KLASIFIKASI BIDANG USAHA MENGGUNAKAN CONVOLUTIONAL NEURAL NETWORK

(*Business Fields Classification using Convolutional Neural Network***)**

Zikry Ramadhan[1], Ramaditya Dwiyansaputra[1],I Gede Pasek Suta Wijaya[1]
[1]Dept Informatics Engineering, Mataram University
Jl. Majapahit 62, Mataram, Lombok NTB, INDONESIA
Email: zikry399b@gmail.com, [rama, gpsutawijaya]@unram.ac.id

Abstract

Business Fields are all forms of business activities carried out to produce goods or services in economic sectors. These fields are located all over the world, with some fields are more abundant than others in certain places. One of the factors to consider when someone plans to open a business is nearby competitors. However, searching for nearby business or competitors can take long time and tedious to do. To approach this issue, we propose this research so people can get a business field from certain business by classifying text available on its website or social media promotion. We hope by using Convolutional Neural Network (CNN) to get the most significant words for certain field, we can get good classification results. We found that using a simple CNN can produce good enough results and saw there can be potential for more complex CNN.

Keywords:*Text Classification, Convolutional Neural Network, Business Fields, Business*

1. PENDAHULUAN

Saat ini perekonomian dunia sedang dalam situasi yang sulit. Mulai dari isu resesi global, perang Rusia – Ukraina, PHK massal yang dilakukan oleh perusahaan – perusahaan kelas dunia seperti Google, Microsoft, dan IBM [1]. Indonesia juga tidak terlepas dari situasi ini. Banyak perusahaan Indonesia yang mengikuti tren PHK massal tersebut, seperti Shopee dan GoTo. Tidak hanya itu, dikabarkan akan ada PHK massal pada sembilan perusahaan lainnya dalam waktu dekat [2]. PHK massal yang terjadi tentunya akan berpengaruh kepada perekonomian Indonesia dan meningkatkan angka pengangguran Indonesia.

Salah satu cara untuk menanggulangi jumlah pengangguran yang meningkat akibat PHK massal tersebut adalah dengan meningkatkan lapangan kerja. Dengan kondisi perusahaan – perusahaan besar saat ini, kecil kemungkinan untuk dibukanya lowongan pekerjaan sehingga yang perlu dilakukan untuk menungkatkan lapangan kerja adalah membuat usaha baru. Akan tetapi, membuat suatu usaha baru bukanlah hal yang mudah. Ada banyak hal yang harus dipertimbangkan dalam riset pasar seperti modal, lokasi, minat dari masyarakat, kompetitor, dan sejenisnya.

Salah satu informasi yang dapat diambil dari sebuah perusahaan kompetitor adalah bidang usaha yang dijalankan. Dengan hal mengetahui itu, seseorang dapat mengetahui usaha yang masih sedikit kompetitornya atau bidang usaha yang populer di suatu tempat. Cara mengetahui bidang usaha tersebut adalah dengan mencari data–data relevan kemudian menarik kesimpulan dari data yang ditemukan. Data-data yang dapat ditemukan umumnya berupa teks yang dapat ditemukan pada situs perusahaan atau akun media sosial. Data lainnya dapat berupa foto atau video namun jenis ini berjumlah lebih sedikit karena membutuhkan lebih banyak waktu untuk membuatnya. Mencari informasi mengenai hal–hal tersebut membutuhkan waktu yang cukup lama terlebih lagi apabila dilakukan secara manual. Oleh karena itu, dibutuhkan suatu alat yang dapat membantu seseorang dalam mengklasifikasikan bidang usaha dari bentuk teks sehingga dapat lebih cepat dalam membuat keputusan.

Salah satu metode yang dapat digunakan untuk melakukan klasifikasi teks adalah dengan menggunakan teknologi *deep learning*. Teknologi *deep learning* merupakan sebuah cabang dari *machine learning* yang menggunakan jaringan syaraf tiruan dengan banyak lapisan (*deep neural networks*) untuk menghasilkan model yang lebih kompleks dan akurat. Dari teknologi deep learning itu sendiri, salah satu metode atau algoritma yang cukup populer untuk

digunakan adalah dengan menggunakan CNN (*Convolutional Neural Network*).

