**PROPOSAL TESIS**

**MACHINE LEARNING MODELING UNTUK KLASIFIKASI BAHASA PADA TEKS**

Program Magister (S2) Ilmu Komputer

Universitas Sumatera Utara

**DEVI HAWANA LUBIS**

**NIM : 177038060**

**Logo

Description automatically generated**

**PROGRAM STUDI MAGISTER ILMU KOMPUTER**

**FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI**

**UNIVERSITAS SUMATERA UTARA**

**2021**

**BAB 1**

**PENDAHULUAN**

* 1. **Latar Belakang**

Bahasa memiliki peranan penting dalam komunikasi. Terdapat beragam bahasa di dunia yang digunakan dalam komunikasi sehari-hari. Setiap individu mampu menguasai lebih dari satu bahasa (Liu et al., 2019). Komunikasi dapat berjalan dengan adanya interaksi antara dua orang atau lebih melalui proses tatap muka dan juga melalui teks. Khususnya untuk komunikasi melalui teks terdapat banyak aplikasi chat, forum dan media sosial yang dapat digunakan. Contohnya twitter, mampu mendapatkan data dalam berbagai bahasa dengan jumlah yang melebihi 6000 *tweet* setiap detik (Shrivastava & Kumar, 2021). Dengan mempelajari teks berbagai informasi dapat diketahui, salah satunya kepribadian penulis serta klasifikasi dokumen (Ishihara, 2021; Shrivastava & Kumar, 2021; Yan et al., 2020). Dalam permasalahan klasifikasi teks, penelitian mengenai analisis sentimen dan *spam filtering* menjadi topik dengan jumlah peminat yang banyak. Klasifikasi merupakan bagian dari permasalah *machine learning*, dimana *input* data di-*train* sehingga menghasilkan *label* data kemudian disimpan dalam bentuk model untuk dapat digunakan kembali. Semakin banyak data yang digunakan dalam proses *training*, semakin baik performa model tersebut (Mouriño García et al., 2017).

Pada orang-orang yang mampu menguasai lebih dari satu bahasa terdapat kendala saat beralih dari satu bahasa ke bahasa lainnya. Kendala ini terjadi pada otak manusia, dimana terdapat momen *switching* atau pengalihan bahasa. Hal ini tidak hanya mempengaruhi pembicara tapi juga pendengar saat melalukan komunikasi (Liu et al., 2019). Di sisi lain, perkembangan teknologi yang pesat memungkinkan penyelesaian masalah ini untuk komunikasi yang dilakukan melalui teks. Metode klasifikasi pada *machine learning* dapat menjadi solusi permasahan tersebut (Mouriño García et al., 2017). Dalam proses *machine learning*, data menjadi *resource* utama yang harus tersedia. Dengan adanya data, algoritma *machine learning* dapat berjalan dengan baik untuk dapat menemukan informasi yang tidak diketahui sebelumnya. Oleh sebab itu, dalam penelitian ini akan dibuat model untuk klasifikasi bahasa pada teks. Model tersebut diharapkan dapat mempermudah penyelesaian masalah peralihan bahasa yang terjadi pada otak manusia.

Pada penelitian oleh (Mouriño García et al., 2017), dilakukan klasifikasi teks dengan menggunakan metode cross-language concept matching (CLCM). Metode ini bekerja dengan cara mengkonversi dokumen dari satu bahasa ke bahasa lainnya tanpa melakukan penerjemahan dokumen terlebih dahulu. Data yang digunakan pada penelitian ini diambil dari wikipedia dokumen dengan menggunakan 2 bahasa, yaitu Inggris dan Spanyol. Hasil akhir penelitian menunjukkan bahwa metode ini sangat bermanfaat untuk klasifikasi dokumen bahkan dengan *training* data dengan jumlah sedikit dapat menghasilkan performa yang sangat baik. Metode ini juga bisa digunakan untuk dokumen dengan ukuran yang berbeda-beda. Selain dari itu, pada penelitian lainnya oleh (Yan et al., 2020), menunjukkan adanya perbedaan hasil klasifikasi teks dengan metode Bag-of-words (BoW) pada *network* yang berbeda-beda. Percobaan dilakukan pada 3 *state network,* yaitu *dynamic*, *static* dan *hybrid network*. Hasil penelitian menyebutkan bahwa penggunaan *dynamic network* menghasilkan klasifikasi yang terbaik dengan metode BoW.

