KLASIFIKASI TOPIK BERITA MENGGUNAKAN *MUTUAL INFORMATION* DAN *BAYESIAN NETWORK*

Fahmi Salman Nurfikri[1], Mohamad Syahrul Mubarok[2], dan Adiwijaya[3]
Fakultas Informatika, Telkom University
Email: [1]fahmisalman@student.telkomuniversity.ac.id, [2]msyahrulmubarok@gmail.com,
[3]adiwijaya@telkomuniversity.ac.id

Abstrak

Seiring dengan meningkatnya perkembangan internet, maka pertumbuhan informasi tekstual di internet terus mengalami peningkatan. Dengan peningkatan informasi tersebut, maka kebutuhan pengklasifikasian berita secara otomatis sangat dibutuhkan untuk menemukan informasi atau berita yang diinginkan. Salah satu cara untuk mengelopokan suatu berita ke dalam kategori tertentu berdasarkan informasi yang terdapat dalam berita tersebut adalah *text classification*. Salah satu metode dalam *text classification* adalah *Bayesian Network*. *Bayesian Network* merupakan salah satu metode reasoning yang memodelkan hubungan antar variabel dalam *Probabilistic Graphical Model* (PGM). Keuntungan *Bayesian Network* dibandingkan dengan metode yang lain yaitu, cocok untuk dataset yang kecil dan tidak lengkap, dapat menangani ketidakpastian dan pengambilan keputusan, dan komputasi yang cepat. Selain itu, dilakukan seleksi fitur dengan menggunakan metode *Mutual Information* untuk mengurangi jumlah dimensi dan untuk meningkatkan performa klasifikasi. Hasil dari klasifikasi ini dinyatakan dalam *F1-measure micro-average* dengan nilai performansi sebesar 75,34%.

Kata kunci :

Text classification, Bayesian Network, Mutual Information, F1-measure

Abstract

Along with the increasing of the internet development, the growth of textual information on internet continues to experience enhancement. With the increase of the information, then the need for automatical classification of news is needed to find the desired information or news. One of ways to disguise a story into a particular category based on the information contained in the news is text classification. One of methods in text classification is Bayesian Network. Bayesian Network is one of reasoning methods which modeled the relationship between variables in the Probabilistic Graphical Model (PGM). Bayesian Network advantages compared by another methods are suitable for small and incomplete datasets, can handle uncertainty and retrieval decisions, and rapid computing. In addition, feature selection is performed using the Mutual Information method for reduce the number of dimensions and to improve the performance of classification. The results of this classification is expressed in F1-measure micro-average with a performance value of 75.34%.

Keywords :

Text classification, Bayesian Network, Mutual Information, F1-measure

I. PENDAHULUAN

Berita suatu informasi mengenai kejadian atau peristiwa yang hangat[1]. Berita tidak hanya menyebar melalui televisi atau surat kabar saja, berita juga menyebar melalui Internet.

1.kbbi.web.id
2.www.ida.or.id
3.www.pcplus.co.id

Seiring dengan meningkatnya perkembangan Internet, maka pertumbuhan informasi tekstual di Internet terus mengalami peningkatan. Menurut data yang diperoleh dari *Indonesian Digital Association*[2] (IDA) bahwa 96% masyarakat Indonesia mengkonsumsi berita melalui online dan dari survey yang dilakukan oleh UC Browser[3] pada tahun 2016 menyatakan bahwa 56,5% pengguna internet di Indonesia rata-rata membaca 4-12 artikel berita per hari. Dari data yang diperoleh, maka kebutuhan pengklasifikasian berita secara otomatis sangat dibutuhkan untuk menemukan informasi atau berita yang diinginkan. Salah satu cara untuk mengelompokan suatu berita ke dalam kategori tertentu berdasarkan informasi yang terdapat dalam berita tersebut adalah *text classification* [8], [14], [16].

