Sentiment Analysis Berbahasa Indonesia Menggunakan Improved Multinomial Naive Bayes

Indonesian Sentiment Analysis Using Improved Multinomial Naïve Bayes

Muhammad Adib Imtiyazi[1], Shaufiah, ST., MT[2], Moch. Arif Bijaksana, Ir., M.Tech[3]

Fakultas Informatika, Universitas Telkom, Bandung

[1]adibim@students.telkomuniversity.ac.id [2]shaufiah@telkomuniversity.ac.id

[3]arifbijaksana@telkomuniversity.com

ABSTRAKSI

Penggunaan *Multinomial Naïve Bayes* sebagai *classifier* dalam kasus *sentiment analysis* sudah jamak dilakukan, namun penggunaan TF-IDF sebagai *feature weighting* dalam kasus tersebut dirasa kurang sesuai karena pada kasus *sentiment analysis*, karena sifat dari TF-IDF itu sendiri yang lebih mementingkan *frequency* kemunculan kata. Oleh karena itu, digunakanlah algoritma *Improved Multinomial Naïve Bayes* yang menggunakan *Improved Gini Index* (TF-iGini) dalam pembobotan yang dianggap lebih tepat sehingga dapat menghasilkan performansi yang lebih baik.

Pada tugas akhir ini dilakukan perbandingan performansi dari Multinomial Naïve Bayes + TF-IDF dan Multinomial Naïve Bayes + TF-iGini. Hasil percobaan menunjukkan bahwa iGini mampu memberikan hasil yang cukup baik jika dibandingankan dengan IDF pada Multinomial Naïve Bayes, namun tidak cukup baik untuk menggantikan fungsi IDF dalam kasus klasifikasi *sentiment*.

Kata kunci: *sentiment analysis, feature weighting, IDF, improved gini, Multinomial Naïve Bayes, Bahasa Indonesia.*

ABSTRACT

Multinomial Naïve Bayes as a classifier for cases in sentiment analysis is commonly used, but the usage of TF-IDF as feature weighting considered unsuitable at those case, this is because the characteristics of TF-IDF that considered term frequency is priority. Thus, Improved Multinomial algorithm is used combined with Improved Gini Index (iGini) in weighting because considered more suitable and able to produce better performances.

In this final project, comparison of performances is done between Multinomial Naïve Bayes + TF-IDF and Multinomial Naïve Bayes + TF-iGini. The results from experiments shows that iGini able to produce quite good performance compared with IDF on Multinomial Naïve Bayes, but not good enough to substitute IDF on sentiment classification cases.

Keywords: *sentiment analysis, feature weighting, IDF, improved gini, Multinomial Naïve Bayes, Indonesian.*

1. PENDAHULUAN

Perkembangan teknologi yang terjadi secara masiv dan cepat menyebabkan banyak perubahan pada kehidupan kita, terutama semenjak peradaban manusia memasuki era digital. Pada era digital ini berkembanglah berbagai macam media yang mampu menghantarkan informasi dengan cepat, terjangkau, dan mudah untuk diakses, diantaranya radio, televisi, telepon genggam, hingga yang paling mutakhir, Internet. Berbagai macam media ini mampu memanjakan kita sebagai pengguna dengan aliran informasi teraktual 24 jam sehari tanpa henti.

Kemudahan dalam pengkasesan media terutamanya internet, membuat pengguna pada akhirnya turut berpartisipasi untuk menyumbang konten [1], dan salah satu konten yang banyak dan mudah untuk ditemui adalah opini atau sentiment. Opini sendiri merupakan sebuah kalimat subjektif yang berisi persepsi seseorang terhadap sebuah objek atau peristiwa [2]. Konten bertipe text ini sangat berguna bagi seorang calon konsumen yang ingin mengetahui pendapat konsumen lainnya yang telah memiliki pengalaman terhadap sebuah objek atau peristiwa. Selain itu, opini menjadi sangat berarti bagi pemilik produk sebagai bahan evaluasi terhadap produk yang mereka miliki. Namun kesempatan ini memunculkan masalah baru, yaitu kebutuhan akan pengkategorian opini positif dan negatif secara cepat dan tepat dikarenakan masifnya data yang ada, sehingga muncul sebuah metode dalam pengkategorian opini yang disebut dengan sentiment analysis [3].

