KLASIFIKASI KATEGORI DAN PELABELAN BERITA BAHASA INDONESIA MENGGUNAKAN MUTUAL INFORMATION DAN K-NEAREST NEIGHBORS

Alfredo Gormantara[1)], Dominikus Boli Watomakin[2)]
[1]Program Studi Teknik Informatika, Teknologi Informasi, Universitas Atma Jaya Makassar
[2]Program Studi Magister Informatika, Universitas Atma Jaya Yogyakarta
Alamat e-mail: alfredo_gormantara@lecturer.uajm.ac.id[1)], j1mmywatomakin@gmail.com[2)]

ABSTRAC

Along with the increasing development of the internet, the growth of textual information on the internet continues to increase. Increased dissemination of information causes the news released every day to increase. The large number of news, especially those in online media, raises problems in categorizing existing news topics. So we need a system that can categorize every news topic that is in online media and also labeling sentiment on a news. This aims to be able to monitor a situation such as political, social or economic from online news by utilizing classifications and also sentiment labeling so that it can predict steps that can be taken in the future. The method used in this research is feature selection and the KNN classification algorithm with Manhattan distance. In this study, there are two main functions of the system, namely news categorization and also news sentiment labeling. The research stages began with data collection, preprocessing, mutual information feature selection and classification. The results of this study show positive results because it can recognize news categorization and news sentiment labeling.

Keywords: *Mutual Information, KNN, Text Classification, Sentiment, News*

1. PENDAHULUAN

Berita merupakan salah satu sarana komunikasi bagi masyarakat. Seiring perkembangan teknologi komunikasi, pendistribusian informasi melalui sarana internet berkembang dengan sangat pesat dan terus meningkat. Berdasarkan hasil riset yang dilakukan oleh lembaga global GFK dan *Indonesia Digital Association* (IDA) yang dilaksanakan di lima kota besar di Indonesia sepanjang tahun 2015, persentase konsumsi berita melalui media *online* mencapai 96 persen [1]. Peningkatan penyebaran informasi menyebabkan berita yang dirilis menjadi sangat banyak tiap hari. Banyaknya berita menimbulkan masalah pada pengkategorian topik maupun sentimen terhadap suatu berita yang dibaca. Kebutuhan pengklasifikasian topik berita sangat dibutuhkan untuk memudahkan layanan media online dalam pengkategorian berita, sehingga masyarakatpun juga akan terbantu dalam pengaksesan berita [2]. Salah satu cara untuk mengelompokan suatu berita ke dalam kategori tertentu berdasarkan informasi yang terdapat dalam berita maupun menentukan sentimen dari suatu berita tersebut adalah *text classification.*

Proses mengidentifikasi dokumen ke dalam kelas-kelas yang sebelumnya sudah terdefinisi inilah yang disebut proses klasifikasi [3]. Dalam proses klasifikasi pemilihan metode yang digunakan menjadi hal yang berpengaruh dalam kinerja klasifikasi [4]. Salah satu masalah dalam *text classification* adalah jumlah atribut yang cukup banyak, yang menyebabkan kompleksitas komputasi menjadi tinggi dan waktu yang dibutuhkan komputasi semakin lama. Untuk mengatasi masalah tersebut digunakan teknik seleksi fitur untuk mengurangi atribut data yang banyak.

Seleksi fitur merupakan bagian penting untuk mengoptimalkan kinerja dari *classifier.* Seleksi fitur dapat didasarkan pada pengurangan ruang fitur yang besar, misalnya dengan mengeliminasi atribut yang kurang relevan [5]. Penggunaan algoritma *feature selection* yang tepat dapat meningkatkan *accuracy.* Salah satu seleksi fitur yang sering digunakan dalam *text mining* adalah *mutual information* (MI). MI merupakan metode seleksi fitur yang cukup efisien untuk

memilih fitur dari suatu dokumen. MI mengukur seberapa banyak informasi atau atribut tersebut berperan untuk membuat klasifikasi benar di dalam kelas manapun sehingga akan menghasilkan atribut yang lebih berpengaruh kepada proses klasifikasi [6]. Klasifikasi bermanfaat untuk mengelompokan data yang jumlahnya sangat banyak dan sulit dilakukan apabila diproses secara manual [7].

