ANALISIS SENTIMEN PADA TWEET BENCANA ALAM MENGGUNAKAN DEEP NEURAL NETWORK DAN *INFORMATION GAIN*

M. Burhanis Sulthan[1], Imam Wahyudi[2], Luluk Suhartini[3]
[1]Teknologi Informasi, Fakultas Teknik, Institut Sains dan Teknologi Annuqayah
[2]Teknologi Informasi, Fakultas Teknik, Institut Sains dan Teknologi Annuqayah
[3]Teknologi Informasi, Fakultas Teknik, Institut Sains dan Teknologi Annuqayah
[1]burhan.sulthan33@gmail.com, [2]wahyudigo94@gmail.com, [3]lulukkhafi@gmail.com

ABSTRAK

Kemajuan teknologi informasi dan komunikasi membuat informasi terkait bencana alam menjadi lebih cepat tersebar, salah satu soial media yang banyak digunakan yaitu twitter. Pada penelitian ini mengklasifikasikan teks terkait analisis sentimen terhadap bencana alam yang terjadi. Metode klasifikasi yang digunakan adalah Deep Neural Network (DNN),. Jadi untuk mempercepat proses klasifikasi digunakan teknik seleksi fitur yaitu Information Gain (IG) untuk memilih fitur-fitur yang terbaik dari hasil ekstrasi. Kemudian evaluasi dan validasi dilakukan untuk mengetahui hasil kinerja klasifikasi. Digunakan confusion matrix dan 10 fold validasi sebagai proses evaluasi dan validasi didalam penelitian ini. Pada penelitian ini menggunakan beberapa metode yaitu Naïve bayes, Random Forest, Decision Tree dan Support Vector Machine. Hasil akurasi dari metode Deep Neural Network dengan Information Gain lebih besar dari metode yang lain.

Kata kunci : Analisis Sentiment, Information Gain, Deep Neural Network

ABSTRACT

Advances in information and communication technology make information related to natural disasters more quickly spread, one of the widely used social media is twitter. In this study, classifying texts related to whether or not natural disasters occurred. The classification method used is Deep Neural Network (DNN). So, to speed up the classification process, a feature selection technique is used, namely Information Gain (IG) to select the best features from the extraction results. Then evaluation and validation are carried out to determine the results of the classification performance. The confusion matrix and 10 fold validation are used as the evaluation and validation process in this study. This research uses several methods, namely Nave Bayes, Random Forest, Decision Tree and Support Vector Machine. The accuracy result of the Deep Neural Network method with Information Gain is greater than the other methods.

Keywords ***:*** *Sentiment Analysis, Information Gain, Deep Neural Network*

1. PENDAHULUAN

Bencana alam adalah fenomena yang diakibatkan oleh peristiwa atau serangkaian peristiwa yang disebabkan oleh alam antara lain berupa gempa bumi, tsunami, gunung meletus, banjir, kekeringan, angin topan, dan tanah longsor. Fenomena bencana alam terjadi dikarenakan alam berupaya dalam menyeimbangkan ekosistem yang rusak oleh manusia maupun proses alam itu sendiri.

Kemajuan teknologi informasi dan komunikasi membuat informasi terkait bencana alam menjadi lebih cepat tersebar. Itulah salah satu keuntungan dari perkembangan masa terutama teknologi komunikasi. Salah satu media komunikasi yang banyak digunakan ialah twitter, yang mana penggunanya terus meningkat. Media memiliki peran penting dalam bencana alam. Melalui media informasi mengenai bencana alam dapat tersebar ke berbagai penjuru dunia. Informasi mengenai jenis bencana, informasi mengenai kapan terjadinya bencana, informasi mengenai lokasi bencana, dampak, dan kebutuhan korban bencana alam dapat terekam dan tersampaikan melalui pemberitaan.

Kawasan rawan bencana adalah suatu wilayah yang memiliki kondisi atau karakteristik geologis, biologis, hidrologis, klimatologis, geografis, sosial, budaya, politik, ekonomi, dan teknologi yang untuk jangka waktu tertentu tidak mampu mencegah, meredam, mencapai kesiapan, sehingga mengurangi kemampuan untuk menanggapi dampak buruk

bencana alam tertentu dan lambat dalam mensosialisi bencana. Sehingga perlu mengedentifikasi bencana alam yang mungkin terjadi dengan menganalisis sentiment pada tweet terkait bencana alam yang nantinya menjadi sebuah sistem.