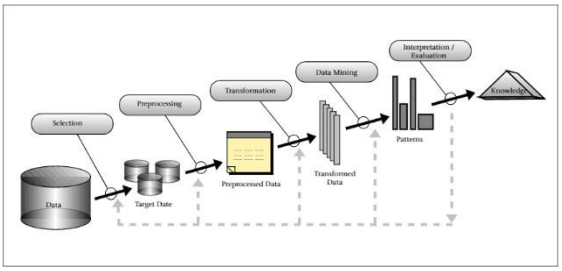
Sentiment analysis sebagai bagian dari text mining mampu menjadi solusi dalam pengkategorian opini secara otomatis berdasarkan learning terhadap data yang telah dipersiapkan sebelumnya. Namun permasalahan besar yang muncul pada kasus text mining adalah tingginya dimensionalitas (feature). Dimensi yang tinggi, terlebih lagi dengan banyaknya yang tidak berkaitan dengan kategorisasi dapat menjadi noise pada data dan menurunkan tingkat akurasi dalam klasifikasi [4]. Multinomial Naïve Bayes sebagai salah satu algoritma yang sering dipakai untuk kasus text mining menggunakan TF-IDF untuk melakukan feature weighting, namun ini masih dianggap kurang maksimal karena pada kasus text mining dikarenakan sifat aslinya yang digunakan pada data bertipe kategorikal. Pada sentiment analysis, setiap features memiliki pengaruh yang berbeda-beda terhadap hasil kategorisasi [5], karenanya diperlukan perubahan terhadap fitur ini agar mampu menyesuaikan terhadap kasus Sentiment Analysis.

Tujuan dari Tugas Akhir ini adalah mencoba menerapkan algoritma Gini Index untuk memodifikasi TF-IDF sebagai feature weighting pada Multinomial Naïve Bayes, yang pada selanjutnya akan disebut dengan algoritma TF-iGini, pada kasus Sentiment Analysis berbahasa Indonesia.

2. DASAR TEORI

2.1 Data Mining

Data mining merupakan sebuah metode yang menjadi inti dari proses Knowledge Discovery. Metode ini mempunyai fokus pada pencarian atau pengekstraksian pola-pola menarik dari sekumpulan data.



Gambar 1. Diagram Proses dari Knowledge Discovery [6]

Data Mining muncul dilatarbelakangi dengan diperlukannya proses penemuan pola unik secara cepat dan tepat pada data-data digital yang dimiliki. Pola yang dihasilkan nantinya akan dirangkum menjadi informasi-informasi yang kemudian dianalisa untuk keperluan lebih lanjut. Data mining sendiri merupakan perpotongan dari berbagai macam ilmu terapan lain yang diantaranya adalah artificial intelligence, machine learning, statistik, dan sistem basis data [7].

2.2 Data Preprocessing

Data yang digunakan dalam proses data mining tidak selamanya dalam kondisi ideal untuk diproses. Terkadang pada data tersebut terdapat berbagai macam permasalahan yang dapat mengganggu hasil daripada data mining itu sendiri seperti diantaranya missing value, data redundant, outliers, ataupun format data yang tidak sesuai dengan system. Untuk mengatasi hal-hal tersebut maka diterapkanlah tahap Data Preprocessing [7].

Data preprocessing merupakan salah satu tahapan di dalam Knowledge Discovery. Proses data preprocessing bertujuan untuk melakukan treatment awal terhadap data demi menghilangkan permasalahan-permasalahan yang dapat mengganggu hasil daripada proses data minig. Dalam kasus sentiment analysis yang menggunakan data bertipe teks.

2.3 Text Mining

Text mining atau yang juga biasa *text data mining* merupakan sebuah metode turunan dari *Data Mining* yang bertujuan untuk mencari pola atau informasi menarik dari sekumpulan data yang

berbentuk *natural language text*. Jika dibandigkan dengan data yang tersimpan di dalam database, data berbentuk teks memiliki karakteristik yang tidak teratur, tidak terstruktur, dan sulit untuk diolah dengan cara biasa, padahal dalam perkembangan dunia modern, teks menjadi salah satu bentuk vital dalam pertukaran data dan informasi antar user.