Tujuan dari penelitian ini adalah untuk dapat mengklasifikasikan sebuah berita menjadi kategori politik, sosial atau ekonomi dan dari itu dilakukan pelabelan sentimen untuk dapat mengetahui berita positif dan negatif. Dari kedua hal tersebut nantinya dapat dilakukan *monitoring* atau prediksi keadaan politik, sosial dan ekonomi berdasarkan berita online sehingga dapat memprediksi langkah yang dapat diambil untuk kedepannya.

Oleh karena itu, dalam penelitian ini menggunakan metode mutual information untuk mengurangi atribut data agar dapat meningkatkan keefektifan dan meningkatkan performa dalam mengklasifikasikan kategori berita dan juga pelabelan pada berita.

2. TINJAUAN PUSTAKA

Feature selection mempunyai peranan penting di data mining dan mesin belajar dimana untuk mengurangi dimensi dari data dan meningkatkan kinerja algoritma, seperti algoritma klasifikasi [8]. *Feature selection* bertujuan untuk memecahkan masalah dengan memilih hanya sebagian kecil fitur yang relevan dari kumpulan fitur utama yang asli. Dilihat dari penelitian sebelumnya penerapan *feature selection* sendiri sudah sangat banyak dilakukan. *Feature selection* pernah dilakukan untuk pengidentifikasian kekuatan tubuh bagian atas untuk para atlit ski [9], kemudian ada penelitian *feature selection* untuk klasifikasi *scene* penginderaan jarak jauh yang berbasis pembelajaran [10]. Sedangkan dalam penelitian [11], sedikit mengemukakan kelemahan yang dimiliki oleh *feature selection* antara lain kurangnya informasi tentang interaksi antara fitur dan penggolong, dan pemilihan fitur yang berlebihan serta tidak relevan. Yang kedua adalah karena keterbatasan fungsi tujuan yang digunakan yang mengarah ke perkiraan berlebihan dari signifikansi fitur.

Kelemahan yang dimiliki *feature selection* bisa diatasi dengan penggunaan metode sebagai pendukung dalam penentuan hasil akhirnya. Salah satu metode yang ditekankan disini adalah mutual information. *Mutual information* menggunakan informasi timbal balik dan kriteria 'maksimum minimum', yang mengurangi masalah terlalu tinggi dari signifikansi fitur seperti yang ditunjukkan baik secara teoritis maupun eksperimental. Mutual information sudah diterapkan juga dalam berbagai penelitian untuk menyelesaikan permasalahan dalam hal seleksi fitur. Dari peneltian sebelumnya mutual information sering digunakan untuk klasifikasi dan mengidentifikasi suatu permasalahan [12] [13].

Dalam penyelesaian untuk algoritma klasifikasi, metode lain yang bisa digunakan adalah *K-Nearest Neighbours* (KNN), *Bayes classification* dan *Support Vector Machine* (SVM). Ketiga metode ini sudah biasa digunakan dalam penelitian sebelumnya. Seperti dalam penelitian [14] [15], membahas tentang klasifikasi teks menggunakan dimana dengan melakukan pendekatan menggunakan teknik data mining dari algoritma KNN untuk memprediksi skor. Algoritma yang diusulkan membagi teks jawaban subjektif untuk banyak kata / frasa menggunakan kamus, yang dapat mengatasi bahasa dalam studi kasus yang dikerjakan. Setelah itu, algoritma KNN diterapkan dengan algoritma kesamaan yang diusulkan. Kemiripan yang diajukan didasarkan pada pencocokan kata dan pengurutan kata. Jika pasangan kata / frasa pencocokan teks dan memiliki urutan yang sama, kesamaannya tinggi.

