan Analisis

   Melakukan pengujian terhadap hasil implementasi sistem yang telah dibangun berdasarkan scenario pengujian yang telah ditentukan.Menggunakan parameter uji yang telah ditentukan seperti *precision, recall, F-measure* dan *Accuracy*untuk melihat tingkat performansi klasifikasi yang dibangun.
6. Kesimpulan dan Saran

   Proses pengambilan kesimpulan, saran dan penyusunan laporan Tugas Akhir dari hasil pengujian dan analisis yang telah dilakukan.

5. Penutup

5.1 Kesimpulan

Berdasarkan analisis hasil pengujian, diperoleh kesimpulan sebagai berikut:

1. Sistem yang dibangun dengan menggunakan metode Naïve Bayes mampu mengklasifikasikan dokumen beropini dengan tingkat performansi yang baik. Hasil kelas prediksi yang dihasilkan sistem memiliki nilai akurasi yang tinggi.

2. Sistem klasifikasi teks yang menggunakan Sentiwordnet dalam menentukan kata-kata yang mengandung sentimen di dalam sebuah korpus dokumen teks dapat mempengaruhi tingkat akurasi sistem dalam mengklasifikasikan dokumen beropini ke dalam kelas positif dan negatif.

3. Metode *Categorical Proportional Difference* dengan menggunakan nilai *threshold* untuk menentukan fitur-fitur yang dianggap penting dalam proses klasifikasi menunjukkan bahwa fitur dengan nilai cpd 1 adalah fitur dengan peranan paling penting karena fitur-fitur dengan nilai cpd 1 menghasilkan tingkat performansi sistem yang semakin baik dibandingkan fitur-fitur dengan nilai cpd lebih kecil dari 1.

5.2 Saran

Berdasarkan hasil analisa dan kesimpulan, terdapat beberapa saran untuk perbaikan pada penelitian berikutnya, di antaranya sebagai berikut:

1. Menggunakan model atau tehnik penggunaan *opinion lexican* yang lain untuk menentukan kata-kata yang mengandung sentimen di dalam sebuah korpus teks dalam hal ini terutama NLP khusus bahasa Inggris yakni Sentiwordnet.

2. Menggunakan contoh pemakaian dataset yang berbeda dalam proses klasifikasi Naïve Bayes dalam penelitian terhadap analisis sentimen.

DAFTAR PUSTAKA

[1] Abbasi.A, Hsinchun,C and Arab,S.2008. Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums. ACM Trans. Inf. Sydney.,26(3):1-34

[2] Baccienella,S. Esuli,A and Sebastiani,F. Sentiwordnet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Istituto di Scienza e Tecnologie dell'Informazione.Consiglio Nazionale delle Ricerche. Via Giuseppe Moruzzi 1, 56124 Pisa, Italy

[3] Darujati,C dan Gumelar, A Bimo. Pemanfaatan Teknik Supervised Untuk Klasifkasi Teks Bahasa Indonesia. ISSN 1858-4667. 2012.

[4] Destuardi,I dan Sumpeno,S. Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan Metode Naïve Bayes. Seminar Nasional Pascasarjana, Institut Teknologi Sepuluh Nopember, 13 Agustus 2009.

[5] Falani,A.Zakki. Knowledge Discovery in Database.www.mfile.narotama.ac.id/files/..

[6] Liu,B (2010). Sentiment Analysis: A Multi-Faceted Problem, *IEEE Intelligent Systems*.

[7] Liu.B. (2010). Sentiment Analysis and Subjectivity, in *Handbook of Natural Language Processing*.

[8] M Ammar Shadiq. Keoptimalan Naïve Bayes Dalam Klasifikasi. Program Ilmu Komputer FPMIPA Universitas pendidikan Indonesia.

[9] O'Keefe,T and Koprinska,I. Feature Selections and Weighting Methods in Sentiment Analysis. Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

[10] Pang,B. Lee,L and Vaithyanathan,S. (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

[11] Simeon,M and Hilderman,R. Categorical Proportional Difference: A Feature Selection Method for Text Categorization. In AusDM, pages 201–208,2008.

[12] Tan, Pang-Ning., Steinbach,M dan Kumar,V .2005. Introduction to Data Mining. Minnesota:University of Minnesota. http://www-users.cs.umn.edu/~kumar/dmbook/index.php

[13] Yusuf, M Nur dan S.Diaz D. ANALISIS SENTIMEN PADA DOKUMEN BERBAHASA INDONESIA DENGAN PENDEKATAN *SUPPORT VECTOR MACHINE*. Konferensi Nasional Sistem dan Informatika 2011; Bali, November 12, 2011.KNS&I11-002

[14] polarity dataset v0.9 (2.8Mb) (includes a README):.700 positive and 700 negative processed reviews.Introduced in Pang/Lee/Vaithyanathan EMNLP 2002.Released July 2002.http://www.cs.cornell.edu/people/pabo/movie-review-data/.