Beberapa penelitian sebelumnya yang memiliki hubungan dengan sentimen teks, ditemukan beberapa kelemahan yang dapat disimpulkan. Kelemahan tersebut diantaranya adalah perbedaan dimensi dataset yang digunakan berpengaruh pada hasil akurasi dan performen dalam penggunaan metode klasifikasi yang diusulkan. Dikarenakan dataset yang digunakan memiliki tipe yang berbeda-beda. Pada penelitian [1],[2],[3],[4],[5] banyak sekali dataset yang digunakan tidak inbalance antara negatif dan positif sehingga sangat mempengaruhi hasil akhir dari proses kinerja klasifikasi. Dan juga dataset yang digunakan masih terbatas, kemungkinan akurasi akan menurun jika diterapkan pada dataset yang lebih besar sekaligus tidak inbalance antara negatif dan positif.

Kelemahan selanjutnya yang paling dominan adalah pada biaya komputasi. Pada penelitian yang berhubungan [1] metode yang digunakan adalah deep learning, yang mebutuhkan biaya komputasi yang sangat besar meskipun hasil akurasi cukup bagus dibandingkan dengan penelitian-penelitian yang lain. Kelemahan dari penelitian ini adalah tidak adanya teknik reduksi fitur untuk proses input dari deep learning tersebut, sehingga menambah biaya komputasi pada proses klasifikasi deep learning.

2. PENELITIAN TERKAIT

Penelitian yang dilakukan sebelumnya, dengan topik terkait dengan pembahasan analisis sentiment pada tweet ketika bencana terjadi dari segi text mining. Sehingga nantinya akan menemukan posisi, perbedaan, metode, kesamaan, dan beberapa aspek lain.

a) *Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification*

Pada penelitian ini maslah yang dibahas ialah tidak efektifnya metode untuk linguistic komputasional untuk evaluasi dan analisis isi bahasa Punjabi yang berkaitan dengan bunuh diri petani yang dilaporkan di surat kabar lokal. metode yang digunakan dalam penelitian ini adalah Deep Neural Network (DNN) untuk mengklasifikasikan sentimen pada punjabi teks yang ada di surat kabar. Hasil akurasi yang didapatkan cukup bagus yaitu 90,29%.

b) *Sentiment analysis during Hurricane Sandy in emergency respons*

Masalah yang diangkat dalam penelitian ini ialah ketidakmampuan untuk menyortir dan mengelompokkan data menjadi tipe yang berguna, ketidakmampuan untuk sepenuhnya mempercayai data atau sumber yang tidak diketahui dan kurangnya hubungan antara lokasi kejadian bencana dari tweeting tentang bencana tersebut. Metode yang digunakan adalah Algoritma SentiStrength digunakan untuk pelabelan polaritas negative dan positive selanjutnya menggunakan metode Naïve bayes dan SVM untuk mengklasifikasiakn negative dan positive. Diteruskan dengan geo mapped sentiment analisis untuk mengetahui situasi dari lokasi bencana yang terjadi. hasil yang di dapat menggunakan metode SVM mencapai akurasi 75,91% sedangkan metode SentiStrength mencapai akurasi 53,54% dan untuk metode naive mencapai kurasi 53.34%.

c) *Towards Twitter Sentiment Classification by Multi-Level Sentiment-Enriched Word Embeddings*

Penelitian ini mengangkat masalah dengan menganggap semua kata memiliki polaritas sentimen dalam tweet yang sama dengan selruh tweet, yang mengabaikan kata polaritas sentimennya sendiri. Dengan masalah yang ada peneliti mengusulkan cara untuk belajar sentimen kata spesifik embedding dengan mengeksploitasi baik sumber leksikon dan informasi yang diawasi. Dan mengembangkan metode pembelajaran yang menggunakan kata

sentimen multi tingkat untuk mengatasi dua kata memiliki polaritas yang berlawanan dan peran sintaksis yang sama untuk tugas klasifikasi sentiment. Metode yang digunakan dalam penelitian ini yaitu Multi-level Sentiment-enriched Word Embedding (MSWE) dan Convolution Neural Network (CNN). Hasil akurasi yang dicapai pada penelitian ini sebesar 85,75%.

