Klasifikasi dalam Pembuatan Portal Berita Online dengan Menggunakan Metode BERT

Jehezkiel Hardwin Tandijaya
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121 – 131
Surabaya 60236, Indonesia
Telp. (031) – 2983455
jehezkiel.ht16@gmail.com

Liliana
Program Studi Informatika
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121 – 131
Surabaya 60236, Indonesia
Telp. (031) – 2983455
lilian@petra.ac.id

Indar Sugiarto
Program Studi Teknik Elektro
Fakultas Teknologi Industri
Universitas Kristen Petra
Jl. Siwalankerto 121 – 131
Surabaya 60236, Indonesia
Telp. (031) – 2983455
indi@petra.ac.id

ABSTRAK

Teknologi internet membantu manusia dengan memberikan kemudahan untuk mendapatkan beragam informasi yang diinginkan yang dapat diakses dari berbagai *platform* berita *online*. Namun, berita yang tersedia sekarang sangat banyak karena dapat diakses di berbagai *platform* berita *online* dan perlu dikategorikan. Berita yang diakses di salah satu sumber juga tidak memiliki kredibilitas yang tinggi terhadap suatu kejadian, karena penerbit menggunakan informasi yang salah dan menyesatkan untuk memajukan kepentingan mereka atau yang dikenal dengan *fake news*. Sehingga untuk mengecek kredibilitas dari suatu kejadian, dibutuhkan membaca dari sumber berita lain tidak hanya di satu tempat saja agar tidak terpengaruh dengan berita yang salah. Hal tersebut kurang efektif karena pembaca harus mencari sumber berita lain dengan alamat *URL* yang berbeda.

Pada penelitian ini akan dilakukan *scraping* untuk mengambil berita yang ada pada sebuah portal berita. Setelah melakukan proses *scraping*¸ berita tersebut akan diklasifikasikan untuk menentukan kategori dari sebuah berita. Metode yang digunakan adalah *Bidirectional Encoder Representations from Transformers.*

Berdasarkan hasil pengujian yang telah dilakukan, berita dapat diambil dan diklasifikasikan. Pengujian dengan *pretrained model* indobenchmark/indobert-base-p1 mendapatkan hasil yang baik dimana akurasi dapat mencapai 87.548%.

Kata Kunci: portal berita, *web scraping*, *text classification*, *Bidirectional Encoder Representations from Transformers.*

ABSTRACT

Internet helps human by making various information from many online news platform accessible. But nowadays, there are a lot of news that can be accessed in different online news platform and needs to be categorized. The news that can be accessed in some of the sources don't have high credibility about an event, because the publishers use false and misleading information to push their agendas. So in order to check the credibility of an event, it is needed to also read from other sources and not only from 1 source. However, this is not effective because the reader has to look for another news source with different URL address.

In this research scraping will be done to retrieve the news that are available in a news platform. After the scraping process is done, the news will be classified to determine the category of the news. The method that will be used is Bidirectional Encoder Representations from Transformers.

From the testing of this research, the news can be retrieved and classified. The testing with a pre-trained model indobenchmark /indobert-base-p1 get a very good result where the accuracy reaches 87.548%.

Keywords: *news portal, web scraping, text classification, Bidirectional Encoder Representations from Transformers.*

1. PENDAHULUAN

Informasi merupakan salah satu kebutuhan bagi manusia.Berita merupakan salah satu satu sumber informasi mengenai kejadian terkini yang ada pada media massa seperti surat kabar, televisi, dan media lainnya [10]. Seiring berkembangnya teknologi internet menyebabkan kemudahan untuk mendapatkan suatu informasi karena dapat diakses oleh siapa saja. Teknologi internet juga membantu manusia dengan memberikan kemudahan untuk mendapatkan beragam informasi yang diinginkan yang dapat diakses dari berbagai *platform* berita online. *Platform* berita online membantu mengatasi batasan temporal dan spasial dari surat kabar cetak tradisional [11] dan memudahkan kita untuk mencari berbagai informasi dari dalam maupun luar negeri yang diinginkan. Namun, berita yang tersedia sekarang sangat banyak di berbagai platform berita online dengan alamat URL yang berbeda-beda [4]. Banyaknya berita yang tersedia perlu diorganisir / dikategorikan yang dapat memudahkan dalam hal akses [16]. Belum juga berita yang diakses di salah satu sumber memiliki kredibilitas yang tinggi terhadap suatu kejadian, karena penerbit menggunakan informasi yang salah dan menyesatkan untuk memajukan kepentingan mereka [1] atau yang dikenal dengan *fake news*. *Fake news* sangat banyak di dunia digital ini, bahkan beberapa pejabat dan individu terlibat dalam penyebarannya agar sesuai dengan tujuan mereka [3]. Sehingga untuk mengecek kredibilitas dari suatu kejadian, dibutuhkan membaca dari sumber berita lain tidak hanya di satu tempat saja agar tidak terpengaruh dengan berita yang salah. Namun, hal tersebut kurang efektif karena pembaca harus mencari sumber berita lain dengan alamat URL yang berbeda. Oleh karena itu, dibutuhkan suatu *website* yang berisi kumpulan berita dari berbagai sumber berita online dan dikategorikan agar memudahkan pengguna dalam hal akses.

