

Klasifikasi Topik Berita Berbahasa Indonesia menggunakan *Multilayer Perceptron*

Grace Tika¹, Adiwijaya²

Fakultas Informatika, Universitas Telkom, Bandung

¹gracetikaa@students.telkomuniversity.ac.id, ²adiwijaya@telkomuniversity.ac.id

Abstrak

Berita adalah salah satu representasi dari kehidupan nyata sosial yang selalu menggambarkan isu-isu yang masih banyak dan sering dibagikan oleh masyarakat. Setiap hari, selalu ada pembaruan berita sehingga menjadikannya berita yang baru dan membuat berita mengalir semakin banyak. Dengan adanya internet membuat berita lebih mudah dan cepat tersebar luaskan. Karena itu dibutuhkan sebuah sistem yang dapat melakukan klasifikasi terhadap berita agar berita tersebut dapat dikelompokkan berdasarkan kategori seperti Teknologi, Budaya, Pendidikan, dan lain-lain. Salah satu cara yang dapat mengatasi permasalahan tersebut adalah dengan adanya sistem klasifikasi menggunakan Jaringan Syaraf Tiruan (JST). Metode jaringan syaraf tiruan ini dapat menciptakan suatu pola pengetahuan dengan kemampuan belajar (*self organizing*) dan secara optimasi, jaringan syaraf tiruan dapat menemukan jawaban terbaik. Sehingga metode jaringan syaraf tiruan ini menarik dan bagus untuk mengatasi permasalahan tersebut. Dari hasil terbaik penelitian yang didapat ini dinyatakan dalam *F1-measure micro-average* dengan nilai performansi mencapai 77,44% dimana data berita berbahasa Indonesia yang digunakan memiliki berbagai macam keragaman yang dapat mempengaruhi proses *stemming*.

Kata kunci : *Berita, text preprocessing, text classification, Multilayer Perceptron*

Abstract

News is one of the representation of real social life that always illustrates issues that are still many and often shared by the community. Every day, there is always a news update so that it makes new news and makes the news flow more and more. With the internet, the news makes it easier and faster to spread. Because of that problem, it requires a system that can classify news so that the news can be grouped by categories such as Technology, Culture, Education, and others. One method that can overcome this problem is the existence of a classification system using Artificial Neural Networks (ANN). This artificial neural network method can create a pattern of knowledge with self-organizing and optimization, artificial neural networks can find the best answer. So this artificial neural network method is interesting and good for overcoming these problems. The results of this research are stated in *F1-measure micro-average* with performance values reaching 77.44% which is the data of Indonesian news has various types that can influence the stemming process.

Keywords: *News, text preprocessing, text classification, Multilayer Perceptron*

1. Pendahuluan

Berita adalah salah satu konstruksi dari kehidupan nyata social yang selalu menggambarkan isu-isu yang masih banyak dan sering dibagikan oleh masyarakat.[1] Berita juga merupakan sarana informasi yang penting dimana didalamnya mencakup kejadian dan peristiwa penting sehari-hari. Dengan majunya teknologi saat ini berita juga menyebar melalui Internet. Menurut data yang diperoleh dari Direktur Eksekutif Nielsen Media, Hellen Katherina mengatakan bahwa “Tingginya frekuensi penggunaan internet diantara pembaca media cetak yang mencapai 86%, yaitu di atas rata-rata yang sebesar 61% semakin memperkuat fakta bahwa pembaca media cetak berasal dari kalangan yang lebih affluent,”¹ Sehingga tidak dapat dipungkiri bahwa saat ini setiap orang memiliki kebutuhan untuk mendapatkan informasi berita dari Internet. Selain itu data teks sangat berkembang luas di Indonesia dan banyak perusahaan Indonesia menggunakan data teks berita sebagai data yang diolah sehingga menjadi suatu usaha yang menjanjikan.