Sedangkan penelitian yang juga menyajikan tentang *text classification* menggunakan *bayes classification.* Dalam penelitian [16] dijelaskan bahwa *Naive Bayes* terus menjadi salah satu metode populer untuk kategorisasi teks karena kesederhanaan, efisiensi, dan kemanjurannya. Dalam semua pendekatan pembobotan fitur yang ada, bobot fitur yang dipelajari hanya diterapkan pada klasifikasi rumus *Naive Bayes*. Hal yang sama juga dilakukan dalam penelitian [17], dimana disajikan teknik klasifikasi semantik Naive Bayes yang didasarkan pada model ruang tensor kami untuk representasi teks. gambar,

peneliti menerapkan tiga algoritma yaitu: KNN, *Naive Bayes* dan *reverse* DBSCAN algoritma, kemudian dilakukan perbandingan kinerja dari ketiga algoritma berdasarkan empat faktor pengukuran yaitu: presisi, sensitivitas, spesifitas dan akurasi.

Sedangkan dalam peneltian [18], untuk mendeteksi teks dalam spam email berbasis Dari perbandingan ketiga algoritma dan empat faktor pengukuran bisa mendapat akurasi yang baik. Dalam penelitian [19], pengunaan metode SVM dalam *Cyberbullying* Klasifikasi Komentar pada Selebgram Indonesia. Metode Support SVM [20] digunakan untuk jenis klasifikasi yang hanya memiliki dua nilai, yaitu -1 dan 1. Dalam proses klasifikasi ini, metode *Support Vector Machine* (SVM) digunakan untuk melihat seberapa jauh metode ini dapat mengklasifikasikan komentar pada akun Indonesia Berisi penindasan maya atau tidak.

3. METODOLOGI PENELITIAN

3.1 Data

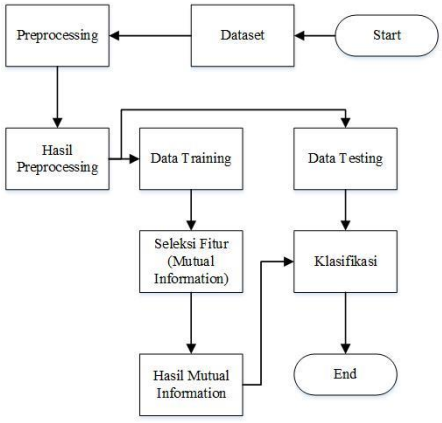
Data yang digunakan dalam penelitian ini 100 data training berita yang terbagi menjadi 40 berita politik, 30 berita ekonomi dan 30 berita sosial. Data berita tersebut di crawler dari website berita setelah itu dilabeli dan dikategorikan oleh kedua penulis. Selain itu juga dalam penelitian ini menggunakan database kata-kata dalam bahasa Indonesia yang digunakan dalam algoritma stemming Nazief dan Adriani.

Tabel 1. Data Training

No	Nama Kelas	Jumlah Dokumen	Positif	Negatif	Netral
1.	Politik	40	11	14	15
2.	Ekonomi	30	6	3	21
3.	Sosial	30	20	9	1

3.2 Rancangan Sistem

Sistem yang dirancang dalam penelitian ini memiliki dua fungsi utama. Pertama sistem mampu mengkategorikan topik berita dan kedua melabelkan sebuah berita positif atau negatif seperti yang tertera pada gambar 1.



Gambar 1. Flowchart Kategorisasi Berita Menggunakan Mutual Information

Berdasarkan gambar 1 ada beberapa proses yang dilakukan sebelum tahap klasifikasi.

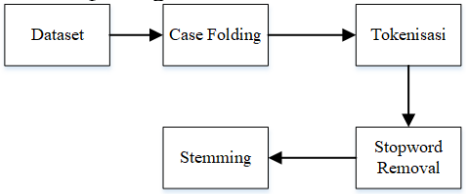
3.3 Dataset

Tahap pertama yang dilakukan adalah dataset, mengumpulkan 100 data training dari beberapa website berita yang terbagi seperti pada tabel 1 di atas ini.

Pada penelitian ini terbagi menjadi dua fungsi utama yaitu kategorisasi berita dibagi menjadi 3 kelas (politik, ekonomi dan sosial) dan pelabelan berita yang dibagi menjadi 3 kelas (positif, negatif dan netral).