Bidirectional Encoder Representations from Transformers merupakan teknik pembelajaran mesin yang dikembangkan oleh

Google berbasis *Transformer* untuk pra-pelatihan pemrosesan Bahasa alami (*Natural Language Processing*). Adapun penelitian – penelitian sebelumnya *Contextual Semantic Embeddings based on Fine-tuned AraBERT Model for Arabic Text Multi-class Categorization* dimana membandingkan AraBERT *pre-trained model* dan melakukan *fine-tunes* dengan AraBERT sebagai *feature extractor* model dalam hubungannya dengan beberapa neural network classifier dan SVM untuk melakukan klasifikasi [13]. *CyberBERT: BERT for cyberbullying identification* dimana model BERT menghasilkan performa lebih baik dari model klasifikasi baik *machine learning* maupun *deep learning* dalam melakukan klasifikasi *cyberbullying* [14]. *News Classification Using Naïve Bayes and Two-Phase Feature Selection Model* dimana data didapat di kompas.com dan peneliti melakukan klasifikasi dengan naïve bayes dan 2 fase *feature selection* dimana akurasi tertinggi didapat sebesar 86% [2]. *Efficient classification model of web news documents using machine learning algorithms for accurate information* dimana penulis membandingkan *K-Nearest Neighbor*(kNN), *Support Vector Machine* (SVM), *Decision Tree* (DT), dan *Long Short-Term Memory* (LSTM) dan SVM menghasilkan akurasi paling tinggi sebesar 95.04% dan akurasi paling rendah didapat KNN sebesar 88.72% [12]. Klasifikasi Berita Olahraga Menggunakan Metode *Naïve Bayes* Dengan *Enhanced Confix Stripping Stemmer* dimana berita dapat diklasifikasi dengan menggunakan metode *naïve bayes* dengan *confix stripping stemmer* menghasilkan akurasi sebesar 77% [16].

Adapun kelebihan dan kekurangan metode-metode yang digunakan sebelumnya. *K-Nearest Neighbor* (KNN) mahal secara komputasi saat diterapkan dalam data berdimensi tinggi terutama jika set pelatihannya banyak, rentan terhadap *overfit* dan kinerja relatif rendah saat dihadapkan oleh teks panjang dan memiliki banyak fitur [8]. Namun, kNN lebih sering digunakan daripada metode lain untuk menangani teks yang pendek. *Naïve Bayes* (NB) bekerja dengan baik ketika fitur saling bergantung, model yang dihasilkan mudah untuk direpresentasikan dan dijelaskan, dan direkomendasikan untuk ukuran sampel yang lebih kecil karena regularisasi yang melekat, membuatnya cenderung tidak terlalu pas dibandingkan dengan pengklasifikasi diskriminatif. *Support Vector Machine* (SVM) dapat menangani dengan baik data berdimensi tinggi, tidak rentan terhadap *overfitting,* tetapi representasi SVM yang terbatas dapat mengakibatkan kurangnya kemampuan untuk memodelkan pola bernuansa dalam data pelatihan. Oleh karena itu, pada skripsi metode *deep learning* yang digunakan adalah *Bidirectional Encoder Representations from Transformers* karena metode klasifikasi dengan menggunakan pendekatan tradisional machine learning memiliki batasan tertentu dalam pelatihan dataset skala besar walaupun memiliki karakteristik efisiensi dan stabilitas tinggi [17].