d) *Classification of Sentiment Reviews using N-gram Machine Learning Approach*

Pada penelitian ini menjelaskan tentang masalah dampak dari metode pembelajaran terawasi pada data berlabel dalam menganalisis sentimen pada review suatu produk. Menggunakan beberapa metode pembelajaran yang terawasi untuk mengetahui tingkat keberhasilan dari metode-metode yang digunakan. Peneliti melakukan komparasi beberapa metode yaitu Support Vector Machine, Naive Bayes, Maximum Entropy dan Stochastic Gradient Descent (SGD) dan fiture ekstrasi dengan unigram, bigram dan trigram. Hasil akurasi yang paling tinggi sebesar 88,89% pada kombinasi SVM dan unigram+bigram+trigram.

e) *Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis*

Pada penelitian ini mebahas tentang masalah untuk menentukan opini dominan dalam sebuah dokumen. Peneliti menentukan polaritas kontektual dari teks, apakah teks itu positif atau negatif. Pada penelitian ini fokus pada analisis sentmen yang dihasilkan dari ulasan produk menggunakan teknik asli penelusuran teks dan menyajikan pendekatan untuk menentukan bagaimana sentimen dapat diklasifikasikan dengan dukungan Vector Machine. Pendekatan yang digunakan dalam penelitian ini yaitu membandingkan Rendom Forest, Support Vector Machine dan metode yang diusulkan oleh peneliti dengan menggabungkan antara dua metode klasifikasi sekaligus Rendom Forest dan Support Vector Machine (RFSVM). Hasil metode yang diusulkan lebih tinggi dari pada metode sebelumnya yaitu dengan nilai F-measure 83,4%.

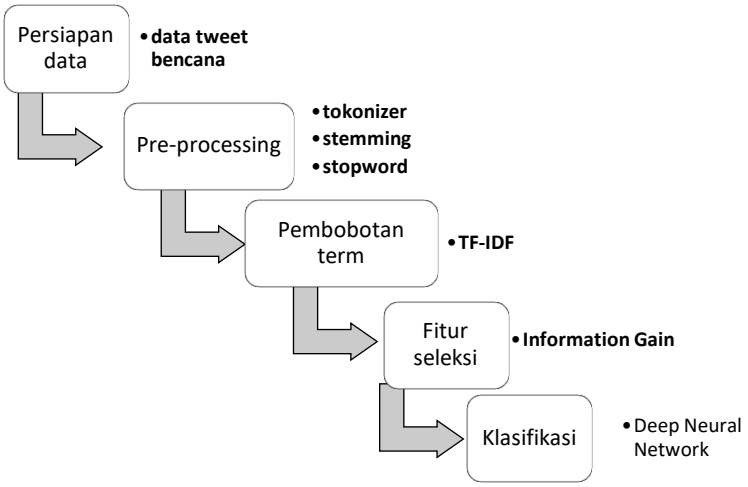
Pada penelitian sebelumnya yang berhubungan masih terdapat peluang untuk posisi penelitian yang akan dilakukan. Penelitian sebelumnya yang kedua tentang identifikasi polaritas sentimen pada kabar berita pada suatu bencana, masih dapat ditingkatkan hasil akurasi dan performa atau kecepatan dari metode klasifikasi yang digunakan. Pada metode Deep Neural Network yang digunakan pada penelitian pertama masih terdapat kelemahan yaitu biaya komputasi yang diperlukan sangat besar, karena tidak mereduksi fitur yang digunakan. Pada penelitian yang akan dilakukan tetap menggunakan deep neural network sebagai klasifikasi tetapi lebih fokus pada meningkatkan hasil akurasi pada klasifikasi dengan fitur seleksi, tidak hanya meningkatkan akurasi tetapi juga dapat mempercepat proses klasisikasi.