Penelitian ini akan membahas klasifikasi teks dengan metode BoW. Metode ini mampu mengubah teks menjadi matriks yang dapat digunakan untuk menentukan *label* bahasa. Klasifikasi yang dimaksud adalah klasifikasi bahasa kedalam 4 jenis bahasa yaitu Inggis, Jerman, Indonesia dan Francis. Berikutnya, data akan diambil dari Application Programming Interface (API) twitter dengan autentikasi melalui *developer* akun. Dengan mengaplikasikan metode ini, diharapkan dapat menghasilkan model untuk klasifikasi teks dengan akurasi yang baik.

* 1. **Rumusan Masalah**

Terdapat ragam bahasa di dunia yang digunakan untuk komunikasi sehari-hari. Tidak semua bahasa dapat diketahui dan dimengerti secara langsung. Pada penelitian terdahulu disebutkan terdapat kendala peralihan bahasa pada saat komunikasi berlangsung. Oleh karena itu, penelitian ini akan membahas mengenai bagaimana membuat model untuk klasifikasi bahasa pada teks dalam *machine learning* dengan metode BoW untuk mempermudah peralihan bahasa.

* 1. **Tujuan Penelitian**

Tujuan akhir penelitian ini adalah untuk dapat membuat model yang mampu melakukan klasifikasi bahasa dengan baik sehingga komunikasi dapat berjalan dengan baik.

* 1. **Batasan Masalah**

Batasan masalah yang diberikan antara lain sebagai berikut.

* Data yang akan digunakan sebagai *training* data dan *test* data diambil dari API twitter.
* Pengenalan bahasa yang dimaksud adalah klasifikasi bahasa ke dalam 4 jenis bahasa yaitu, bahasa Indonesia, bahasa Prancis, Bahasa Jerman, dan bahasa Inggris.
* Jumlah keseluruhan data twitter yang digunakan adalah 4000 *records*.
* Klasifikasi akan dilakukan dengan menggunakan metode BoW (Bag-of-Words)
  1. **Manfaat Penelitian**

Pada akhirnya, penelitian ini diharapkan dapat menghasilkan model dengan akurasi yang baik, sehingga dapat membantu mempermudah peralihan bahasa saat komunikasi melalui teks berlangsung.

**BAB 2**

**LANDASAN TEORI**

* 1. **Machine learning**

*Machine learning* merupakan jenis AI (Artificial Intelligence) yang mampu membantu sistem untuk mempelajari data secara langsung tanpa programming. Machine learning banyak digunakan dalam dunia bisnis sebagai alat untuk melakukan inovasi baru terhadap pelayanan ataupun produk dengan mempelajari *asset* data yang mereka miliki (Hurwitz & Kirsch, 2018). *Machine learning* digunakan untuk sistem yang terlalu kompleks untuk di *program* secara manual dan juga untuk sistem yang membutuhkan adaptasi secara terus-menerus. Penyelesaian masalah dengan mempertimbangkan banyak faktor, seperti prediksi harga rumah dapat dilakukan oleh manusia, tapi tidak oleh mesin. Maka dari itu, algoritma dibutuhkan oleh komputer untuk dapat menyelesaikan masalah tersebut.

Selanjutnya, ketersediaan data serta kemampuan komputasi yang baik menjadikan alasan perkembangan *machine learning* menjadi semakin cepat. Pengguna *device* *digital* yang semakin banyak, memungkinkan tersedianya data dalam jumlah besar yang dapat digunakan untuk menemukan informasi baru yang tersembunyi di dalam data tersebut. Pada sisi lain, model merupakan data yang diproses sedemikian rupa untuk dapat menentukan jawaban atas suatu permasalahan. Sedangkan *machine learning* model merupakan kumpulan data yang telah diproses dengan perhitungan matematika *machine learning* serta di-*train* dan kemudian diuji dengan *testing* data untuk mendapatkan jawaban dari permasalahan tertentu (Maass & Storey, 2021).  Model ini dapat digunakan kembali untuk menyelesaikan permasalahan yang sama.