Untuk menjawab permasalahan diatas, dibutuhkan *website* yang berisi kumpulan berita dari berbagai sumber berita online dengan fitur dasar dan mengkategorikan berita tersebut agar memudahkan pembaca untuk membaca dari satu *website* dengan berita-berita yang disediakan dan dikategorikan. Metode yang akan digunakan untuk mengkategorikan berita adalah *Bidirectional Encoder Representations from Transformers*.

2. LANDASAN TEORI

2.1 *Preprocessing*

Preprocessing memiliki peranan yang sangat penting dalam aplikasi *text mining* [9]. *Preprocessing* dapat membersihkan data dari *noise* seperti kesalahan ejaan, pengurangan karakter yang direplikasi dan disambiguasi akronim yang ambigu) [7]. Pada umumnya, terdapat beberapa proses *preprocessing* seperti *case folding, tokenization, stopwords removal dan stemming* [18]. *Case Folding* mengubah semua text menjadi huruf kecil. *Tokenization* merupakan proses dimana teks dipecah menjadi sebuah kata, frasa, simbol atau elemen lainnya yang disebut *token*. *Stopword Removal* merupakan proses dimana kata – kata yang umum dan sering muncul dihilangkan seperti kata penghubung, dll) atau kata-kata yang tidak penting yang didefinisikan oleh pembuat program. *Stemming* adalah salah satu fase dalam pra-pemrosesan teks yang menerapkan bahasa alami untuk menghilangkan imbuhan dari kata untuk mengubahnya menjadi kata dasar [19]. Contoh kata "mengunjungi" menjadi kata "kunjung". *Filtering* adalah melakukan pembersihan pada *string* seperti menghapus tanda baca, angka, *special character* dan sebagainya.

2.2 *Bidirectional Encoder Representations from Transformers* (BERT)

Bidirectional Encoder Representations from Transformers merupakan model representasi kata kontekstual yang dilatih sebelumnya berdasarkan MLM (Masked Language Model), menggunakan *Transformers* dua arah [15]. Arsitektur model BERT adalah struktur *encoder-decorder* transformator dua arah *multi-layer* [6]. *Transformer* mengikuti keseluruhan arsitektur ini menggunakan *stacked self-attention* dan *point-wise*, terhubung sepenuhnya untuk *encoder* dan *decoder* [20].

Ada 2 langkah dalam kinerja framework BERT yaitu *pre-training* dan *fine-tuning. Pre-training* BERT tidak menggunakan cara tradisional *left-to-right* atau *right-to-left*, tetapi menggunakan *Masked Language Modeling* (MLM) dan *Next Sentence Prediction* (NSP) untuk *pre-training* data [5]. MLM mengisi tempat kosong, dimana model menggunakan *context word* disekitar *mask token* untuk memprediksi kata yang seharusnya sedangkan NSP merupakan prediksi kalimat berikutnya dengan dua model yang diberikan. Setelah melakukan *pre-training* data, BERT akan melakukan *fine-tuning* dimana *fine-tuning* diinisialisasi dengan parameter yang telah dilatih sebelumnya, dan semua parameter *fine-tuning* menggunakan data berlabel dari tugas-tugas *downstream*.

3. DESAIN SISTEM

3.1 Analisis Data

Pada penelitian ini, data yang akan digunakan adalah data dari portal berita berbahasa Indonesia yaitu kompas.com dan sindonews.com dengan menggunakan *web scraping* dengan bantuan library *BeautifulSoup*. Data yang diambil adalah judul berita, isi dan kategori berita dan data tersebut disimpan dengan format .csv. Kategori yang akan dianalisa pada penelitian ini adalah edukasi, tekno, *sports*, *health* dan *lifestyle*. Data keseluruhan yang digunakan berjumlah 6309 dan akan dibagi menjadi 2 tipe data yaitu data *training* dan data *testing* dengan perbandingan 80% data *training* dan 20% data *testing*. Data *training* ini digunakan untuk membuat *knowledge* baru dimana dalam penelitian ini adalah klasifikasi. Data *testing* ini digunakan untuk mengevaluasi model yang dibangun.