CNN digunakan karena kemampuannya dalam mengenali fitur – fitur penting yang akan sangat berguna dalam klasifikasi. Keunggulan lainnya yaitu dapat memberikan hasil yang baik walaupun dengan arsitektur sederhana saja [3]. CNN juga memiliki kecepatan pembelajaran yang lebih cepat jika dibandingkan dengan metode lainnya seperti RNN, LSTM, dan Bi-LSTM [4].

Berdasarkan paparan tersebut, penulis mengajukan penelitian untuk merancang sebuah model machine learning untuk melakukan Klasifikasi Bidang Perusahaan Menggunakan CNN. Diharapkan penelitian ini dapat membantu masyarakat untuk mendapatkan kesimpulan dalah membuat suatu usaha ataupun menjadi referensi penelitian selanjutnya.

2. TINJAUAN PUSTAKA

Penelitian yang dilakukan Yoon Kim menunjukkan bahwa CNN yang sederhana dapat memberikan hasil yang baik dan bahkan kompetitif dengan beberapa metode lainnya [3]. Ada bermacam-macam model yang digunakan dalam penelitian dan didapatkan hasil yang paling baik adalah model yang menggunakan *pre-trained word vector* untuk kata-katanya.

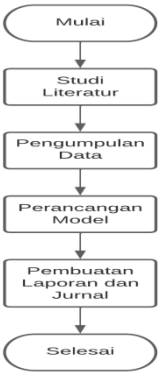
Penelitian [5] melakukan percobaan untuk klasifikasi perusahaan hanya menggunakan *pre-trained transformer model* tanpa melatih terhadap data tertentu. Penelitian dilakukan menggunakan dataset yang didapatkan dari Wharton Research Data Services (WRDS) dan dikategorikan berdasarkan Global Industry Classification Standard (GICS). Didapatkan hasil yang menunjukkan terdapat potensi untuk melakukan klasifikasi secara otomatis.

Penelitian [6] membandingkan beberapa pendekatan *machine learning* untuk klasifikasi industri berdasarkan teks deskripsi suatu perusahaan. Dataset yang digunakan berjumlah sekitar 300.000 data terdiri dari artikel ensiklopedis mengenai berbagai perusahaan dan kegiatan ekonomi mereka yang diklasifikasikan berdasarkan Wikipedia. Didapatkan hasil bahwa semua pengujian memberikan hasil yang relatif dekat tanpa ada yang memiliki hasil jauh lebih baik dibanding lainnya.

Penelitian [7] merupakan penelitian yang serupa dan dataset yang sama dengan [6] namun menggunakan beberapa pendekatan yang berbeda. Hasil yang didapatkan juga serupa namun untuk model yang menggunakan XLNet memiliki hasil yang sebagian besar lebih baik.

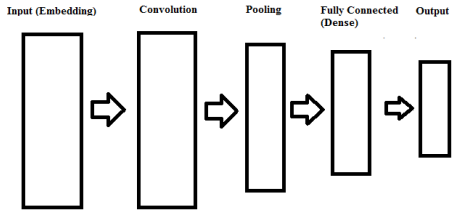
3. METODE PENELITIAN

Alur pelaksanaan penelitian dari awal hingga akhir dapat dilihat pada gambar sebagai berikut:



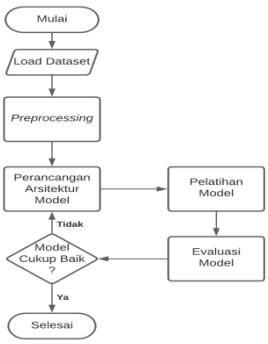
Gambar 1. Alur penelitian

Dataset yang digunakan dalam penelitian adalah Company Classification yang tersedia pada Kaggle. Dataset didapatkan dari *scraping* berbagai *website* dan terbagi menjadi 13 kategori. Model dirancang dengan arsitektur CNN sederhana seperti pada gambar berikut:



Gambar 2. Rancangan struktur CNN

Pembangunan model dilakukan berulang kali hingga model dinilai sudah cukup untuk digunakan dalam pengujian dengan alurnya sebagai berikut:



Gambar 3. Alur pengembangan model

Terdapat tiga skenario pengujian yaitu pengaruh penggunaan *stemming*, pengaruh dimensi *word embedding*, dan pengaruh tipe *pooling.*

4. HASIL DAN PEMBAHASAN

4.1 Pengujian Stemming

Pada pengujian stemming diharapkan model yang menggunakan stemming memiliki performa yang lebih baik karena banyak kata dalam berbagai bentuk yang bermakna sama dengan kata dasarnya.