3.4 Preprocessing

Preprocesing merupakan proses untuk mempersiapkan 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. Proses Preprocess

Case Folding merupakan tahapan untuk mengubah semua huruf menjadi lower-case.

Tujuan dari case folding adalah untuk menanggulangi ketidak konsistenan penggunaan huruf kapital dan huruf kecil (case sensitive) dalam sebuah artikel berita. Selain itu, pada tahap ini juga dilakukan penghapusan karakter selain huruf seperti tanda baca (seperti (,), (.), dll) dan angka (seperti 4, 64, 35, dll) 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 merupakan suatu proses untuk menghapus kata-kata yang terdapat dalam stopword list dari token list seperti kami, saya, dan, dari, dll [6]. Tujuan dari penggunaan Stopword removal ini adalah untuk mengurangi jumlah kata yang akan diproses. Pada penelitian ini, penulis menggunakan daftar stopword untuk Bahasa Indonesia Stemming Nazief-Andriani. 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 [21].

3.5 Seleksi Fitur (Mutual Information)

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. *Mutual Information* (MI) adalah salah satu cara untuk melakukan seleksi fitur yang digunakan pada penelitian ini. Perhitungan MI dapat dilihat dalam persamaan 1 [22].

$$I(U;C) = \frac{N_{11}}{N}\ log_2\ \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\ log_2\ \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\ log_2\ \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\ log_2\ \frac{NN_{00}}{N_{0.}N_{.0}}$$

dimana setiap N memiliki dua kondisi yang direpresentasikan dengan angka 1 artinya ada dan 0 artinya tidak ada. Sebagai contoh, N_{10} artinya ada pada kondisi pertama dan tidak ada pada kondisi kedua. Untuk $N_{1.} = N_{10} + N_{11}$ adalah jumlah dokumen yang mengandung kondisi pertama. $N = N_{11} + N_{10} + N_{01} + N_{00}$ adalah jumlah keseluruhan dari dokumen.
Hasil dari perhitungan mutual information berisi nilai keterkaitan antara satu atribut dengan atribut lainnya, semakin besar nilai MI, maka semakin besar keterkaitan antar atribut tersebut. Pada penelitian ini akan mencari nilai MI untuk semua atribut terhadap setiap kelas. Setelah itu diambil 10 atribut dengan nilai keterkaitan yang tertinggi untuk mewakili setiap kelas dalam proses pengkategorian sebuah berita.

3.6 Manhattan Distance

Manhattan distance merupakan salah satu cara untuk menghitung jarak suatu atribut dengan atribut lainnya yang biasa juga disebut city blok distance. Perhitungan manhattan distance dapat dilihat pada persamaan 2.

$$(X, Y) = (|X1 - Y1| + |X2 - Y2| + \ldots + |XN - 1 - YN - 1| + |XN - YN|)$$

dimana X dan Y merupakan atribut yang ingin dihitung jaraknya. Pada penelitian ini X dan Y adalah sebuah kata yang akan dibandingkan dengan kata lain pada setiap dokumen. X dan Y bernilai 1 jika ada dan 0 tidak ada maka dari perhitungan ini didapatkan sebuah nilai similaritas dari sebuah data testing. Semakin tinggi nilai similaritas data traninig terhadap data testing, semakin dekat pula pelabelan terhadap data testing.

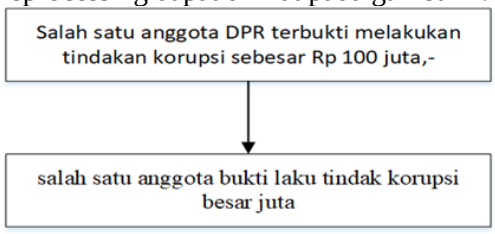
3.7 Klasifikasi KNN

Gambaran secara umum pada KNN, yaitu perhitungan jarak dari data testing terhadap semua data training. Kemudian diambilah sejumlah ‘k’ pada data training terhadap jarak yang terdekat. Data testing dapat ditentukan berdasarkan kategori mayoritas dari tetangga yang terdekat. Jika jumlah mayoritas dari tetangga sama maka akan dihitung nilai rata-rata dari nilai similarity yang telah didapatkan sebelumnya.