Gambar 1 merupakan data yang akan digunakan pada penelitian ini. Pada gambar tersebut memiliki 3 kolom, dimana kolom pertama berisi judul berita. Pada kolom kedua, berisi isi dari

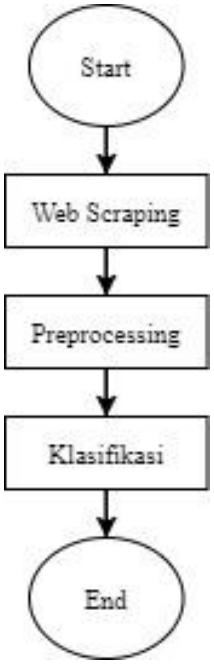
berita yang akan digunakan dalam membangun model. Kolom terakhir, berisi kategori dari berita tersebut.

	A	C	D	E	F	G	H
1	judul	isi	kategori				
2	Bantu Per	Pemasara	edukasi				
3	Mahasisw	Mahasisw	edukasi				
4	Asosiasi G	Presiden	edukasi				
5	Mau Belaj	Kementer	edukasi				
6	Inovasi Ini	Online	edukasi				
7	IPB Unive	Sebanyak	edukasi				
8	Ancaman	Federasi	edukasi				
9	Survei FSC	Federasi	edukasi				
10	Jangan Sa	Proses	edukasi				
11	Maksimal	Lembaga	edukasi				
12	Pandemi I	Pemerint	edukasi				
13	Siap-siap,	Senin	edukasi				
14	Ini Respor	Rektor	edukasi				
15	Tak Komp	Rencana	edukasi				
16	Pertamina	Selama	edukasi				
17	Catat! Ini	Tahap	edukasi				
18	Mahasisw	Tim	edukasi				
19	Rencana F	Federasi	edukasi				
20	DKI Perpa	Pemprov	edukasi				
21	Sekolah D	Sekolah	edukasi				
22	Sekolah S	Sekolah	edukasi				
23	Tim Maha	Tim	edukasi				
24	Mahasisw	Lembaga	edukasi				
25	IIK Bhakti	Institut	edukasi				

Gambar 1. Dataset yang digunakan

3.2 Analisis Sistem

Sistem mencakup pengambilan data dengan bantuan *Web Scraping*, mengolah data yang didapat dengan melalui tahapan *preprocessing* dan tahapan terakhir yang dilakukan adalah klasifikasi dengan menggunakan metode *Bidirectional Encoder Representations from Transformers* (BERT).

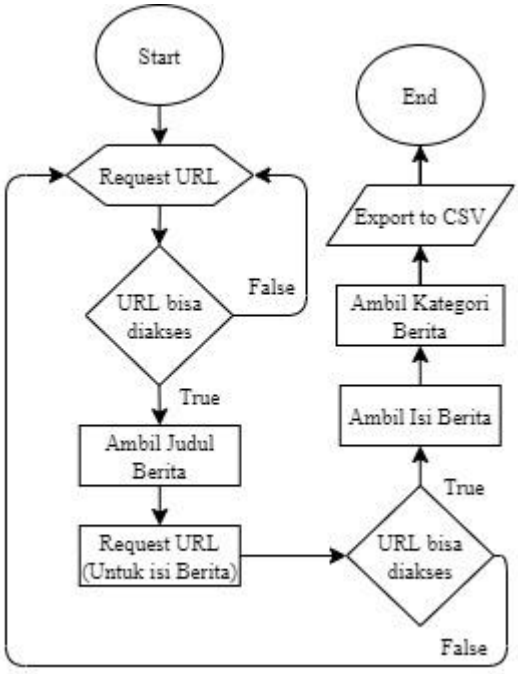


Gambar 2. Arsitektur Sistem

Gambar 2 merupakan arsitektur sistem yang akan dilakukan. Hal pertama yang dilakukan dalam penelitian ini adalah dengan mengambil data dari portal berita Bahasa Indonesia yaitu kompas.com dan sindonews.com dengan cara *web scraping* dengan bantuan *library BeautifulSoup*. Setelah *mendapatkan* data yang dibutuhkan, dilakukan *preprocessing* yang bertujuan untuk membersihkan data dari *noise*. Tahapan terakhir yang dilakukan adalah dengan melakukan klasifikasi dengan menggunakan metode *Bidirectional Encoder Representations from Transformers.*

3.2.1 *Web Scraping*

Pada penelitian ini, data yang akan diambil yaitu data dari portal berita berbahasa Indonesia yaitu kompas.com dan sindonews.com dengan menggunakan *web* scraping dengan bantuan library BeautifulSoup.