3. METODOLOGI PENELITIAN

3.1. Persiapan data

Dalam penelitian ini menggunakan data public yang di ambil dari website kaggle dengan link berikut https://www.kaggle.com/jannesklaas/disasters-on social media. Pada dataset tersebut memiliki 13 atribut yaitu unit_id,golden,unit_state,trusted_judgments,last_judgment_at,choose_one,choose_one:confidence,choose_one_gold,keyword,location,text,tweeted dan userid, tetapi ada beberapa atribut yang tidak berpengaruh pada proses klasifikasi yang nantinya akan dibuang atau tidak digunakan.

Data tersebut hanya digunakan beberapa atribut yaitu text dan chose_one, yang mana atribut chose_one akan dijadikan sebagai target klasifikasi. Atribut text tersebut di klasifikasikan sebagai isu terjadinya bencana yaitu relavant dan not relavant pada atribut chose_one.



Gambar 1 : Proses Model

3.2. Pre-pocessing data

Beberapa tahap yang digunakan dalam preprocessing diantaranya, yaitu tokenization, stopword, dan stemming [6]. Tekonization adalah proses untuk menghilangkan tanda baca pada suatu kalimat dalam suatu dokumen yang akan menghasilkan kata-kata yang berdiri sendiri [7]. Stopword adalah proses penghilangan kata sambung yang tidak relavan dalam penentuan topik sebuah dokumen, misalnya kata "dan", "yaitu", "atau" [1]. Stemming adalah tahap untuk menghilangkan imbuhan-imbuhan pada sebuah kata sampai diperoleh kata dasarnya dalam suatu dokumen [7].

Pada tahap ini yaitu pre-pocessing data, yang mana data mentah atau text dilakukan proses tokenizer, stemming dan stopwords. Hasil dari tahapan tersebut akan menghasilkan fitur-fitur yang nantinya akan digunakan sebagai input pembelajaran mesin oleh Deep Neural Network (DNN).

3.3. Pembobotan term

Menurut penelitian yang dilakukan oleh Mohamed Abdel Fattah [8] mengatakan bahwa frekuensi kata/term adalah seberapa banyak kata-kata yang muncul dalam suatu dokumen. Dalam penelitian ini menghitung term untuk mengetahui tingkat kemunculan term/kata dalam dokumen. Sehingga nantinya akan meningkatkan keefektifan dalam proses selanjutnya. Penelitian ini menerapkan Term Frecuency dan Inverse Dokument Frequency (TF-IDF) sebagai teknik penghitungan frekuensi kata pada dokumen untuk ekstrasi fitur.

Pada tahap pembobotan term ini dilakukan setelah data mentah menjadi suatu dokument matrix melalui proses sebeumnya yaitu pre-pocessing. Pembobotan term merupakan term documents matrix yang representasi kumpulan dokumen yang digunakan untuk melakukan proses klasifikasi dokumen teks. Pada penelitian ini akan digunakan metode TF-IDF sebagai proses pembobotan, yaitu akan dilakukan pembobotan pada tiap term berdasarkan tingkat kemunculan term tersebut di dalam sekumpulan dokumen yang ada.

3.4. Seleksi fitur dengan information gain

Berdasarkan penelitian yang dilakukan Changxing Shang, dkk [9], akan mengalami kesulitan dan ketidakefektifan jika seluruh kata/term langsung digunakan sebagai dokumen vektor. Jadi untuk mengurangi biaya komputasi, memilih fitur yang penting yang nantinya akan mewakili karakter dari semua dokumen. Dalam pemilihan fitur menggunakan

Information Gain (IG), untuk memilih fitur-fitur yang terbaik dari hasil ekstrasi. Information Gain salah satu teknik seleksi fitur yang menggunakan metode scoring untuk nominal ataupun pembobotan atribut kontinu yang menggunakan maksimal entropy. Entropy merupakan definisi dari nilai Information Gain. Banyaknya informasi digambarkan dengan entropy yang dibutuhkan untuk menafsirkan suatu kelas [10]. Information Gain (IG) dari term didapatkan dengan menghitung jumlah bit informasi yang diambil dari prediksi kelas dengan ada atau tidaknya term dalam dokumen.