Konfigurasi	*Precision*	*Recall*	*Accuracy*
Tanpa stemming	76.03%	74.20%	75.89%
Dengan stemming	75.76%	73.71%	75.55%

Dari hasil pengujian didapatkan model yang tidak menggunakan stemming memiliki performa yang sedikit lebih baik. Hasil tersebut serupa dengan penelitian yang dilakukan oleh Bhustomy Hakim yang mendapatkan performa yang lebih baik saat tidak menggunakan stemming.

4.2 Pengujian Dimensi Word Embedding

Pengujian dimensi word embedding dilakukan dikarenakan dimensi yang lebih besar diharapkan lebih mampu dalam mengenali keterkaitan tiap kata.

Konfigurasi	*Precision*	*Recall*	*Accuracy*
Dimensi = 2	67.22%	64.43%	68.26%
Dimensi = 4	75.69%	73.55%	75.11%
Dimensi = 5	76.03%	74.20%	75.89%
Dimensi = 6	75.13%	73.80%	75.53%
Dimensi = 10	73.79%	72.76%	74.50%
Dimensi = 20	73.23%	71.99%	73.63%

Dari pengujian dimensi word embedding terlihat performa model perlahan naik hingga dimensi sebesar 5 dan menurun setelah semakin ditambah, yang menunjukkan bahwa ukuran dimensi yang besar tidak menjamin performa yang maksimal.

4.3 Pengujian Tipe Pooling

Pengujian tipe pooling dilakukan karena average pooling dan max pooling dapat memberikan hasil yang berbeda.

Konfigurasi	*Precision*	*Recall*	*Accuracy*
max pooling	75.18%	73.72%	75.38%
average pooling	76.03%	74.20%	75.89%

Dari pengujian terlihat bahwa average pooling memiliki performa yang sedikit lebih baik. Hal ini dapat disebabkan oleh cara kerja max pooling yang mengambil nilai maksimum sehingga dapat menghilangkan fitur penting yang bernilai kecil atau negatif.

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian yang dilakukan, dapat diambil beberapa kesimpulan seperti model CNN sederhana dapat dinilai sebagai cukup baik dalam melakukan klasifikasi bidang usaha, pengujian dimensi word embedding merupakan pengujian yang memberikan perbedaan performa paling signifikan, dan tidak menutup kemungkinan CNN yang lebih kompleks dapat memberikan hasil yang lebih baik.

Saran yang dapat diberikan untuk penelitian selanjutnya dapat mencoba arsiterktur yang lebih kompleks sehingga memaksimalkan performa model.

DAFTAR PUSTAKA

[1] K. Alfonseca and M. Zahn, “Tech layoffs 2023: Companies that have made cuts,” ABC News, https://abcnews.go.com/Business/tech-layoffs-2023-companies-made-cuts/story?id=96564792 (accessed Jun. 13, 2023).

[2] D. C. Emeria, “Peringatan Bahaya, 9 Pabrik SIAP-SIAP PHK Massal,” CNBC Indonesia, https://www.cnbcindonesia.com/news/20230605082404-4-442904/peringatan-bahaya-9-pabrik-siap-siap-phk-massal (accessed Jun. 13, 2023).

[3] Y. Kim, “Convolutional neural networks for sentence classification,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Sep. 2014. doi:10.3115/v1/d14-1181

[4] H. Lu, L. Ehwerhemuepha, and C. Rakovski, “A comparative study on Deep Learning models for text classification of unstructured medical notes with various levels of class imbalance,” *BMC Medical Research Methodology*, vol. 22, no. 1, Jul. 2022. doi:10.1186/s12874-022-01665-y

[5] M. Rizinski et al., “Company classification using zero-shot learning,” arXiv.org, https://doi.org/10.48550/arXiv.2305.01028 (accessed Nov. 10, 2023).

[6] A. Tagarev, N. Tulechki, and S. Boytcheva, “Comparison of machine learning approaches for Industry Classification based on textual descriptions of companies,” ACL Anthology, https://aclanthology.org/R19-1134/ (accessed Nov. 10, 2023).

[7] S. Slavov, A. Tagarev, N. Tulechki and S. Boytcheva, "Company Industry Classification with Neural and Attention-Based Learning Models," 2019 Big Data,

Knowledge and Control Systems Engineering (BdKCSE), Sofia, Bulgaria, 2019, pp. 1-7, doi: 10.1109/BdKCSE48644.2019.9010667.