Gambar 3. Tahapan *Web Scraping*

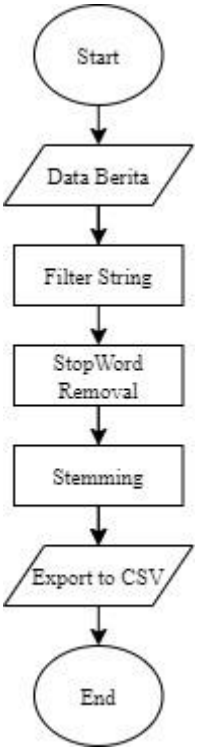
Gambar 3 merupakan tahapan *web scraping*. Pada proses *web scraping*, pertama-tama akan dilakukan *request* terhadap *website* yang akan digunakan. *Request* ini bertujuan untuk mengetahui apakah *website* tersebut diperbolehkan untuk dilakukan *scraping* atau tidak. Setelah mendapatkan respon code [200] dari *website* tersebut, langkah berikutnya adalah mengambil judul berita dari *website* tersebut. Setelah judul berita didapatkan, akan dilakukan *request url* kembali untuk mendapatkan isi berita karena isi berita berada pada url lain. Setelah semua data yang dibutuhkan didapatkan, langkah terakhir yang akan dilakukan adalah dengan mengexport ke format .csv.

3.2.2 *Preprocessing*

Setelah isi berita didapatkan, akan dilakukan tahapan *preprocessing* terhadap isi berita tersebut untuk membersihkan data dari *noise*. tahapan *preprocessing* meliputi *filter string* (menghilangkan simbol, angka, hanya menerima string), *stopword removal*, *stemming* (mengubah menjadi kata dasar seperti memakan -> makan).

Gambar 4 merupakan tahapan *Preprocessing*. Pada proses *preprocessing*, pertama-tama akan dilakukan *filter string* untuk

melakukan pembersihan pada *string* seperti menghapus tanda baca, angka, *special character* dan sebagainya. Tahapan ini juga mengubah semua huruf besar menjadi huruf kecil (*Lowercase text*). Setelah itu, akan dilakukan *stopword removal* yang bertujuan untuk menghapus kata-kata yang sering bermunculan, kata yang tidak memiliki makna yang diinisalisasi oleh pembuat program seperti kata hubung (dan, atau, dengan, dan sebagainya). Langkah terakhir, akan dilakukan *stemming* untuk menghilangkan imbuhan dan merubahnya menjadi kata dasar seperti kata "berkunjung" menjadi "kunjung". Setelah dilakukan *preprocessing*, data tersebut akan diubah kategorinya menjadi list agar memudahkan pada saat *training* dan diekspor menjadi file dengan ekstensi .csv dikarenakan saat dilakukan *preprocessing* waktu yang dibutuhkan sangat lama sehingga diubah ke csv agar memudahkan untuk melanjutkan ke tahapan berikutnya.

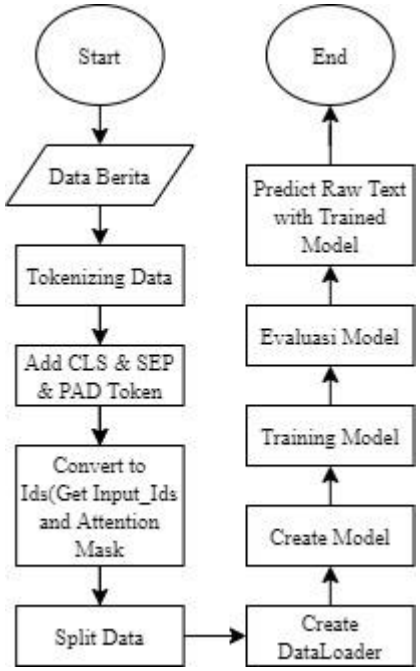


Gambar 4. Tahapan *Preprocessing*

3.2.3 *Klasifikasi*

Tahapan terakhir yang akan dilakukan adalah klasifikasi dengan metode *Bidirectional Encoder Representations from Transformers*.