Pada tahap ini dilakukan pemilihan fitur terbaik dari data mentah yang sudah diekstrasi menjadi fitur atau term. Sangat tidak efektif jika seluruh kata atau termlangsung digunakan sebagai vektor dokumen. Fitur yang tidak relavan akan dihapus atau tidak digunakan agar mempercepat proses komputasi, karena fitur yang terpilih tersebut nantinya akan menjadi input pada Deep Neural Network (DNN). Melalui penerapan tahap pemilihan fitur yang sangat penting untuk membentuk sekumpulan kata untuk mewakili karakter dari suatu kelas. Fitur yang dipilih adalah fitur dengan nilai Information Gain yang tidak sama dengan nol dan lebih besar dari suatu nilai threshold tertentu.

$$\mathrm{i}\, g(K,L) = Entropy(K) - \sum_{v \in value(L)} \frac{|K_v|}{K} Entropy(K_v)$$

$$Entropy(K) = -\sum \frac{|K_i|}{K} \log \frac{K_i}{K}$$

Dimana K adalah jumlah seluruh fitur, L adalah kategori, Kv adalah jumlah sampel untuk nilai v, v adalah nilai yang mungkin untuk kategori L, Ki adalah fitur ke i dan Value(L) adalah himpunan nilai-nilai yang mungkin untuk kategori L.

3.5. Klasifikasi menggunakan Deep Neural Network

Berdasarkan penelitian yang dilakukan oleh Hosen [11], Pembelajaran mendalam (deep learning) adalah suatu pembelajaran mesin dengan banyak tingkat representasi, fitur yang lebih tinggi ditentukan dari tingkat fitur yang rendah dan sebaliknya fitur yang sama dapat mendefinisikan banyak fitur yang tinggi. Deep learning berkembang dari jaringan saraf(NN) dengan memperbanyak hidden layer pada arsitektur jaringan salah satunya adalah Deep Neural Network. Output dari satu lapisan berfungsi sebagai input ke lapisan berikutnya, dengan pembatasan pada semua jenis loop dalam arsitektur jaringan [12].

Pada tahap ini menggunakan metode deep learning yaitu deep neural network untuk mengklasifikasikan data yang diproleh dari tahap-tahap sebelumnya. Fitur yang dibentuk sebagai representasi fitur Bagof-words (BoW) sebagai masukan dari beberapa fitur yang sudah diproses menggunakan teknik prepcessing terhadapnya, diekstrasi dengan fitur ekstrasi TF-IDF dan fitur-fitur tersebut dipilih menggunakan Information Gain. Dari fitur yang dimasukkan, kemudian dilakukan algoritma forward propagation dengan memasukkan semua fitur yang sudah dipilih oleh Information Gain. Bertahap dari lapisan input dilanjutkan ke lapisan tersembunyi sampai lapisan output. Dari hasil output, dilakukan perhitungan menggunakan fungsi kesalahan untuk mengukur tingkat kesalahan yang terjadi antara proses fitur yang dimasukkan melalui setiap lapisan input hingga ke lapisan output yang dihasilkan. Penggunaan fungsi kesalahan, yang juga didasarkan pada label yang sudah ditentukan adalah relavant dan non relavant, maka dari hasil kesalahan yang diperoleh diterapkan algoritma back propagation untuk mencapai kesalahan minimum hingga ditemukan konvergensi nilai fitur.

3.5. Evaluasi dan Validasi

Setelah semua proses dilakukan, yaitu dari proses pre-processing hingga proses klasifikasi menggunakan Deep Neural Network, selanjutnya dilakukan evaluasi dan validasi dari hasil kinerja klasifikasi. Pada penelitian ini, evaluasi untuk mengetahui performa dari hasil kinerja klasifikasi menggunakan confusion matrix pada tabel 3.1 dan validasi menggunakan 10-fold cross validation. Pada saat melakukan validasi, urutan dari kumpulan dokumen yang ada akan diacak. Hal ini bertujuan untuk menghindari adanya pengelompokkan dokumen yang berasal dari kategori tertentu. Dan juga waktu (t) yang dibutuhkan untuk proses komputasi antara menggunakan fitur seleksi dan tidak yaitu dengan Information Gain.