Pada penelitian ini akan digunakan algoritma klasifikasi Bag-of-Words (BoW). Metode tersebut menggunakan geometri matematika sebagai *classifier*. Setiap kata akan disusun dalam bentuk matriks sesuai dengan kata kunci yang dipilih sebagai penentu. Selanjutnya data akan diklasifikasikan ke *label* yang paling mendekati sebagai hasilnya. Untuk lebih jelasnya lagi (Yan et al., 2020) menyebutkan bahwa BoW model bekerja dengan melakukan *mapping* menjadi suatu vektor sebagai berikut.

*v* = [ *x1 , x2 , …, xn* ] dimana *xi* menunjukkan frekuensi munculnya kata pada *array*. *Array* tersebut terbentuk dari *dataset* dan merupakan kata-kata penentu yang banyak digunakan pada teks.

Chart, diagram, bubble chart

Description automatically generated

Gambar 2.1 Machine Learning merupakan bagian dari AI ( Hurwitz & Kirsch, 2018)

Gambar di atas merupakan gambaran umum AI, dimana machine learning merupakan salah satu bagiannya. Selain machine learning, terdapat *subset* lain dari AI yaitu (1) machine reasoning untuk mengambil kesimpulan dari data, (2) Natural Language Processing (NLP) untuk pembelajaran teks dan bahasa, (3) planning yang merupakan sistem pintar yang dapat menciptakan suatu tindakan sekuensial, serta (4) machine learning yang memiliki 4 macam pendekatan (Hurwitz & Kirsch, 2018) .

NLP (Natural Language Processing) mampu menemukan manfaat dari data dalam bentuk teks. NLP menggunakan algoritma komputer dalam pembelajarannya untuk dapat mengerti dan mengenali inti dari bahasa yang digunakan oleh manusia (Cai, 2021). Pada smartphone terdapat *virtual assistant* yang dapat digunakan untuk memberikan perintah tertentu yang dapat dimengerti oleh perangkat tersebut. Siri, Cortana, Google Assistant merupakan contoh *virtual assistant* yang ada saat ini (Booth, 2018). *Virtual assistant* tersebut adalah contoh implementasi NLP yang sangat menarik dan masih akan dikembangkan dalam beberapa waktu kedepan.

Pada gambar 2.1 dijelaskan bahwa terdapat 4 pendekatan dalam machine learning. Berikut adalah pembahasannya.

* Supervised Learning

Pendekatan ini merupakan pembelajaran data dimana data tersebut akan diproses ke dalam *label* tertentu. *Label* tersebut sudah ditentukan terlebih dahulu oleh peneliti. Contohnya, model yang dapat menentukan apakah kanker yang terdapat pada pasien mematikan atau tidak.

* Unsupervised Learning

*Unsupervised learning* adalah mempelajari data dalam jumlah besar yang belum diketahui *label*-nya. Dimana proses ini mengelompokkan hasil secara langsung tanpa bantuan peneliti.

* Reinforcement Learning

Pada *reinforcement learning*, agen berusaha mempelajari lingkungan sekitar. Agen akan dilatih dengan cara eksplorasi dan pembelajaran dari *experience* atau *input* yang pernah diberikan oleh *user* sebelumnya. Melalui *experience* tersebut, agen diharapkan dapat beradaptasi untuk melakukan tindakan terbaik dalam menyelesaikan permasalahan yang terjadi saat ini (Asadulaev et al., 2020). Contoh *reinforcement learning* adalah permainan catur pada komputer.

* Deep Learning

*Neural network* kompleks atau deep learning digunakan untuk mempelajari big data abstrak dan tidak terstruktur. Perkembangan *deep learning* sangat pesat dimana berbagai perusahaan besar seperti Mercedez, Baidu, dan perusahan lainnya menggunakan metode ini untuk *self-driving car*  (Chen et al., 2021).