Oleh sebab itu, dibutuhkan suatu sistem yang dapat melakukan klasifikasi kategori berita secara otomatis dan sistematis sehingga dapat mempermudah para pembaca berita melalui internet tanpa harus mencari ribuan berita. Bentuk dari solusinya yaitu membagi berita-berita tersebut kedalam beberapa kategorinya yaitu olahraga, kesehatan, otomotif, pendidikan, teknologi, ekonomi, dan lain-lain. Salah satu cara yang dapat dilakukan dengan menggunakan teknik *Artificial Intelligence* yaitu *K-Nearest Neighbor*, *Artificial Neural Network*, *Support Vector Machine*, dan sebagainya. Berdasarkan penelitian paper yang berjudul “*Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction*”[2], dinyatakan bahwa metode jaringan syaraf tiruan ini memberikan nilai akurasi yang lebih tinggi dimana hasil akurasi yang didapatkan sebesar 86,9%. Sehingga pada tugas akhir ini akan dibangun sebuah sistem klasifikasi topik berita menggunakan metode Jaringan Syaraf Tiruan (JST).

¹ Nielsen.com

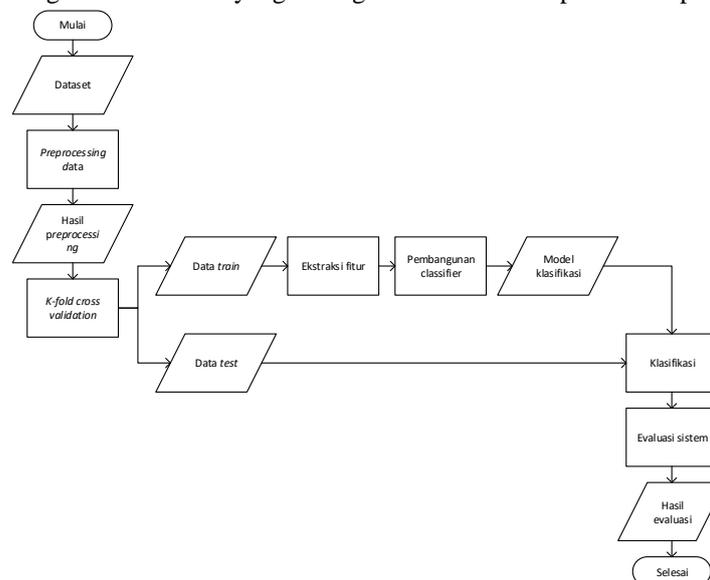
2. Studi Terkait

Pada penelitian yang dilakukan oleh Sigit Bagus, metode yang digunakan adalah metode TF-IDF dan *weighted k-nearest neighbor* yang memperoleh akurasi sebesar 75,86%. Pada penelitian tersebut juga menyimpulkan bahwa hasil dari data latih sangat berpengaruh untuk pembelajaran sistem, sehingga akurasi terbaik pada penelitian tersebut dimana data latih dan data tes adalah 80% berbanding 20%. Selain daripada jumlah data latih, parameter k dan pembobotan pada metode klasifikasi *weighted k-NN* mempengaruhi akurasi sistem [3]. Dan pada penelitian yang dilakukan oleh Komang Ari Widani, metode yang digunakan adalah metode *Chi-Square* dan *multinomial naïve bayes* yang memperoleh akurasi sebesar 85,28%. Pada penelitian tersebut penggunaan tanpa *stemming* menghasilkan akurasi yang lebih baik dibandingkan menggunakan *stemming* dan menggunakan seleksi fitur yaitu *Chi-Square* belum mampu untuk meningkatkan performansi *Multinomial Naive Bayes Classifier*, karena semakin kecil *level of significance* yang digunakan pada seleksi fitur membuat semakin sedikit pula jumlah fitur yang digunakan untuk *learning* model klasifikasi. Tetapi penggunaan seleksi fitur *Chi-Square* mampu mengurangi waktu komputasi model klasifikasi *Multinomial Naive Bayes* dengan kecepatan yang dapat ditingkatkan yaitu sebesar 6%-34% [4]. Pada penelitian yang dilakukan oleh Fahmi Salman Nurfikri, metode yang digunakan pada topik berita yaitu menggunakan metode *mutual information* dan *bayesian network* yang menghasilkan akurasi sebesar 75,34% dimana fitur seleksi *Mutual Information* mampu meningkatkan akurasi dari klasifikasi *Bayesian Network* [5].