4. HASIL DAN ANALISIS

Berdasarkan metode penelitian yang telah dirancang maka pada bagian ini akan dibahas tentang hasil – hasil yang telah didapatkan. Dalam implementasi suatu sistem dapat diketahui cara kerja suatu sistem yang dijalankan, apakah telah berjalan baik atau tidak.

Untuk mengetahuinya, program ini dibangun dengan menggunakan bahasa pemrograman *PHP* dan menggunakan *database SQL*. Pada tahap pertama dalam pembangunan sistem dilakukan preprocessing terhadap seluruh data training begitu pula untuk data testing. Contoh proses preprocessing dapat dilihat pada gambar 4.



Gambar 4. Contoh Proses Preprocessing

Pada gambar 4 dapat dilihat bahwa tahapan-tahapan preprocessing yang telah dilakukan adalah case folding mengkoversi semua menjadi huruf kecil dan juga penghapusan tanda baca maupun angka. Setelah itu dilakukan tokenisasi dengan membagi teks menjadi beberapa bagian berdasarkan kata tanpa melihat makna keseluruhan teks. Selanjutnya dilakukan penghapusan kata yang termasuk golongan stopword, pada kalimat di atas DPR termasuk kata stopword karena kata yang subjektif. Terakhir dilakukan proses stemming terhadap seluruh kata yang ada, sehingga seluruh kata yang awalnya memiliki imbuhan berubah menjadi kata dasar seperti terbukti menjadi bukti.

4.1 Klasifikasi Kategori Berita

Setelah itu akan dilakukan seleksi fitur mutual information terhadap seluruh data training yang ada berdasarkan tiap kelas (politik, ekonomi dan sosial) yang telah melalui proses preprocessing. Hasil dari mutual information dapat dilihat pada 3 tabel berikut.

Kata-kata yang telah didapatkan ini nantinya akan digunakan dalam proses klasifikasi berita berdasarkan kategori. Data testing akan dicocokan dengan kata-kata yang telah didapatkan dari seleksi fitur mutual information. Semakin banyak kata yang sama, semakin dekat pula klasifikasi berita tersebut.

Table 2. Hasil Mutual Information Kategori Politik

Kata	Hasil Mutual Information
guna	0.1001
hak	0.1001
luar	0.0846
mau	0.0776
sedang	0.0666
baru	0.0597
dukung	0.0535
temu	0.0455
upaya	0.0435
salah	0.0411

Tabel 3. Hasil Mutual Information Kategori Ekonomi

Kata	Hasil Mutual Information
mau	0.3227
arus	0.3227
jalan	0.2312
jasa	0.1666
harus	0.1666
siap	0.1666
atur	0.1329
total	0.1329
diri	0.1329
baik	0.1329

Table 4. Hasil Mutual Information Kategori Sosial

Kata	Hasil Mutual Information
api	0.1527
toleransi	0.1354
guna	0.1001
hak	0.1001
toleran	0.0856
luar	0.0846
arif	0.0818
mau	0.0776
sedang	0.0666
baru	0.0597

4.2 Pelabelan Sentimen Berita

Pada proses pelabelan sentimen berita, dilakukan proses perhitungan jarak antara data testing dengan data training yang

menggunakan manhattan distance. Pada gambar 5 menampilkan salah satu uji coba data testing terhadap sistem dengan menampilkan nilai dari manhattan distance.

Dari gambar dapat dilihat nilai similarity yang didapatkan atas perhitungan jarak antara data testing dan data training. Karena pada percobaan ini menggunakan algoritma klasifikasi KNN dengan nilai 'k = 5' maka ditampilkan 5 berita dengan nilai similarity tertinggi. Dari hasil tersebut dapat dilihat sentimen dari 5 data training yang memiliki nilai similarity tertinggi adalah 4 negatif dan 1 positif. Maka dari itu hasil sentimen dari data testing yang diujji coba adalah negatif karena 5 tetangga yang memiliki sentimen negatif lebih banyak daripada positif.