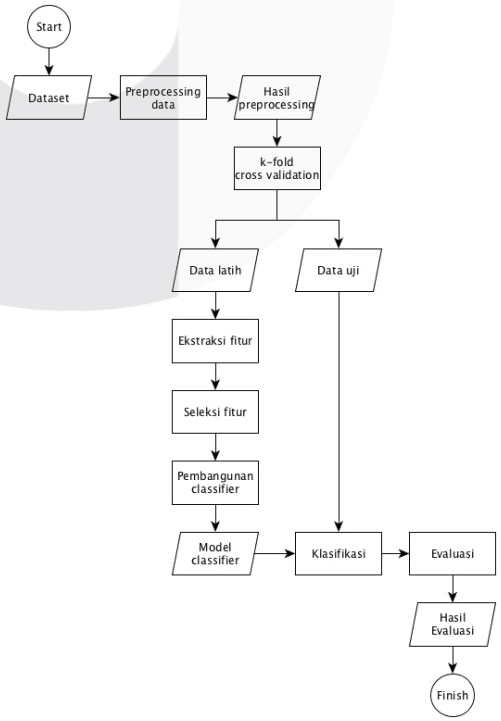
Banyak cara pendekatan dalam *text classification*, antara lain pendekatan probabilistik, *Support Vector Machine*, dan *Artificial Neural Network* [7]. Salah satu metode dalam text classification dengan pendekatan probabilistik adalah *Bayesian Network* [7], [5], [9], [17]. *Bayesian Network* merupakan salah satu metode *reasoning* yang memodelkan hubungan antar variabel dalam *Probabilistic Graphical Model* (PGM). *Bayesian Network* menggunakan *Directed Acyclic Graph* (DAG) untuk merepresentasikan sekumpulan variabel dan conditional dependency antar variable tersebut. DAG tersebut digunakan untuk dijadikan sebuah model dalam *Bayesian Network*. *Bayesian Network* telah banyak diimplementasi sebagai classifier pada banyak penelitian [2], [6], [11], [15]. Keuntungan *Bayesian Network* dibandingkan dengan metode yang lain yaitu, cocok untuk dataset yang kecil dan tidak lengkap, dapat menangani ketidakpastian dan pengambilan keputusan, dan komputasi yang cepat.

Permasalahan lain dalam *text classification* adalah jumlah dimensi yang cukup besar, yang menyebabkan kompleksitas komputasi menjadi tinggi dan akurasi yang rendah [12]. Untuk mengatasi hal tersebut diusulkan teknik seleksi fitur untuk mengurangi tingginya dimensi data.

Salah satu metode seleksi fitur adalah dengan menggunakan *Mutual Information* (MI) [12]. MI merupakan metode seleksi fitur yang cukup efisien untuk memilih fitur dari suatu dokumen [20], [21]. MI mengukur berapabanyak informasi atau atribut tersebut berperan untuk membuat klasifikasi benar didalam class manapun sehingga akan menghasilkan inputan yang lebih berpengaruh kepada proses klasifikasi [21]. Oleh karena itu, dalam penelitian ini penulis menggunakan metode MI untuk mengurangi dimensi data agar dapat meningkatkan keefektifan dan meningkatkan performansi *Bayesian Network Classifier* dalam mengklasifikasikan artikel berita bahasa Indonesia.

II. KAJIAN LITERATUR

Sistem yang diajukan untuk dalam penelitian ini adalah sistem yang mampu menglasifikasikan topik berita secara otomatis. Proses-proses tersebut digambarkan menggunakan flowchart pada gambar 1.



Gambar 1. Gambaran Umum Sistem

Berdasarkan gambar 1, alur sistem secara umum pada penelitian ini dapat diuraikan sebagai berikut: 2. Gambaran Umum Sistem. Berikut rincian penjelasan dari setiap tahapan yang telah disebutkan pada gambaran umum sistem dari gambar 1:

1. Dataset

Tahap awal yang dilakukan adalah pengumpulan dataset berbasis teks yang diperoleh dari situs penyedia berita berbahasa Indonesia yang terdapat di internet. Data tersebut diambil secara manual dari berbagai macam situs seperti detik.com, kompas.com, tribunnews.com, beranews.com, okezone.com, sindonews.com, nationalgeographic.co.id, liputan6.com, tempo.co, aktual.com dan republika.co.id dengan rentang berita dari mulai bulan Agustus 2016 sampai dengan Februari 2017.