Seperti halnya *data mining* yang bertujuan untuk mencari pola unik di dalam data, text mining juga memiliki tujuan untuk mencari pola yang unik dalam teks. Perbedaannya adalah, *data mining* dapat diartikan mencari informasi yang implisit atau tersembunyi, sebelumnya tidak diketahui, dan berpotensi guna dalam sebuah data [9]. Sedangkan *text mining* mencari informasi yang sebenarnya sudah secara eksplisit dapat terlihat dan dimengerti oleh manusia, namun tantangannya adalah bagaimana caranya informasi ini mampu direpresentasikan dengan baik tanpa menghilangkan data penting sehingga mampu diolah oleh komputer menggunakan algoritma yang ada tanpa melalui perantara manusia [10]. Dengan potensi yang sangat besar tersebut, *text mining* dapat diaplikasikan ke dalam berbagai macam kondisi dan kasus di dunia nyata, salah satunya adalah *sentiment analysis*.

2.4 Feature Weighting

Feature weighting merupakan sebuah proses memberikan nilai pada setiap feature berdasarkan relevansi dan pengaruhnya terhadap hasil kategorisasi. Nilai tersebut nantinya dapat digunakan sebagai dasar untuk melakukan seleksi *feature* berdasarkan minimum bobot yang telah dihitung dari setiap *feature*. [12]Tujuan akhir dari metode ini adalah untuk meningkatkan performansi daripada algoritma utama yang digunakan dalam proses *Data Mining.* Metode yang digunakan dalam tugas akhir ini adalah IDF dan iGini

a. Algoritma TF-IDF pertama kali dicetuskan oleh Salton dan Buckley pada tahun 1988 dan digunakan untuk kepentingan *information retrieval*, yang kemudian turut dimanfaatkan sebagai salah satu algoritma dalam metode *feature weighting* dalam *text mining.* TF-IDF memiliki formula sebagai berikut:

$$W_{ij} = tf_{ij} \times idf_i \quad (2.1)$$

Formula tersebut dapat dijabarkan menjadi *term frequency* dari *feature* i pada dokumen j dikalikan dengan IDF dari *feature i* pada dokumen j, dimana IDF sendiri merupakan kepanjangan dari Inverse Document Frequency. IDF sendiri dapat dihitung dengan cara:

$$idf_i = \log \left(\frac{\text{jumlah dokumen}}{\text{jumlah dokumen yang mengandung } term_i} \right)$$

Sebagai contoh, terdapat dokumen yang berisi "Saya suka sepak bola karena bola itu bundar". Cara menghitung TF-IDF dari kata "bola" dalam kalimat itu jika total dokumen ada 100 dan jumlah dokumen yang mengandung kata bola ada 25 adalah:

$$W_{ij} = 2 \times \log \left(\frac{100}{25} \right) = 1{,}2$$

Bisa terlihat dari formula tersebut bahwa semakin sering sebuah *feature* muncul dalam sebuah teks, maka semakin besar pula *weight* yang akan didapat, yang artinya maka akan semakin penting pula *features* tersebut. Metode ini dianggap efektif untuk *information retrieval* tetapi tidak untuk kategorisasi teks. Pada kategorisasi teks, sebuah *feature* dengan *document frequency* yang tinggi dianggap lebih penting dibandingkan dengan yang lebih rendah, dimana hal ini berbandung terbalik dengan prinsip yang ada pada *information retrieval*. Ditambah lagi, TF-IDF hanya merepresentasikan kemampuan dari *feature* untuk membedakan sebuah teks, bukan untuk membedakan antara sebuah kelas dengan kelas yang lainnya. Sehingga dapat disimpulkan, IDF dianggap kurang cocok untuk karegorisasi teks [13].

b. Gini index merupakan metode pemisahan atau split pada *decision tree*. Metode ini sesuai untuk pengurutan, sistem biner, nilai kontinu, dan lain sebagainya. Breiman, Freidman, dan Olshen memperkenalkan metode ini pada tahun 1984 dan banyak digunakan pada algoritma CART, SLIQ, SPRINT, dan Intelligent Miner [7]. Ide utama dari Gini Index untuk menghitung *impurity* dari sebuah *branch* untuk melakukan splitsebagai berikut:

$$Gini(S) = 1 - \sum_{i=1} p_i^2 \quad (2.3)$$

Dimana P_I merupakan kemungkinan sebuah sample tergolong dalam C_i dan mengestimasi menggunakan S_i/S. Nilai minimum dari Gini(S) adalah 0, hal itu terjadi jika semua anggota digolongkan kepada kelas yang sama. Gini Index digunakan untuk menghitung impurity dari atribut terhadap kategori. Jika semakin kecil nilai *impurity*, maka semakin baik pula atribut tersebut. Ketika semua sample terdistribusikan secara merata ke semua kelas, maka nilai Gini(S) adalah maksimum. Jika dilakukan perubahan dalam tujuan perhitungan menjadi menghitung *purity* pada sebuah cabang, maka rumus tersebut dapat dirubah menjadi:

$$Gini(S) = \sum_{i=1} p_i^2 \quad (2.4)$$

T. Dong menawarkan perubahan terhadap formula Gini Index untuk mengukur kadar *purity* yang merupakan karakteristik dari

kategorisasi, dimana semakin besar *purity* maka semakin baik. Formula penghitungan *purity* dengan menggunakan Gini Index menjadi seperti berikut [13]:

$$\text{[illegible]} = \sum_{\text{[illegible]}=1}^{|\text{[illegible]}|} \sqrt{\text{[illegible]}}$$

Selanjutnya algoritma *improved gini* akan digunakan untuk mensubtitusi algoritma IDF dalam *feature weighting*, sehingga bentuk *feature weighting* dalam *improved multinomial naïve bayes* dapat dirubah menjadi:

$$W_{\text{[illegible]}} = \text{[illegible]} \times \text{[illegible]}$$

2.5 Multinomial Naïve Bayes

Algoritma Multinomial Naïve Bayes merupakan algoritma yang dikembangkan dari teorema *naïve bayes classifier*. Algoritma ini menggunakan distribusi multinomial pada fungsi *conditional probabilities*. Walaupun menggunakan distribusi multinomial, algoritma ini bisa untuk diterapkan pada kasus *text mining* dengan cara merubah data teks menjadi bentuk nominal yang bisa dihitung dengan nilai integer.

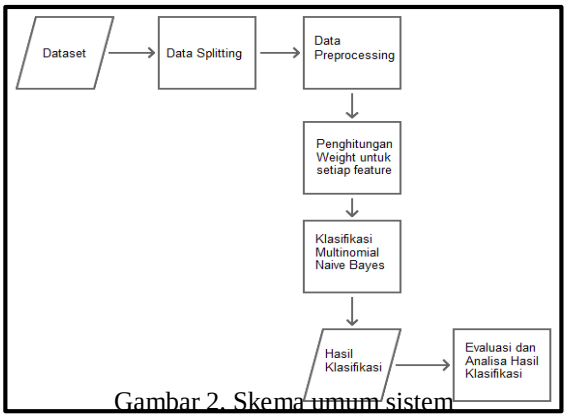
Secara umum, algoritma Multinomial Naïve Bayes untuk kasus *text classification* dimana dokumen d dapat dikategorikan pada kelas c dapat dihitung dengan:

$$\text{[illegible]} \propto \text{[illegible]} \prod_{\text{[illegible]}=1}^{\text{[illegible]}} \text{[illegible]}$$

Dimana [illegible] merupakan conditional probability dari *feature* [illegible] muncul dalam sebuah dokumen d di kelas c. Persamaan ini menghitung seberapa besar kontribusi dari setiap [illegible] pada dokumen d terhadap kelas c.

2. PERANCANGAN SISTEM

Tujuan dari tugas akhir ini adalah membuat sistem yang mampu melakukann proses *sentiment analysis* terhadap data teks menggunakan menggunakan *Improved Multinomial Naïve Bayes*. *Output* dari sistem menerapkan hasil *training* terhadap *data testing* dengan nilai output berupa negatif atau positif berdasarkan tipe *feature weighting* masing-masing.



Gambar 2. Skema umum sistem

4. PENGUJIAN

4.1 Dataset yang digunakan

Dalam tugas akhir ini digunakan dua tipe dataset, yaitu dataset dengan teks dengan jumlah karakter banyak dan teks dengan jumlah karakter sedikit. Untuk dataset teks dengan tipe karakter banyak menggunakan dataset teks beropini dari *website* "Pertamina Pasti Pas" [11], sedangkan pada *dataset* teks dengan tipe karakter sedikit menggunakan *dataset* teks beropini dari *media social* Twitter dengan tema "Provider Telekomunikasi" [16]. Tujuan penggunaan dua tipe dataset ini adalah untuk meneliti efek dari tipe panjang dataset terhadap sistem.