Pada Artificial Neural Network (ANN) terdapat berbagai model yang dapat digunakan untuk masalah klasifikasi dan regresi. Multi-layer Perceptron (MLP) merupakan model paling sederhana yang sering digunakan (Feng et al., 2020; Zhou et al., 2021). Pada model ini terdapat input, output dan satu atau lebih *hidden* layer. Setiap layer terhubung satu sama lain dan setiap neuron pada layer yang sama tidak terhubung sama sekali. Berikut adalah rumus MLP dari (Feng et al., 2020):

**( 1 )**

Penjelasan rumus di atas adalah, *g* merupakan fungsi aktivasi, *wij* merupakan weight pada input layer *j* dan hidden layer *i, R* adalah jumlah input, *xi* adalah neuron ke *i* pada input layer, dan *bj* yang merupakan bias ke *j* pada hidden layer.

* 1. **Bag of Words ( BoW )**

BoW merupakan metode sederhana yang banyak digunakan untuk pengolahan teks dalam NLP. Metode ini bekerja dengan menggunakan perhitungan frekuensi munculnya kata dalam suatu kalimat ataupun dokumen. Metode ini telah digunakan di berbagai bidang seperti analisa forensik kepemilikan data, sentiment analysis, dan lain sebagainya (Ishihara, 2021; Silva et al., 2018). Menurut (Ghalyan et al., 2018) terdapat tiga fitur utama dalam BoW, yaitu *extraction*, menemukan bagian terpenting dalam suatu teks maupun gambar, *quantization* yang merupakan *assigning* setiap kata ke dalam *descriptor* dan yang terakhir *classification*, yaitu penentuan hasil melalui pembelajaran *boundary* dan perbedaan antara setiap *class*.

Berikut adalah kelebihan dari metode BoW menurut (Ofer et al., 2021):

* BoW dapat menghasilkan *fixed-sized* input vektor.
* Mampu menormalisasi hasil vektor untuk menampilkan frekuensi token.
* Bisa digunakan untuk teks dengan ukuran panjang yang berbeda-beda.
* Merupakan metode yang sederhana dan menghasilkan output yang baik untuk analisis teks yang tidak berhubungan dengan struktur bahasa.

Dalam metode BoW, kata menjadi penentu utama sehingga struktur kalimat menjadi masalah yang dikesampingkan, hal tersebut membuat BoW menjadi pilihan yang sangat terbatas untuk menyelesaikan permasalahan teks yang mendalam seperti *summarization* ataupun pengumpulan data yang spesifik dari dokumen (Lauriola et al., 2021; Verberne et al., 2010).

* 1. **Twitter API**

Twitter API (Application Programming Interface) merupakan aplikasi yang dibuat oleh twitter yang dapat digunakan untuk mengambil data dengan menggunakan akun twitter developer. Twitter sendiri merupakan aplikasi sosial media berisi berbagai macam tulisan unggahan dari penggunanya. Dengan menggunakan API ini, kita bisa mendapatkan data *tweet*, *timeline*, *direct* *message*, *trend twitter* dan lain sebagainya. Terdapat SDK (Software Development Kit) atau *libraries* dalam bahasa pemrograman Python, Ruby dan Javascript / Node JS.

Twitter digunakan secara luas, sehingga data yang terdapat di dalamnya memiliki jumlah yang besar. Data dari social media banyak digunakan untuk mempelajari berbagai macam permasalahan seperti masalah pandemi yang terjadi saat ini, bencana alam, dan berbagai masalah lainnya (Hou et al., 2020). Untuk dapat mengakses data twitter dibutuhkan akun *developer* dari twitter. Selanjutnya melalui portal *developer, developer* diharuskan untuk membuat proyek baru yang di dalamnya terdiri dari beberapa *apps*. Setiap *apps* memiliki *keys* and *token* yang akan digunakan untuk *authentication* dalam *request* penarikan data dari twitter.

* 1. **Google Cloud Platform**

*Cloud* merupakan *platform* yang menyediakan berbagai service untuk mengembangkan aplikasi dan hosting. Dalam setiap cloud provider, terdapat beberapa *essentials* *service* yang dimiliki secara umum. Berikut *service* dasar yang dimiliki *cloud*:

* *Compute*, yang merupakan kegiatan komputasi melalui internet, seperti *Virtual Machine* (VM).
* *Storage*, yaitu menyimpan data pada penyimpanan *virtual* melalui internet.
* *Networking*, untuk hosting dengan network yang stabil, cepat dan terpercaya serta mampu mengurangi efek dari pemadaman server.
* *Database*, untuk akses data lebih mudah, keamanan yang baik, sentralisasi data dan menghindari *database* *sprawl*.