Metode *multilayer perceptron* sangat menarik untuk diteliti karena berbagai masalah yang dapat dipecahkan oleh metode *multilayer perceptron* ini dapat menghasilkan akurasi yang cukup tinggi. Pada penelitian Novianty yang bertema pengenalan *gender* pada pasfoto menggunakan *backpropagation* mendapatkan hasil akurasi sebesar 100% dimana pada paper tersebut melakukan percobaan terhadap *hidden layer*. Apabila jumlah *hidden layer* semakin banyak membuat waktu komputasi semakin lama tetapi membuat performansi sistem semakin baik dan nilai MSE semakin kecil. Sehingga penggunaan metode jaringan syaraf tiruan *Backpropagation* membuat sebuah sistem yang dapat mengenali jenis kelamin dengan baik [6]. Salah satu paper acuan lainnya yaitu pada penelitian Yuslan Abu Bakar yang berjudul "*Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language translation) using Information Gain and Backpropagation Neural Network*" mendapatkan hasil akurasi sebesar 88.42% untuk *multi label* pada *hadith* dan 65.275% dari *single label hadith* dimana fitur seleksi *Information Gain* mampu meningkatkan performansi untuk *multi label* dibandingkan *single label* [7].

3. Sistem yang Dibangun

Berikut merupakan gambaran sistem yang dibangun mulai dari tahap awal sampai tahap akhir:



Gambar 1 Flowchart sistem secara umum

3.1 Data Set

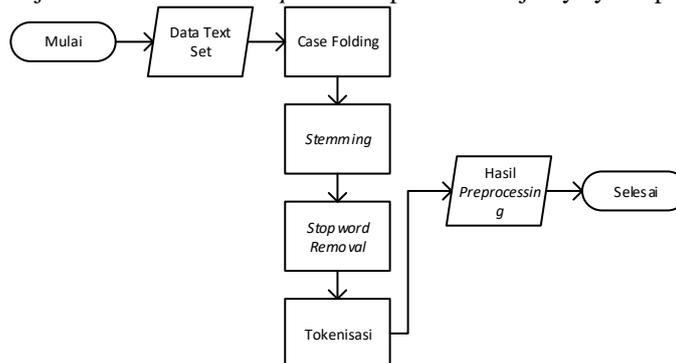
Data yang digunakan untuk penelitian tugas akhir ini yaitu sekumpulan data teks yang berisi berita dari berbagai macam sumber di Internet. Beberapa contoh sumber berita dari internet yaitu: nationalgeographic.co.id, kompas.com, tribunnews.com, detik.com, sindonews.com, aktual.com, republika.co.id, tempo.co, okezone.com, cnnindonesia.com, liputan6.com, merdeka.com, dan suara.com.

Berita yang akan diambil dimulai dari bulan Agustus 2016 sampai dengan bulan Februari 2017 yang sudah digunakan pada penelitian yang berjudul “Klasifikasi Topik Berita menggunakan *Mutual Information* dan *Bayesian Network*” [5]

Pada penelitian ini data yang dikumpulkan berupa berita sesuai dengan 12 kategori yaitu Olahraga, Kesehatan, Otomotif, Pendidikan, Teknologi, Wisata, Politik, Properti, Hukum, Hiburan, Ekonomi, dan Gaya Hidup. Masing-masing kategori terdapat 30-32 data teks berita dengan total keseluruhan data yaitu 368 *data set*.

3.2 Text Preprocessing

Preprocessing pada teks merupakan proses yang dilakukan pertama kali untuk mengolah sebuah data yang banyak sampai menjadikan sebuah hasil *input* untuk proses selanjutnya yaitu proses klasifikasi.