Pada penelitian ini terdapat 12 kelas, yaitu Ekonomi, GayaHidup, Kesehatan, Hiburan, Hukum, Olahraga, Otomotif, Pendidikan, Politik, Properti, Teknologi dan Wisata. Dalam satu kelas terdapat 30 data teks berita sehingga total keseluruhan dataset berjumlah 360 dataset.

Table 1. Komposisi Data

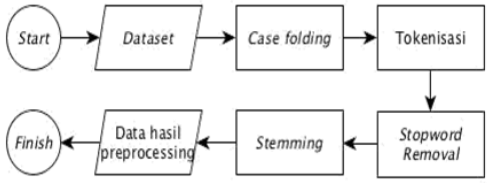
No	Nama Kelas	Jumlah Dokumen
1	Ekonomi	30
2	GayaHidup	30
3	Kesehatan	30
4	Hiburan	30
5	Hukum	30
6	Olahraga	30
7	Otomotif	30
8	Pendidikan	30
9	Politik	30
10	Budaya	30
11	Teknologi	30
12	Wisata	30

Sebanyak 360 dataset tersebut akan dipisah menjadi data latih dan data uji dengan menggunakan *k-fold cross validation* dengan tujuan untuk melakukan persebaran data dan membagi data (*training* dan *testing*) sebanyak k *segment/fold*.

2. *Preprocessing*

Preprocesing merupakan proses untuk mempersi-apkan data yang akan diklasifikasikan. *Preprocessing* merupakan tahapan untuk menghilangkan noise yang terdapat dalam data teks sehingga dapat menghapus data yang kurang relevan dalam pengklasifikasian. Selain itu, tujuan dari *preprocessing* adalah untuk meningkatkan performansi sistem

yang dibangun dalam penelitian ini. Tahap-tahap *preprocessing* pada penelitian ini tergambar dalam flowchart pada gambar 2.



Gambar 2. Preprcessing

Pada subbab berikut akan dijelaskan lebih detail mengenai preprocessing data teks pada penelitian ini.

- *Case Folding*
  Merupakan tahapan untuk mengubah semua huruf menjadi *lower-case.* Tujuan dari *case folding* adalah untuk menanggulangi ketidakkonsistennan peng-gunaan huruf kapital dan huruf kecil (*case sensitive*) dalam sebuah artikel berita. Selain itu, pada tahap ini juga dilakukan penghapu-san karakter selain huruf seperti tanda baca (seperti (,), (.), dll) dan angka (seperti 4, 64, 35, dll).
- Tokenisasi
  Setelah dilakukan *case folding* lalu setelah itu dilakukan proses tokenisasi atau *lexing*. Tokenisasi merupakan proses untuk merubah suatu kalimat menjadi potongan-potongan kata. Tokenisasi dilakukan untuk membuat data lebih terstruktur dan juga memudahkan sistem dalam pengolahan kata.
- *Stopword Removal*
  Stopword removal merupakan suatu proses untuk menghapus kata-kata yang terdapat dalam stopword list dari token list seperti kami, saya, dan, dari, dll. Tujuan dari peng-gunaan *Stopword removal* ini adalah untuk mengurangi jumlah kata yang akan diproses. Pada penelitian ini, penulis menggunakan daftar *stopword* untuk Bahasa Indonesia hasil penelitian yang dilakukan oleh F. Z. Tala [15].
- *Stemming*
  Stemming merupakan proses untuk mengembalikan kata-kata yang ada di token list menjadi bentuk awal atau kata dasar dari kata tersebut. *Stemming* menghilangkan imbuhan awalan, sisipan, akhiran ataupun kom-binasi awalan dan akhiran. Pada penelitian ini, penulis menggunakan algoritma *Stemming* Nazief-Andriani karena mempunyai performa yang lebih baik jika dibandingkan dengan algoritma *stemming* lainnya [1], [14].

3. K-Flod Cross Validation

Pada tahap ini dilakukan pembagian data latih dan data uji dengan menggunakan 5-*fold cross validation*. Tujuan dari *cross validation* ini adalah untuk melakukan persebaran data dan membagi data (latih dan uji) sebanyak 5 *segment*/*fold*.