4.2 Pengujian Sistem

Pengujian yang dilakukan pada sistem ini pada umumnya adalah untuk mengetahui pengaruh dari iGini sebagai *feature weighting* menggunakan *Multinomial Naïve Bayes* terhadap klasifikasi pada kasus *sentiment analysis*. Rincian pengujian dan analisis yang akan dilakukan adalah sebagai berikut:

1. Menganalisis pengaruh algoritma iGini terhadap performansi klasifikasi bila dibandingkan dengan IDF yang telah lebih umum digunakan.
2. Mengetahui pengaruh rasio training dan testing pada kasus *sentiment analysis* yang menggunakan algoritma iGini.
3. Melihat pengaruh dari panjang karakter terhadap performansi dalam kasus *sentiment analysis.*

Skenario pengujian difokuskan kepada variasi daripada rasio data training dan data testing yang berkisar antara 70:20, 80:20, dan 90:10. Ketiga rasio tersebut akan diberlakukan kepada kedua dataset dengan metode *random stratified sampling*.

5. KESIMPULAN

1. Dari segi performansi secara keseluruhan, algoritma IDF untuk *feature weighting* masih lebih baik jika dibandingkan dengan iGini dengan rata-rata selisih preformansi mencapai 5%.

2. Algoritma iGini sebagai *feature weighting* justru lebih baik jika digunakan dalam kasus *information retrieval*, dimana hal ini dibuktikan dengan performansi pada parameter *recall* yang pada mayoritas skenario mampu mengungguli IDF hingga mencapai 10%.
3. Dataset "Pertamina Pasti Pas" memberikan hasil akurasi yang lebih rendah disebabkan dengan banyaknya kemiripan *feature* dari satu kelas ke kelas lainnya. Sedangkan dataset "Provider Telekomunikasi" mampu memberikan hasil yang lebih baik karena *feature* yang lebih homogen antar kelasnya seperti kata-kata umpatan.
4. Dalam pengujian ini, dataset dengan karakteristik rata-rata karakter pendek ("Provider Telekomunikasi") memberikan hasil yang lebih baik dikarenakan dengan jumlah kata yang sedikit maka kemungkinan jumlah kata yang saling ketergantungan akan semakin rendah.

6. DAFTAR PUSTAKA

[1] W. Steve, Y. Peter and W. Dawn, "The good, the bad and the wiki: Evaluating," *British Journal of Educational Technology*, p. 987–995, 2008.

[2] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trend's in Information Retreival*, pp. 1-135, 2008.

[3] H. Kanayama, T. Nasukawa and H. Watanabe, "Deeper Sentiment Analysis Using Machine Translation Technology," *Tokyo Research Laboratory.*

[4] A. M. Kiribiya, E. Frank, B. Pfahringer and G. Holmes, "Multinomial Naive Bayes for Text Classification Revisited".

[5] A.-H. Tan, "Text Mining: The State of The Art and Challenges," in *PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999.

[6] U. Fayyad, G. P Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases".

[7] J. Han and M. Kamber, "Data Mining Concept and Technique".

[8] R. K. Rahardi, Dimensi-Dimensi Kebahasaan: Aneka Masalah Bahasa Indonesia Terkini, 2006.

[9] I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 2011.

[10] I. H. Witten, "Text Mining," Computer Science, University of Waikato, Hamilton, New Zealand.

[11] N. A. Vidya, "Opinion Mining dengan Menggunakan Multinomial Naive Bayes Classifier pada Blog".

[12] M. Keyvanpou and R. Tavoli, "Feature Weighting for Improving Document Image Retrieval System Performance".

[13] D. Tao and S. Wenqian, "An Improved Algorithm of Bayesian Text," *JOURNAL OF SOFTWARE*, 2011.

[14] M. J. Zaki and M. J. W, "Data Mining and Analysis".

[15] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," School of Informatics and Engineering Flinders University, 2007.

[16] S. M. Istiqomah, "Opinion Mining pada Twitter Menggunakan Klasifikasi Sentimen pada Hashtag Berbasis Graf," 2014.

[17] S. Abright, W. Winston and C. Zappe, Data Analysis and Decision Making with Microsoft Excel, Revised, 2009.

[18]Unknown, Telaah bahasa dan sastra : persembahan kepada Prof. Dr. Anton M. Moeliono, Pusat Pembinaan dan Pengembangan Bahasa, Departemen Pendidikan dan Kebudayaan, 1999.