*Cloud Platform* semakin diminati dalam dunia bisnis karena lebih efektif dan tidak memerlukan adanya perangkat keras secara fisik sehingga biaya *maintenance* menjadi sangat kecil. Selain itu, dengan menggunakan GCP seluruh *service* menjadi *reliable* dan *accessible* bagi penggunanya.

* 1. **Big Query**

Data menjadi bagian penting dari *machine learning* karena data merupakan aset utama yang akan dipelajari dan dijadikan indikasi untuk diolah dengan tujuan mendapatkan hasil yang diinginkan. Salah satu tempat penyimpanan dan pengolahan data milik GCP adalah bigQuery. BigQuery adalah *enterprise* *data* *warehouse* yang membantu pengolahan dan analisa data. Big query menggunakan *sql command* untuk melakukan *query* data pada database. Dengan menggunakan bigQuery, *developer* dapat langsung mengakses data dengan menggunakan REST (Representational State Transfer) API ataupun RPC (Remote Procedure Call) API.

* 1. **Tensorflow**

Tensorflow merupakan *software* yang dikembangkan oleh tim riset Google Brain yang digunakan sebagai *platform* untuk *machine learning* dan *deep neural network*. *Software* ini menggabungkan komputasi aljabar dengan teknik optimasi untuk mempermudah perhitungan matematika yang rumit (Zaccone, 2016). Fitur - fitur yang tersedia dalam tensorflow adalah sebagai berikut :

* Mendefinisikan dan mengoptimalkan ekspresi matematika dengan efisien.
* Mendukung programming untuk teknik deep neural network dan machine learning.
* Penggunaan GPU (Graphic Processing Unit) *computing* yang transparan, optimasi serta manajemen memori dan data.
* Komputasi machine learning dan data set dalam ukuran besar dengan skalabilitas tinggi.

**BAB 3**

**METODOLOGI PENELITIAN**

* 1. **Kerangka Penelitian**

Pada penelitian ini, terdapat beberapa langkah yang dilakukan untuk mencapai tujuan akhir penelitian. Langkah tersebut adalah sebagai berikut.

* + 1. *Mengumpulkan data*

Proses pengumpulan data diambil dari twitter API. Dengan menggunakan python akan dilakukan penarikan data ke salah satu *end-point* twitter. Data yang akan diambil di filter sesuai kebutuhan. Jumlah keseluruhan data yang akan digunakan adalah 4000 data. Dalam pengumpulan data, GCP BigQuery digunakan untuk membentuk database yang selanjutnya akan digunakan sebagai data untuk *training* dan *testing*. Keseluruhan data akan digabungkan dan diteliti agar tidak ada data yang tidak normal.

* + 1. *Menyiapkan data*

Setelah data diambil dari twitter, selanjutnya data tersebut akan dibersihkan serta diberikan label dengan menggunakan *library preprocessor*. *Library* ini merupakan *library* untuk python yang berfungsi untuk membersihkan, *parsing* serta *tokenizer* data.

* + 1. *Pembagian data training dan testing*

Data yang telah didapat sebelumnya, kemudian akan dibagi menjadi data training dan data testing. Untuk training set akan digunakan delapan puluh persen dari keseluruhan data, dan sisanya yaitu dua puluh persen akan digunakan sebagai training dataset.

* + 1. *Create model*

Selanjutnya, model akan dibuat dengan tensorflow, dengan menggunakan *library* keras. Keras adalah *tools* untuk deep learning yang di desain untuk meminimalisir user action. Model yang akan dibuat akan menggunakan klasifikasi Multi Layer Perceptron dengan fungsi aktivasi sigmoid, sedangkan metode yang digunakan adalah metode Bag-of-words.