Gambar 5 merupakan tahapan klasifikasi. Pada Tahapan klasifikasi, pertama input data yang didapat diubah kedalam bentuk input yang dapat dibaca oleh BERT. Untuk dapat dibaca oleh BERT, perlu ditambahkan token [CLS] di awal kalimat dan token [SEP] diakhir kalimat, dan menentukan panjang kalimat untuk menambahkan [PAD] token sebanyak sisa token. Setelah token [CLS], [SEP], dan [PAD] ditambahkan, BERT akan mengubah setiap token kata menjadi token ids dan mengembalikan hasil input_ids dan attention_mask untuk nantinya akan diteruskan kedalam model BERT. Setelah mengubah data menjadi input yang dapat dibaca oleh BERT, akan dilakukan pembuatan *dataloader* membantu mempercepat pengambilan data. Setelah membuat *dataloader*, tahapan selanjutnya adalah pembuatan model. Setelah model dibuat, akan dilakukan *training* dan evaluasi terhadap model yang telah dibuat. Model yang telah dilatih sebelumnya akan digunakan untuk melakukan klasifikasi terhadap berita baru.



Gambar 5. Tahapan Klasifikasi

4. HASIL PENGUJIAN

Pengujian ini menggunakan data keseluruhan hasil *web scraping* dengan jumlah 6309 data dan dibagi menjadi 80% *training* dan 20% *testing* dan menggunakan data baru berjumlah 1285. *Training* ini berguna untuk membangun model untuk dapat melakukan klasifikasi. Testing ini berguna untuk mengevaluasi model yang dibangun.

Tabel 1. Konfigurasi *hyperparameter*

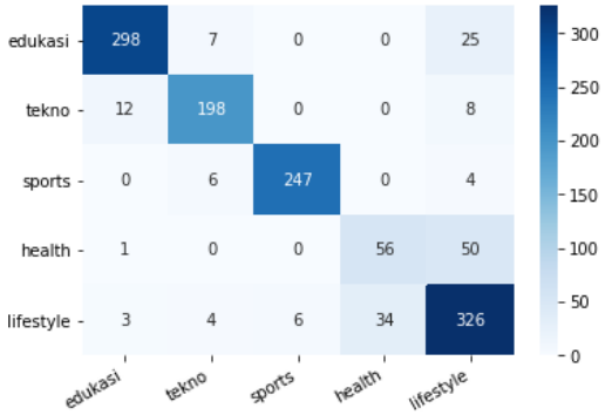
Model	Indobenchmark/bert-base-p1	Bert-base-multilingual-uncased
Dropout	0.1-0.5	
Learning rate	1e-5, 2e-5, 3e-5, 4e-5	
Batch size	16	
Epoch	4	

Tabel 2. Metrics pengukuran tiap kategori

Kategori	Precision	Recall	F1-Score
Edukasi	0.95	0.90	0.93
Tekno	0.92	0.91	0.91
Sports	0.98	0.96	0.97
Health	0.62	0.52	0.57
Lifestyle	0.79	0.87	0.83

Setelah melakukan *tuning parameter*, akurasi dapat mencapai 87.548% pada konfigurasi *learning rate* 5e-05, *dropout* 0.1.

Tabel 2 merupakan metrics pengukuran terhadap data baru. Dari tabel tersebut, dapat disimpulkan bahwa model yang dibangun masih belum bisa melakukan prediksi terhadap kategori *health* dengan baik. Hal ini dikarenakan data yang dipakai pada saat *training* sedikit dan pada situs sindonews.com tidak memiliki kategori *health*.



Gambar 6. *Confusion Matrix*

Gambar 6 merupakan *confusion matrix* dalam melakukan klasifikasi terhadap data baru. Dari gambar tersebut, dapat disimpulkan bahwa kategori *health* sering salah diprediksi menjadi kategori *lifestyle*. Begitu sebaliknya, hal ini disebabkan karena adanya kemiripan berita antara berita dengan kategori *health* dan *lifestyle*. Sebagai contoh, pada berita yang memiliki judul "12 Makanan yang Bisa Meredakan Stress", model yang dibangun tidak dapat mengerti bahwa berita tersebut berkategori *health* dan model akan memprediksi bahwa berita tersebut memiliki kategori *lifestyle*.