Gambar 2 Flowchart sistem pada text preprocessing

- *Case Folding*

Case Folding merupakan langkah untuk menghilangkan *Uppercase* atau huruf kapital menjadi *Lowercase* atau huruf kecil. Pada saat melakukan langkah ini juga menghilangkan karakter-karakter seperti contoh koma(,), tanda seru(!), *ampersand*(&), garis miring(/), dan sebagainya.

- *Stemming*

Pada tahap *stemming* adalah proses memotong kata berimbuhan menjadi kata dasar dengan menerapkan aturan-aturan tertentu. Salah satu contohnya adalah kata “kesukaan”, “disukai” dan “menyukai” akan diubah menjadi kata dasarnya yaitu “suka”. *Stemming* juga merupakan proses menghilangkan imbuhan(*affixes*) yang terdiri dari awalan(*prefix*), akhiran(*suffixes*), sisipan(*infixes*) dan kombinasi awal dan akhir(*confixes*).[8]

- *Stopword Removal*

Ketika suatu kata yang sering muncul tetapi tidak memiliki *value* atau informasi penting seperti contoh “saya”, ”kamu”, ”yang”, “di”, dan sebagainya. Sehingga kata tersebut akan dihilangkan atau tidak diproses.

- *Tokenisasi*

Pada tahap tokenisasi dilakukan pemecahan setiap kalimat menjadi berbagai kata atau string.

3.3 K-Fold Cross Validation

K-fold cross validation merupakan salah satu cara untuk melakukan *testing* pada program yang dibangun. Pada tugas akhir ini *k-fold* dilakukan dengan cara membagi data *train* menjadi lima bagian dan secara bergantian setiap bagiannya menjadi data tes.

3.4 Ekstraksi Fitur

Ekstraksi fitur digunakan untuk membuat hasil dari teks atau dokumen menjadi sebuah representasi vektor dan ekstraksi fitur yang digunakan pada penelitian ini yaitu *term frequency* (TF) dan TF-IDF. *Term frequency* adalah ekstraksi fitur dimana TF menghitung banyaknya keberadaan suatu *term* atau fitur. Sedangkan TF-IDF adalah ekstraksi fitur dimana menghitung pembobotan dari setiap fitur. Ekstraksi fitur dilakukan setelah melalui *preprocessing* data dimana akan dilakukan percobaan terhadap keduanya.

	tuhan	mobil	renang	pantai
D1	2	0	2	0
D2	4	0	0	0

D3	0	2	4	1
----	---	---	---	---

Tabel 1 Contoh Term Frequency

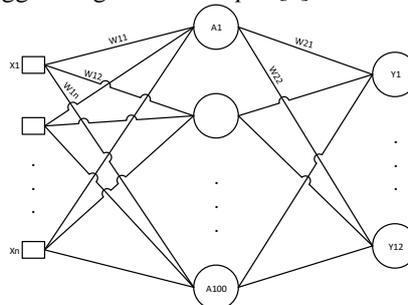
	tuhan	mobil	renang	pantai
D1	0.2432	0	4.4321	0
D2	0.7787	0	0	0.7899
D3	2.230	3.1230	3.2344	1.8773

Tabel 2 Contoh TF-IDF

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{1}$$

3.5 Multilayer Perceptron

Pada tahap ini merupakan tahap dimana hasil dari tahap sebelumnya (*preprocessing* dan ekstraksi fitur) kemudian diproses menggunakan berbagai metode untuk klasifikasi. Klasifikasi merupakan teknik untuk menetapkan kelas atau kategori pada suatu data. Pada tugas akhir ini metode yang digunakan adalah metode klasifikasi menggunakan *Multilayer Perceptron* (MLP). Metode ini adalah suatu metode pembelajaran mesin yang merepresentasikan cara kerja yang hampir menyerupai sistem saraf manusia. Sistem ini mempunyai banyak *input* dan satu *output*, serta terdiri dari proses-proses dalam jumlah besar yang terhubung satu sama lain, serta memiliki layer. Sama seperti jaringan saraf pada manusia, *Multilayer Perceptron* juga memiliki neuron yang berfungsi menerima *input* dari neuron lainnya, lalu dihitung sedemikian rupa dengan menggunakan suatu fungsi sehingga menghasilkan *output*. [9]



Gambar 3 Design model MLP

Pada penelitian ini struktur model yang digunakan yaitu dimana terdapat tiga *layer*: *layer input*, *hidden layer*, dan *layer output*.