Tabel 3.1 Cofution matrix

	yes	**no**	**precision**
yes	True positive	False Positive	Precision yes
no	False Negative	True Negative	Precision no
recall	Recall yes	Recall no	

$$\boldsymbol{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4. HASIL DAN PEMBAHASAN

Ekperimen yang dilakukan adalah membandingkan metode yang diusulkan dengan metode-metode yang lain seperti Naïve bayes (NB), Rendom forest(RF), Support vector machine(SVM), Decision Tree (DT) dan metode yang diusulkan yaitu Deep Neural Network (DNN) dengan Information gain(IG).

Tabel 4.1 Hasil komparasi dengan metode yang lain

Method	**Precision**	**Recall**	**Accuracy**
NB	79.60%	58.82%	66.08%
RF	62.53%	100.00%	63.86%
SVM	70.56%	97.79%	74.06%
DT	71.98%	96.32%	75.17%
DNN+IG	**81.21%**	**93.75%**	**83.15%**

Pada tabel 4.1 menunjukkan hasil komparasi dari bebearapa metode dengan metode Deep Neural Network dan Information Gain. Naïve bayes mencapai nilai akurasi sebesar 66.08% dan random forest 63.86%, sedangkan hasil akurasi dari Support Vector Machine 74.06% dan decision tree 75.17% lebih besar dari hasil metode Naive bayes. Hasil untuk Deep Neural Network dengan Information gain sebesar 83.15%.

5. KESIMPULAN

Hasil akurasi yang didapatkan untuk metode Naive Bayes lebih rendah dari metode yang lain, dengan hasil akurasi sebesar 66.08%. Pada hasil akurasi metode Deep Neural Network dengan Information Gain (DNN+IG) lebih tinggi dari metode yang lain yaitu sebesar 83.15%, karena adanya proses seleksi fitur yang dilakukan sebelum diklasifikasikan dengan metode Deep Neural Network.

DAFTAR PUSTAKA

[1] J. Singh, G. Singh, R. Singh, and P. Singh, "Morphological evaluation and sentiment analysis of Punjabi text using deep learning classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 5, pp. 508–517, 2021, doi: 10.1016/j.jksuci.2018.04.003.

[2] A. Squicciarini, A. Tapia, and S. Stehle, "Sentiment analysis during Hurricane Sandy in emergency response," *Int. J. Disaster Risk Reduct.*, vol. 21, no. December 2016, pp. 213–222, 2017, doi: 10.1016/j.ijdrr.2016.12.011.

[3] S. Xiong, H. Lv, W. Zhao, and D. Ji, "Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings," *Neurocomputing*, vol. 275, pp. 2459–2466, 2018, doi: 10.1016/j.neucom.2017.11.023.

[4] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.

[5] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," *Procedia Computer Science*, vol. 127. pp. 511–520, 2018, doi: 10.1016/j.procs.2018.01.150.

[6] A. Luthfiarta, J. Zeniarja, and A. Salam, "Algoritma Latent Semantic Analysis (LSA) Pada Peringkas Dokumen Otomatis Untuk Proses Clustering Dokumen," *Semin. Nas. Teknol. Inf. Komun. Terap. 2013 (SEMANTIK 2013)*, vol. 2013, no. November, pp. 13–18, 2013.

[7] G. Karyono, F. S. Utomo, A. Sistem, and T. Balik, "Temu Balik Informasi Pada Dokumen Teks Berbahasa Indonesia Dengan Metode Vector Space Retrieval Model," *Semin. Nas. Teknol. Inf. dan Terap. 2012*, vol. 2012, no. Semantik, pp. 282–289, 2012.

[8] M. Abdel Fattah, "New term weighting schemes with combination of multiple classifiers for sentiment analysis," *Neurocomputing*, vol. 167, pp. 434–442, 2015, doi: 10.1016/j.neucom.2015.04.051.

[9] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," *Knowledge-Based Syst.*, vol. 54, pp. 298–309, 2013, doi: 10.1016/j.knosys.2013.09.019.

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

[11] H. Mohsen, E.-S. A. El-Dahshan, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Classification using deep learning neural networks for brain tumors," *Futur. Comput. Informatics J.*, vol. 3, no. 1, pp. 68–71, 2018, doi: 10.1016/j.fcij.2017.12.001.