4. Ekstraksi Fitur

Pada tahap ini, kata-kata yang telah dihasilhan dari hasil *preprocessing* data teks dilakukan proses *features extraction* dengan representasi model *bag of words*. Representasi dengan model *bag of words* dilakukan dengan cara memeriksa ada atau tidaknya masing-masing kata pada data latih untuk masing-masing dokumen. Pada penelitian ini, pendekatan yang dipakai adalah pendekatan binomial.

5. Seleksi Fitur

Seleksi fitur adalah proses pemilihan subset dari fitur yang relevan untuk digunakan dalam pembangunan model klasifikasi. Seleksi fitur digunakan untuk melakukan seleksi atribut yang akan dimasukan untuk proses klasifikasi agar

lebih informatif dan efektif [11]. *Mutual Information* (MI) adalah salah satu cara untuk melakukan seleksi fitur yang digunakan pada penelitian ini [13]. Perhitungan MI dapat dilihat dalam persamaan 1.

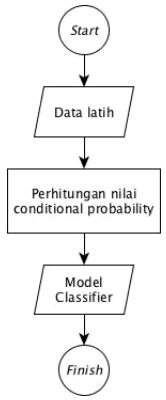
$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right), \qquad (1)$$

Keluaran dari MI merupakan suatu matrix yang berukuran n x n dimana n adalah banyaknya atau jumlah atribut dimana nilai dalam matriks tersebut merupakan nilai keterkaitan antar dua atribut [2][3]. Matrix tersebut berisi nilai keterkaitan antara satu atribut dengan atribut lainnya, semakin besar nilai MI, maka semakin besar keterkaitan antar atribut tersebut. Pada penelitian ini, penulis akan mencari nilai MI untuk semua atribut terhadap kelas.

Setelah didapatkan nilai MI dari persamaan 1, penulis melakukan pengurutan nilai MI dari besar ke kecil, dimana semakin besar nilai MI, maka semakin besar suatu atribut tersebut mempengaruhi suatu kelas [12]. Dalam penelitian ini, penulis mengambil informasi nilai MI hanya untuk atribut kelas saja, lalu mengurutkan nilainya dari nilai MI terbesar sampai nilai MI terkecil.

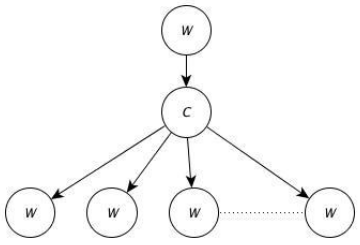
6. Pembangunan *Classifier*

Pada tahap ini, data latih hasil preprocessing akan digunakan dalam pembangunan model *classifier*. Dimana pembangunan model classifier ini akan dilakukan dengan menggunakan metode *Bayesian Network Classifier*. Alur dari pembangunan classifier ini digambarkan oleh flowchart pada gambar 3.



Gambar 3. Pembangunan Classifier

Model *Directed Asyclic Graph* (DAG) *Bayesian Network* yang dibangun berdasarkan hubungan *causal relationship* atau hubungan sebab akibat [10]. DAG *Bayesian Network* yang digunakan pada penelitian ini seperti gambar 4 [11].



Gambar 4. DAG Bayesian Network

Pada gambar 4, C merupakan node untuk variable kelas dimana node tersebut adalah keluaran yang akan dituju dalam proses klasifikasi. Node W_1 sampai dengan W_n merupakan kata dalam dokumen yang akan menjadi masukan dalam

proses klasifikasi dan n menyatakan jumlah kata. W_1 adalah parent dari node C dan W_2 sampai dengan W_n adalah child dari node kelas. Dimana kata yang terdapat pada node W_1 adalah kata yang mempunyai nilai MI terbesar.
Pembangunan classifier dilakukan dengan menghitung nilai *conditional probability* pada data latih hasil seleksi fitur yang dihasilkan dari mutual information dengan menggunakan parameter *Maximum a Posteriori* (MAP). Keuntungan dari penggunaan parameter MAP dapat menghindari *zero probability* dimana *zero probability* merupakan suatu kejadian dimana probabilitasnya sama dengan nol. Perhitungan *conditional probability* dihitung dengan persamaan 2,

$$\theta_{ijk} = \frac{\alpha_{ijk} + n_{ijk}}{\sum_k (\alpha_{ijk} + n_{ijk})} \quad (2)$$