Model MLP	Keterangan
<i>Input layer</i>	Sesuai dengan jumlah dari hasil ekstraksi fitur
<i>Hidden layer</i>	Satu <i>hidden layer</i> dengan jumlah 100 <i>neuron</i>
<i>Output layer</i>	Jumlah 12 <i>neuron</i>
<i>Epoch</i>	75
Fungsi aktivasi	<i>Sigmoid</i>
<i>Learning rate</i>	0.001
<i>Optimizer</i>	<i>Stochastic gradient descent</i> (SGD)
<i>Momentum</i>	0.9
<i>Nesterovs_momentum</i>	<i>True</i>

Tabel 3 Struktur model MLP

3.6 Evaluasi Sistem

Evaluasi ini dilakukan dengan menghitung nilai akurasi, *precision*, *recall* dan *f-measure*(F1-Score).

	Prediksi Sistem (Positif)	Prediksi Sistem (Negatif)
Aktual (Positif)	<i>True Positive</i> (TP)	<i>False Negative</i> (FN)
Aktual (Negatif)	<i>False Positive</i> (FP)	<i>True Negative</i> (TN)

Tabel 4 Confusion Matrix

$$Precision = \frac{TP}{TP + FP} = \frac{\sum_{i=0}^{11} TP_i}{\sum_{i=0}^{11} (TP_i + FP_i)} \quad (18)$$

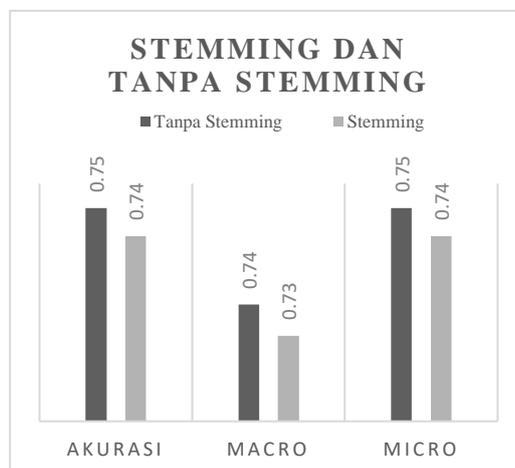
$$Recall = \frac{TP}{TP + FN} = \frac{\sum_{i=0}^{11} TP_i}{\sum_{i=0}^{11} (TP_i + FN_i)} \quad (19)$$

$$f1(\text{micro \& macro}) = \frac{2 * precision * recall}{precision + recall} \quad (20)$$

Selanjutnya mengukur nilai *Precision*. *Precision* sendiri adalah nilai rasio perbandingan jumlah dokumen positif yang terklasifikasikan dengan benar. Nilai TP dan FP yang digunakan adalah jumlah dari nilai TP kelas 0 sampai 11, dan sama dengan FP dari jumlah nilai FP dari kelas 0 sampai 11. Rumus yang digunakan untuk perhitungan *Precision* dapat dilihat pada rumus (18). Setelah itu menghitung *recall* yang dapat dilihat pada rumus (19). Terakhir melakukan kombinasi rata-rata *harmonic* dari nilai *precision* dan *recall* dengan menggunakan rumus *f1 micro-averaged* yang dapat dilihat pada rumus (20)[10]. Perbandingan penggunaan *f1 macro-averaged* yang serupa dengan rumus pada (18) sampai (20), hanya saja penggunaan *macro-averaged* menghitung rata-rata dari setiap nilai *f1* dari setiap kelas. Untuk perhitungan *macro* dengan *micro*, rumus yang digunakan sama hanya saja ada beberapa tahap yang berbeda. Pada perhitungan *macro* pertama yaitu mencari setiap *confusion matrix* setiap kelas, setelah itu masing hitung *precision*, *recall*, dan *f1* sesuai rumus (18) dan (20) lalu setelah mendapatkan setiap *f1* dari masing-masing kelas selanjutnya nilai *macro* didapatkan dari rata-rata nilai *f1*. Sedangkan perhitungan *micro* membuat sebuah *confusion matrix* dimana isinya penjumlahan dari *true positif* seluruh kelas, *true negatif* seluruh kelas, *false negatif* seluruh kelas, dan *false positif* seluruh kelas setelah itu baru menghitung *precision*, *recall*, dan *precision* sesuai rumus (18) sampai (20).