5. KESIMPULAN

Dari hasil pembuatan sistem yang dilakukan, dapat ditarik beberapa kesimpulan, antara lain:

- Setiap *website* yang digunakan pada saat melakukan *scraping* memiliki struktur yang berbeda sehingga *code* yang digunakan untuk mengambil isi dari sebuah *website* yang diinginkan berbeda.
- Kategori yang dipilih lebih baik menggunakan kategori yang terdapat pada semua *website* yang akan digunakan agar tidak kekurangan data tertentu.
- *Pretrained* model yang digunakan akan mempengaruhi performa dari model yang dibangun. Berdasarkan kombinasi *learning rate* dan *dropout* yang digunakan oleh penulis, dapat disimpulkan bahwa *pretrained* model indobenchmark/indobert-base-p1 memiliki performa yang lebih baik daripada *pretrained* model bert-base-multilingual-uncased dalam melakukan klasifikasi.
- Setiap kombinasi *hyperparameter* yang digunakan dapat mengaruhi kinerja model yang dibangun.
- Akurasi terbaik dalam melakukan klasifikasi dapat mencapai 87.548% dengan konfigurasi *learning rate* 5e-05, *dropout* 0.1, dan *epoch* 4.

6. DAFTAR PUSTAKA

[1] Aldwairi, M., & Alwahedi, A. 2018. Detecting fake news in social media networks. *Procedia Computer Science*, *141*, 215–222. https://doi.org/10.1016/j.procs.2018.10.171

[2] Ali Fauzi, M., Arifin, A. Z., Gosaria, S. C., & Prabowo, I. S. 2017. Indonesian news classification using naïve bayes and two-phase feature selection model. *Indonesian Journal of Electrical Engineering and Computer Science*, *8*(3), 610–615. https://doi.org/10.11591/ijeecs.v8.i3.pp610-615

[3] Apuke, O. D., & Omar, B. 2021. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, *56*(July), 101475. https://doi.org/10.1016/j.tele.2020.101475

[4] Aziz, A., & Rahmah, Y. 2017. Portal system for Indonesian online newspaper - Based feed parser simple pie. *Proceedings - 2016 International Seminar on Application of Technology for Information and Communication, ISEMANTIC 2016*, 169–173. https://doi.org/10.1109/ISEMANTIC.2016.7873832

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

[6] Fang, W., Luo, H., Xu, S., Love, P. E. D., Lu, Z., & Ye, C. 2020. Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*, *44*(March 2019), 101060. https://doi.org/10.1016/j.aei.2020.101060

[7] HaCohen-Kerner, Y., Miller, D., & Yigal, Y. 2020. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, *15*(5), 1–22. https://doi.org/10.1371/journal.pone.0232525

[8] Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

[9] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S., & Gurusamy, V. 2015. Preprocessing Techniques for Text Mining. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

[10] Kasanah, A. N., Muladi, M., & Pujianto, U. 2019. Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

[11] Kwak, K. T., Hong, S. C., & Lee, S. W. 2020. A study of repetitive news display and news consumption in Korea. *Telematics and Informatics*, *46*(October 2019), 101313. https://doi.org/10.1016/j.tele.2019.101313

[12] Mulahuwaish, A., Gyorick, K., Ghafoor, K. Z., Maghdid, H. S., & Rawat, D. B. 2020. Efficient classification model of web news documents using machine learning algorithms for

accurate information. *Computers and Security*, *98*. https://doi.org/10.1016/j.cose.2020.102006

[13] Ouatik, S., Alaoui, E., & Nahnahi, N. E. 2021. Contextual Semantic Embeddings based on Fine-tuned AraBERT Model for Arabic Text Multi-class Categorization. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2021.02.005

[14] Paul, S., & Saha, S. 2020. CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. *Multimedia Systems*, *0123456789*. https://doi.org/10.1007/s00530-020-00710-4

[15] Peng, Y., Yan, S., & Lu, Z. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *ArXiv*, *iv*. https://doi.org/10.18653/v1/w19-5006

[16] Pramudita, Y. D., Putro, S. S., & Makhmud, N. 2018. Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(3), 269. https://doi.org/10.25126/jtiik.201853810

[17] Sari, W. K., Rini, D. P., Malik, R. F., & Azhar, I. S. B. 2017. Klasifikasi Teks Multilabel pada Artikel Berita Menggunakan Long Short- Term Memory dengan Word2Vec. 1(10), 276–285.

[18] Sistem, R. 2021. Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia. *JURNAL RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *1*(10), 648–654.

[19] Utomo, F. S., Suryana, N., & Azmi, M. S. 2020. Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances. *IIUM Engineering Journal*, *21*(1), 33–50. https://doi.org/10.31436/iiumej.v21i1.1170

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.