Dimana θ_{ijk} merupakan parameter ketika sebuah variabel X_i bernilai k dan PA_i dari X_i bernilai j, n_{ijk}adalah berapa kali variabel X_i bernilai k dan PA_i adalah j dalam sebuah dokumen D dan α_{ijk} dinyatakan dalam persamaan 3,

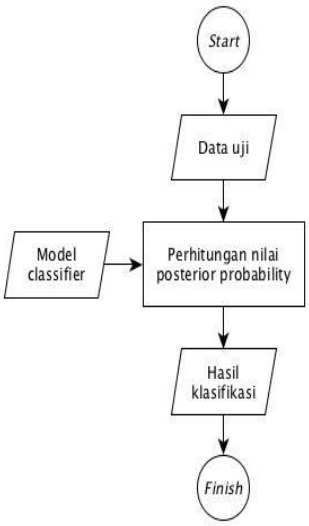
$$\alpha_{ijk} = \frac{\propto}{r_i . q_i} \quad (3)$$

Dimana α adalah nilai kecil untuk menghindari probabilitas nol, r_i adalah penjumlahan nilai variabel X_i dan q_i adalah jumlah instantiasi PA_i.

Hasil dari perhitungan persamaan 3 dinyatakan dalam *Conditional Probability Table* (CPT). CPT akan digunakan dalam proses klasifikasi menggunakan *Bayesian Network Classifier*.

7. Bayesian Network Classifier

Tahap klasifikasi ini dilakukan untuk melakukan pengujian terhadap classifier yang sudah dibangun. Tahapan proses klasifikasi dapat dilihat pada gambar 5.



Gambar 5. Proses Klasifikasi

Berdasarkan Gambar 5, proses klasifikasi dilakukan pada data uji dengan memperhitungkan nilai posterior probability untuk semua kelas. Perhitungan posterior probability dilakukan dengan menggunakan nilai conditional probability dari model classifier yang didapatkan pada tahap pembangunan klasifikasi yang direpresentasikan dalam bentuk CPT. Setelah mendapatkan nilai posterior probability untuk masing-masing kelas, kemudian untuk menentukan kelasnya ditentukan berdasarkan kelas yang memiliki nilai posterior probability tertinggi atau *Maximum a Posteriori* (MAP).

8. Evaluasi

Tahap akhir dari penelitian ini adalah dengan melakukan evaluasi sistem. Evaluasi sistem dilakukan untuk mengetahui performa dari sistem yang telah dibangun pada penelitian ini. Hasil evaluasi dari classifier dinyatakan dalam akurasi, micro-average precision, micro-average recall dan F1-micro average untuk mengukur performansi dari classifier seperti dalam persamaan 4, 5, dan 6.

$$Micro - average\ of\ precision = \frac{\sum_i^n TP_i}{\sum_i^n (TP_i + FP_i)} \quad (4)$$

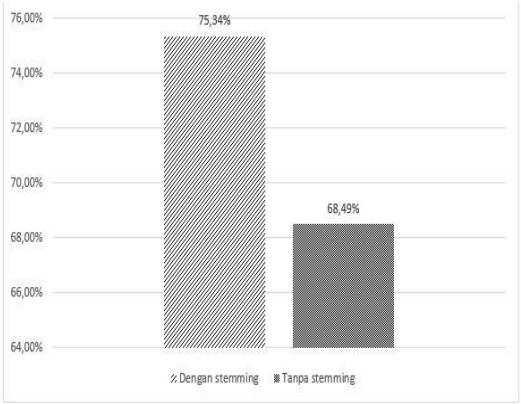
$$Micro - average\ of\ recall = \frac{\sum_i^n TP_i}{\sum_i^n (TP_i + FN_i)} \quad (5)$$

$$F1 - measure = \frac{2 * precision * recall}{precision + recall} \quad (6)$$

III. ANALISIS DAN PERANCANGAN

Untuk mengetahui performa sistem yang dibangun maka dilakukan pengujian menggunakan 5-fold cross validation. dimana 4 fold data, atau setara dengan 286 data digunakan sebagai data latih dan 1 fold data, atau setara dengan 74 data digunakan sebagai data tes atau validasi. Adapun analisis hasil pengujian yang dilakukan:

1) Pengaruh penggunaan Steeming terhadap *Bayesian Network Classifier*
Gambar 6 merupakan hasil perbandingan nilai *F1-Measure Micro Average* terhadap scenario pengujian pertama.