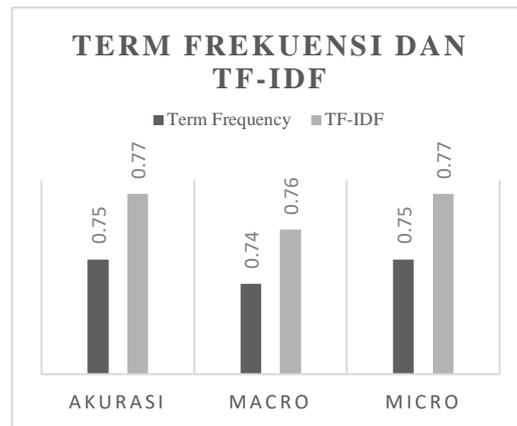
4. Evaluasi

Pada penelitian yang dibangun ada beberapa percobaan, *scenario* pertama yaitu percobaan dengan menggunakan *stemming* dan tanpa *stemming* untuk mengetahui seberapa baik penggunaan *stemming* dalam kasus ini. Skenario kedua yaitu percobaan dengan menggunakan ekstraksi fitur yaitu menggunakan *term frequency* dan TF-IDF untuk mengetahui ekstraksi fitur manakah yang lebih baik dalam performansi klasifikasi. Pada setiap percobaan yang dilakukan nilai akurasi, *macro*, dan *micro* didapatkan dari perhitungan rumus (18) sampai (20) yang dilakukan pada program.



Gambar 4 Skenario pertama, perbandingan penggunaan stemming dan tanpa stemming

Dapat dilihat dari hasil yang diperoleh gambar 3 dapat disimpulkan bahwa hasil menggunakan tanpa *stemming* jauh lebih baik daripada menggunakan *stemming*. Hal tersebut dapat terjadi yang disebabkan pemotongan yang tidak sesuai dengan kata dasarnya, seperti contohnya penggunaan *stemming* pada kata “menyayangi” sehingga hasil dari *stemming* membuang *me-* dan *-I* menjadi “nyayang” yang sebenarnya kata “nyayang” tidak ada dalam kamus bahasa Indonesia. Hal ini dapat mengakibatkan penggunaan *stemming* membuat performansi menurun.



Gambar 5 Skenario perbandingan menggunakan ekstraksi fitur TF-IDF dan term frekuensi

Pada gambar 4 dapat diketahui bahwa penggunaan TF-IDF menghasilkan grafik yang lebih tinggi dimana hasil dari akurasi dan f1-score mencapai persentase tertinggi dari percobaan ini. Hal ini dikarenakan pada ekstraksi fitur TF-IDF melakukan proses dimana fitur yang diolah berdasarkan pembobotan TF-IDF. Semakin besar nilai perhitungan bobot yang diperoleh maka semakin tinggi tingkat similaritas dokumen terhadap kata kunci, sehingga sistem menghasilkan akurasi yang lebih baik. Untuk perhitungan pembandingan f1-score menggunakan *macro average performance* dan *macro average performance* dimana hasil dari *micro average* mendapatkan nilai evaluasi yang lebih baik daripada *macro average*

Dengan menggunakan TF-IDF dapat menghasilkan akurasi dan F1-Score yang lebih baik dibandingkan dengan *term frequency* saja. Sedangkan pada *preprocessing* data menggunakan tanpa *stemming* menghasilkan akurasi dan F1-Score yang lebih baik daripada menggunakan *stemming*. Sama seperti percobaan sebelumnya pengukuran F1-Score dibagi menjadi 2 yaitu *macro* dan *micro average performance*, dimana hasil dari *micro average performance* tetap dinyatakan sebagai akurasi terbaik untuk kasus ini. Sehingga dapat disimpulkan bahwa tanpa menggunakan *stemming* dan fitur ekstraksi TF-IDF adalah metode yang cukup baik untuk permasalahan pada klasifikasi topik berita berbahasa Indonesia ini.