No	JUDUL	SENTIMENT	SIMILARITY
1	PSI Heran Pergantian Jabatan Pemprov DKI Tidak Transparan - Tribunnewscom	negatif	17
2	Mantan Panglima GAM dan Eks Kadensus 88 Demo Desak KPK Bebaskan Gubernur Aceh Kompascom	negatif	15
3	Kemdikbud Sarankan Regrouping Bagi Sekolah Sepi Peminat Republika Online	positif	15
4	51 Persen Kasus Properti karena Pembangunan Tak Jelas Kompascom	negatif	13
5	Harga Pertalite di Sumbar Naik Ini Penyebabnya Republika Online	negatif	12

Gambar 5. Contoh Pengujian Data testing untuk Pelabelan Sentimen Berita

5. Kesimpulan

Dari hasil penelitian yang telah dilakukan dapat ditarik kesimpulan bahwa penggunaan seleksi fitur mutual information pada pengklasifikasikan berita politik, ekonomi dan sosial berhasil begitu pula dengan pelabelan sentimen berita menggunakan algoritma KNN dengan manhattan distance. Sistem dapat mengklasifikasikan data testing yang diuji coba dan memberika pelabelan sentimen berita. Ada sebuah kondisi dimana jika kata yang dihasilkan pada proses mutual information tidak terdapat data testing maka hasil pengkategorian tidak diketahui. Ini disebabkan karena jumlah data training yang masih kurang.

Akan tetapi salah satu hambatan pada penelitian ini adalah jumlah data training yang digunakan masih sedikit sehingga kata-kata yang didapatkan pada proses mutual information masih banyak kata yang umum. Maka dari itu diharapakan mengembangkan penelitian ini dengan jumlah data training yang besar sehingga dapat menghasilkan keluaran sistem yang lebih akurat. Selain itu penggunaan algoritma klasifikasi dapat pula digunakan pada pengkategorian berita sehingga menambah akurasi dengan menggabungkan seleksi fitur dan algoritma klasifikasi. Selain itu pula diharapkan dapat menghitung akurasi dari sistem yang telah dirancang sehingga dapat diperbandingkan dengan algoritma dan seleksi fitur yang lainnya.

6. DAFTAR PUSTAKA

[1] D. Afrianto, "96% Masyarakat Indonesia Konsumsi Berita Online," Okezone.Com, 2018. .

[2] P. Widodo, J. A. Putra, S. Afiadi, A. Z. Arifin, and D. Herumurti, "Klasifikasi Kategori Dokumen Berita Berbahasa Indonesia dengan Metode Kategorisasi Multi-Label Berbasis Domain Specific Ontology," J. Teknosains, vol. II, no. 2, pp. 101–112, 2017, doi: 10.22146/teknosains.86111.

[3] S. B. Setiawan and M. S. Mubarok, "Klasifikasi Topik Berita Berbahasa Indonesia menggunakan Weighted K-Nearest Neighbor," vol. 5, no. 1, pp. 1–7, 2015.

[4] I. Maulida, A. Suyatno, and H. R. Hatta, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," JSM STMIK Mikroskil, vol. 17, no. 2, pp. 249–258, 2016.

[5] S. Anisah, A. Pujiastuti, P. S. Informatika, S. Tinggi, and T. Adisutjipto, "Klasifikasi Teks Menggunakan Chi Square Feature Selection Untuk Menentukan Komik Berdasarkan Periode , Materi Dan Fisikdengan Algoritma Naivebayes," pp. 59–66, doi: 10.1007/s10903-015-0186-0.

[6] F. S. Nurfikri and M. S. Mubarok, "Klasifikasi Topik Berita Menggunakan," vol. 5, no. 1, pp. 1579–1588, 2018.

[7] A. Rahman, "Online News Classification Using Multinomial Naive Bayes," vol. 6, no. 1, 2017, doi: 10.1177/1096348015584441.