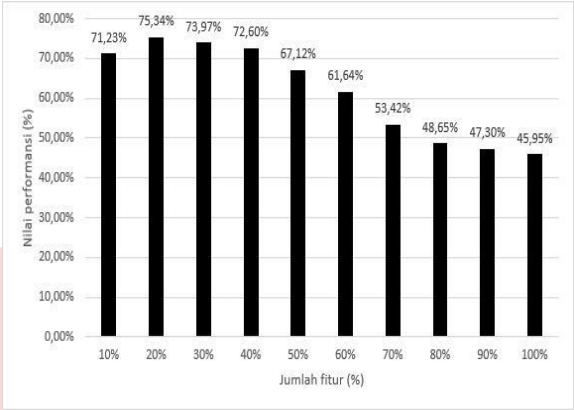


Gambar 6. Perbandingan nilai F1-Measure Micro Average pada Preprocessing dengan Stemming dan tanpa Stemming

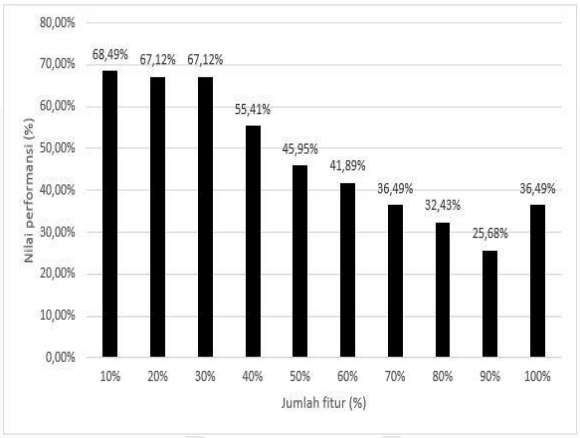
Berdasarkan gambar 6 dapat terlihat bahwa skor performansi dengan menggunakan steming lebih besar daripada tanpa menggunakan stemming pada preprocessing, dimana nilai f1-measure micro average dengan menggunakan stemming adalah sebesar 75,34% sedangkan tanpa menggunakan stemming sebesar 68,49%. Dapat disimpulkan bahwa preprocessing dengan menggunakan stemming dapat meningkatkan performansi klasifikasi sebesar 7%. Hal tersebut dikarenakan pada *preprocessing* dengan menggunakan stemming dapat mengurangi variasi fitur-fitur sehingga mengurangi kemungkinan perbedaan kata yang mempunyai makna yang sama. Hal inilah yang membuat performa sistem yang dibangun dengan proses stemming lebih besar.

2) Pengaruh penggunaan Mutual Information terhadap *Bayesian Network Classifier*
Pada tahap ini dibahas mengenai pengaruh pengunaan mutual information terhadap performansi *dari Bayesian Network Classifier*. Gambar 7 merupakan grafik perbandingan penggunaan *mutual information* dengan menggunakan stemming, gambar 8 merupakan grafik perbandingan penggunaan *mutual information* tanpa menggunakan stemming

dan gambar 9 merupakan grafik perbandingan *mutual information* dengan menggunakan *stemming* dan tanpa menggunakan stemming.

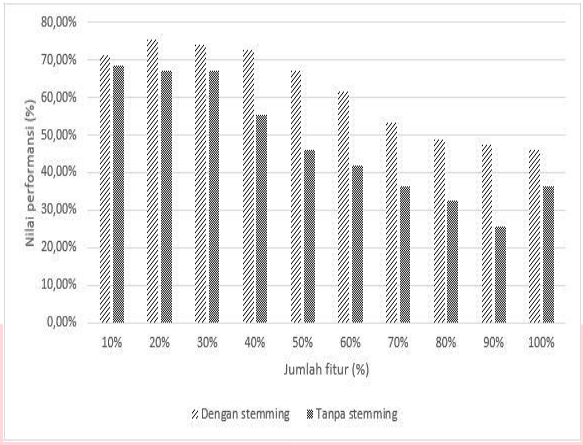


Gambar 7. Perbandingan nilai mutual information yang diambil dengan proses stemming