5. Kesimpulan

Berdasarkan penelitian dan pengujian yang sudah dilakukan, dapat disimpulkan bahwa penggunaan metode *Multilayer Perceptron* dengan ekstraksi fitur TF-IDF dan tanpa *stemming* dapat menangani permasalahan mengenai klasifikasi topik berita berbahasa Indonesia, karena telah mencapai hasil akurasi F1-Score tertinggi yaitu sebesar 77,44%. Untuk penelitian berikutnya, sangat diharapkan untuk menggunakan seleksi fitur yang mungkin saja dapat membuat sistem klasifikasi topik berita berbahasa Indonesia ini menjadi lebih baik.

Daftar Pustaka

- [1] M. Fitriah and F. El'Arsya, "Berita Utama Surat Kabar Lokal di Bogor Studi Analisis Isi pada Jurnal Bogor dan Radar Bogor," *J. Komun. Pembang.*, vol. 9, no. 1, pp. 11–19, 2012.
- [2] P. H. Goodman *et al.*, "Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction," pp. 857–862, 1996.
- [3] S. B. Setiawan and M. S. Mubarak, "Klasifikasi Topik Berita Berbahasa Indonesia menggunakan Weighted K-Nearest Neighbor," pp. 1–7, 2015.
- [4] M. S. Komang Ari Widani; Mubarak, "Klasifikasi Artikel Berita Berbahasa Indonesia menggunakan Chi-Square Feature Selection dan Multinomial Naive Bayes," no. September 2017, p. 2017.
- [5] F. S. Nurfikri and M. S. Mubarak, "KLASIFIKASI TOPIK BERITA MENGGUNAKAN."
- [6] A. Novianty, F. T. Elektro, and T. Matching, "PENGENALAN GENDER PADA PASFOTO BERBASIS CIRI PADA WAJAH MENGGUNAKAN ALGORITMA JARINGAN SYARAF TIRUAN BACK PROPAGATION GENDER RECOGNITION ON THE PHOTOGRAPH WITH FACE FEATURES BASED," pp. 1–8.
- [7] M. Yuslan, A. Bakar, and S. Al Faraby, "Multi-Label Topic Classification of Hadith of Bukhari (Indonesian Language translation) using Information Gain and Backpropagation Neural Network," no. 1.
- [8] A. Pratama, M. A. Bijaksana, and M. S. Mubarak, "Evaluasi Eksplicit dan Implisit Algoritma-Algoritma Stemming Bahasa Indonesia Explicit and Implicit Evaluation of Indonesian Stemming Algorithms."
- [9] V. Sharma, S. Rai, and A. Dev, "A Comprehensive Study of Artificial Neural Networks," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 2, no. 10, pp. 278–284, 2012.

- [10] D. Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 1–24, 2011.
- [11] Suyanto, S.T, M.Sc. 2007. Artificial Intelligence: Searching, Reasoning, Planning and Learning. Penerbit Informatika, Bandung, Indonesia
- [12] S. Al Faraby, E. Riviera, and R. Jasin, "Classification of hadith into positive suggestion , negative suggestion , and information Classification of hadith into positive suggestion , negative suggestion , and information," 2018.
- [13] M. Syahrul and M. Dwi, "Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes," vol. 20060, 2017.
- [14] A. I. Pratiwi, "On the Feature Selection and Classification Based on," vol. 2018, 2018.
- [15] R. Abdul Aziz and M. Syahrul Mubarak, "Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naïve Bayes," *Indosc 2016*, no. August, pp. 139–148, 2016.