[8] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A Survey on Evolutionary Computation Approaches to Feature Selection," IEEE Trans. Evol. Comput., vol. 20, no. 4, pp. 606–626, 2016, doi: 10.1109/TEVC.2015.2504420.

[9] M. F. Akay, F. Abut, M. Özçiloğlu, and D. Heil, "Identifying the discriminative predictors of upper body power of cross-country skiers using support vector machines combined with feature selection," Neural Comput. Appl., vol. 27, no. 6, pp. 1785–1796, 2016, doi: 10.1007/s00521-015-1986-9.

[10] and Q. W. Qin Zou , Lihao Ni , Tong Zhang, "Deep Learning Based Feature Selection for Remote Sensing Scene Classification," IEEE Trans. Magn., vol. 44, no. 11 PART 2, pp. 4045–4048, 2015, doi: 10.1109/LGRS.2015.2475299.

[11] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," Expert Syst. Appl., vol. 42, no. 22, pp. 8520–8532, 2015, doi: 10.1016/j.eswa.2015.07.007.

[12] C. Devi Arockia Vanitha, D. Devaraj, and M. Venkatesulu, "Gene expression data classification using Support Vector Machine and mutual information-based gene selection," Procedia Comput. Sci., vol. 47, no. C, pp. 13–21, 2014, doi: 10.1016/j.procs.2015.03.178.

[13] R. Martínez-Cancino, J. Heng, A. Delorme, K. Kreutz-Delgado, R. C. Sotero, and S. Makeig, "Measuring transient phase-amplitude coupling using local mutual information," Neuroimage, vol. 185, no. October 2018, pp. 361–378, 2019, doi: 10.1016/j.neuroimage.2018.10.034.

[14] C. Wang, "A K-Nearest Neighbor Algorithm Based on Cluster in Text Classification Chun-Y," pp. 225–228, 2010, doi: 10.1109/CMCE.2010.5610477.

[15] K. Neighbors, "Text Classification for Subjective Scoring Using," 2018 Int. Conf. Digit. Arts, Media Technol., pp. 139–142, 2018.

[16] Q. Jiang, W. Wang, X. Han, S. Zhang, X. Wang, and C. Wang, "Deep feature weighting in Naive Bayes for Chinese text classification," Proc. 2016 4th IEEE Int. Conf. Cloud Comput. Intell. Syst. CCIS 2016, pp. 160–164, 2016, doi: 10.1109/CCIS.2016.7790245.

[17] H. J. Kim, J. Kim, and J. Kim, "Semantic text classification with tensor space model-based naïve Bayes," 2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc., pp. 4206–4210, 2017, doi: 10.1109/SMC.2016.7844892.

[18] A. Harisinghaney, A. Dixit, S. Gupta, and A. Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm," ICROIT 2014 - Proc. 2014 Int. Conf. Reliab. Optim. Inf. Technol., pp. 153–155, 2014, doi: 10.1109/ICROIT.2014.6798302.

[19] M. Andriansyah et al., "Cyberbullying comment classification on Indonesian Selebgram using support vector machine method," Proc. 2nd Int. Conf. Informatics Comput. ICIC 2017, vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/IAC.2017.8280617.

[20] A. Gormantara, “Analisis Sentimen Terhadap New Normal Era di Indonesia pada Twitter Analisis Sentimen Terhadap New Normal Era di Indonesia pada Twitter Menggunakan Metode Support Vector Machine,” no. July, pp. 0–5, 2020.

[21] D. Wahyudi, T. Susyanto, and D. Nugroho, “IMPLEMENTASI DAN ANALISIS ALGORITMA STEMMING NAZIEF & ADRIANI DAN PORTER PADA DOKUMEN BERBAHASA INDONESIA,” J. Ilm. SINUS, vol. 15, no. 2, 2017, doi: 10.30646/sinus.v15i2.305.

[22] H. S. Christopher D. Manning , Prabhakar Raghavan, Introduction, An Retrieval, Information, no. c. 2009.