Gambar 8. Perbandingan nilai mutual information yang diambil tanpa proses stemming

Berdasarkan gambar 7 dan 8 dapat terlihat bahwa semakin besar fitur yang diambil dari nilai mutual information maka terjadi kecenderungan performansi akan menurun, dapat disimpulkan bahwa semakin banyak fitur yang diambil, akan semakin kecil nilai performansi klasifikasi *Bayesian Network Classifier*. Hal ini dikarenakan semakin banyaknya fitur yang dipakai dalam proses klasifikasi, semakin banyak fitur yang mempunyai nilai mutual information rendah ikut dalam proses klasifikasi.



Gambar 9. Perbandingan nilai mutual information yang diambil dengan dan tanpa proses stemming

Pada gambar 9, dapat terlihat bahwa perbandingan nilai *mutual information* yang diambil dengan proses stemming menghasilkan rata-rata performansi yang lebih besar dibandingakan tanpa proses stemming, hal tersebut seperti yang telah dijelaskan pada pemaparan sebelumnya, bahwa stemming dapat mengurangi variasi fitur-fitur sehingga mengurangi kemungkinan perbedaan kata yang mempunyai makna yang sama.

Dari penjelasan diatas dapat disimpulkan bahwa proses stemming dapat mengurangi noise dalam proses klasifikasi sehingga dapat meningkatkan performansi dari proses klasifikasi pada Bayesian Network Classifier. Selain itu, penggunaan jumlah fitur juga berpengaruh dalam performansi dari proses klasifikasi dimana pada proses yang menggunakan stemming grafik performansi naik pada awalnya, lalu kemudian menurun, sedangkan pada proses tanpa menggunakan stemming grafik cenderung menurun, namun pada akhirnya naik.

Pada kondisi pertama pada proses yang hanya menggunakan 10% fitur performansinya lebih kecil dibandingkan dengan proses yang menggunakan 20% fitur hal ini disebabkan pada proses yang menggunakan 10% fitur, kata yang digunakan sangat sedikit sehingga lebih sedikit memberikan variasi fitur atau kata untuk masuk dalam model klasifikasi. Lalu grafik cenderung menurun setelah proses yang menggunakan 20% fitur. Hal ini dikarenakan fitur-fitur yang mempunyai nilai MI kecil merupakan fitur-fitur yang tidak diinginkan namun masuk ke dalam model klasifikasi sehingga mengurangi nilai performansi dari proses klasifikasi.

Pada kondisi kedua, dari awal grafik cenderung menurun dari penggunaan fitur 10% sampai dengan 90% kemudian naik lagi pada penggunaan fitur 100%. Hal ini dikarenakan karena ada beberapa fitur yang dianggap tidak diinginkan pada proses MI namun sebenarnya dibutuhkan pada proses klasifikasi.

IV. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang telah dilakukan maka didapatkan kesimpulan bahwa sistem yang dibangun dengan proses stemming menghasilkan performa f1-measure micro average yang lebih baik dibandingkan tanpa menggunakan stemming yaitu sebesar 75,34%. Selain itu, *Mutual infotmation* mampu meningkatkan performansi dari *Bayesian Network Classifier*, karena semakin kecil persentase fitur yang diambil pada proses klasifikasi, semakin besar nilai performansi dari *Bayesian Network Classifier*.

Saran untuk penelitian selanjutnya adalah dengan menambah jumlah dataset yang digunakan, karena dengan penambahan dataset dapat meningkatkan keragaman informasi sehingga diharapkan dapat meningkatkan performansi sistem dan juga membuat sistem klasifikasi yang dapat menangani kasus multi label classification, karena topik pada

suatu berita dapat melibatkan lebih dari satu label kelas. Selain itu, perlu adanya optimasi pada struktur *Bayesian Network* agar dapat menghasilkan performansi yang lebih baik.

REFERENSI

[1] A. Pratama, M. Syahrul. Mubarok., and Bijaksana, A. Evaluasi eksplisit dan implisit algoritma-algoritma stemming bahasa indonesia.

[2] A Saputra, Adiwijaya, MS Mubarok. Klasifikasi sentimen pada level aspek terhadap ulasan produk berbahasa inggris menggunakan bayesian network (case study: Data ulasan produk amazon). eProceedings of Engineering 4 (2017). Adiwijaya. Matematika Diskrit dan Aplikasinya. Alfabeta: Ban-dung, 2016.

[3] Luis M. de Campos, A. E. R. Bayesian network models for hierar-chical text classification from a thesaurus. International Journal of Approximate Reasoning 50 (2009), 932944.

[4] Adiwijaya. Aplikasi matriks dan Ruang Vektor. Graha Ilmu: Yogyakarta, 2014.

[5] Adiwijaya. Matematika Diskrit dan Aplikasinya. Alfabeta: Bandung, 2016.

[6] B Julianto, A Adiwijaya, M. M. Identifikasi parafrasa bahasa indonesia menggunakan naive bayes. eProceedings of Engineering 4 (2017).

[7] D Sitompul, A Adiwijaya, M. M. Analisis sentimen level kalimat pada ulasan produk menggunakan bayesian networks. eProceedings of Engineering 4 (2017).

[8] Hamzah, A. Klasifikasi teks dengan naive bayes classifier (nbc) untuk pengelompokan teks berita dan abstract akademis.

[9] Luis M. de Campos, A. E. R. Bayesian network models for hierarchical text classification from a thesaurus. International Journal of Approximate Reasoning 50 (2009), 932944.

[10] L Putri, M Mubarok, A. A. Klasifikasi sentimen pada ulasan buku berbahasa inggris menggunakan information gain dan na ve bayes. ePro- ceedings of Engineering 4 (2017).

[11] Ivana Clairine Irsan, M. L. K. Hierarchical multilabel classification for indonesian news articles.

[12] R.A.Aziz, M.Syahrul.Mubarok.,andAdiwijaya.Klasifikasitopik pada lirik lagu dengan metode multinomial naive bayes. Indonesia Symposium on Computing (IndoSC).

[13] MSMubarok, KC Widiastuti, A.A. Implementasi mutual information dan naive bayes untuk klasifikasi data microarray. eProceedings of Engineering 4 (2017).

[14] Nanda Maulina Firdaus, M. Syahrul. Mubarok., and Adiwijaya. Klasifikasi huruf isyarat tangan menggunakan naive bayes (2017).

[15] Asriyanti Indah Pratiwi, Adiwijaya. On the feature selection and classification based on information gain for document sentiment analysis. Applied Computational Intelligence and Soft Computing (2018).

[16] MS Mubarok, M. A. Klasifikasi emosi pada twitter menggunakan bayesian network.

[17] M. Syahrul Mubarok, Adiwijaya, and Aldhi, M. Aspect-based sentiment analysis to review products using na¨ıve bayes. AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060) (2017).

[18] N Prayuga, A Adiwijaya, M. M. Klasifikasi polycystic ovary syndrome berdasarkan citra ultra sonografi menggunakan principal component analysis dan naïve bayes untuk membantu mendeteksi kesuburan wanita. eProceedings of Engineering 4 (2017).

[19] A.H.R.Z. Arifin, M. Syahrul. Mubarok., and Adiwijaya. Learning struktur bayesian networks menggunakan novel modified binary differential evolution pada klasifikasi data. Indonesia Symposium on Computing (IndoSC) (2016).

[20] M. Syahrul Mubarok, M. D. Purbolaksono, Adiwijaya. Implementasi mutual information dan bayesian network untuk klasifikasi data microarray. e-Proceeding of Engineering: Vol.4, No.2 Agustus 2017 (2017).

[21] Jie Cheng, R. G. Comparing bayesian network classifers.

[22] Marius Vila, Anton Bardera, M. F., and Sbert, M. Tsallis mutual information for document classification. Entropy 2011 (2011).

[23] Dwi Wahyudi, Teguh Susyanto, D. N. Implementasi dan analisis algoritma stemming nazeif-andriani dan porter pada dokumen berbahasa indonesia. Jurnal Ilmiah SINUS.

[24] F. Z. Tala, A study of stemming effects on information retrieval in Bahasa Indonesia, Inst. Log. Lang. Comput. Univ. Van Amst. Neth., 2003.