* + 1. *Training the data*

Setelah berhasil membuat model, selanjutnya proses training akan dijalankan melalui GCP (Google Cloud Platform). Berikutnya akan dilakukan evaluasi hasil training data pada platform yang sama. Training data bertujuan untuk menentukan kombinasi parameter terbaik yang dapat digunakan sebagai penentu hasil output sehingga didapatkan prediksi dengan hasil akurasi yang lebih baik (Maldonado-Cruz & Pyrcz, 2021).

* + 1. *Testing and Analysis*

Testing data selanjutnya akan di input ke model yang telah di deploy. Model tersebut akan di analisis untuk melihat hasil kerjanya.

A picture containing graphical user interface

Description automatically generated

Gambar 3. 1 Proses Penelitian

Gambar diatas merupakan proses yang akan dilakukan untuk menghasilkan model pada penelitian ini. Secara keseluruhan akan digunakan python sebagai programming language, GCP sebagai platform pengembangannya, bigquery untuk pengolahan data serta tensorflow untuk fungsi machine learning.

Sedangkan model akan diprogram dengan bahasa python. Gambar 3.2 dibawah ini, menggambarkan bagaimana proses dalam pembuatan model. Mulai dari penarikan data dari API twitter, proses *cleaning* data, membuat model, hingga *train* dan *test* data.

Diagram

Description automatically generated

Gambar 3.2 Flowchart Program

**DAFTAR PUSTAKA**

Asadulaev, A., Kuznetsov, I., Stein, G., & Filchenkov, A. (2020). Exploring and Exploiting Conditioning of Reinforcement Learning Agents. *IEEE Access*, *8*, 211951–211960. https://doi.org/10.1109/ACCESS.2020.3037276

Booth, J. D. (2018). *Natural Language Processing Succinctly Foreword by Daniel Jebaraj*. www.syncfusion.com.

Cai, M. (2021). Natural language processing for urban research: A systematic review. In *Heliyon* (Vol. 7, Issue 3). Elsevier Ltd. https://doi.org/10.1016/j.heliyon.2021.e06322

Chen, H., Zhang, Y., Cao, Y., & Xie, J. (2021). *Security Issues and Defensive Approaches in Deep Learning Frameworks* (Vol. 26, Issue 6). http://creativecommons.org/licenses/by/4.0/

Feng, X., Ma, G., Su, S. F., Huang, C., Boswell, M. K., & Xue, P. (2020). A multi-layer perceptron approach for accelerated wave forecasting in Lake Michigan. *Ocean Engineering*, *211*. https://doi.org/10.1016/j.oceaneng.2020.107526

Ghalyan, I. F. J., Chacko, S. M., & Kapila, V. (2018). Simultaneous robustness against random initialization and optimal order selection in Bag-of-Words modeling. *Pattern Recognition Letters*, *116*, 135–142. https://doi.org/10.1016/j.patrec.2018.09.010

Hou, Q., Han, M., & Cai, Z. (2020). Survey on data analysis in social media: A practical application aspect. In *Big Data Mining and Analytics* (Vol. 3, Issue 4, pp. 259–279). Tsinghua University Press. https://doi.org/10.26599/BDMA.2020.9020006

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning IBM Limited Edition*. http://www.wiley.com/go/permissions.

Ishihara, S. (2021). Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International*, *327*. https://doi.org/10.1016/j.forsciint.2021.110980

Kriesel, D. (n.d.). *Neural Networks*. http://www.dkriesel.com/en/science/neural\_networkshttp://www.dkriesel.com/en/tech/snipe

Lauriola, I., Lavelli, A., & Aiolli, F. (2021). An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2021.05.103

Liu, H., Zhang, M., Pérez, A., Xie, N., Li, B., & Liu, Q. (2019). Role of language control during interbrain phase synchronization of cross-language communication. *Neuropsychologia*, *131*, 316–324. https://doi.org/10.1016/j.neuropsychologia.2019.05.014

Maass, W., & Storey, V. C. (2021). Pairing conceptual modeling with machine learning. *Data and Knowledge Engineering*, *134*. https://doi.org/10.1016/j.datak.2021.101909

Maldonado-Cruz, E., & Pyrcz, M. J. (2021). Tuning machine learning dropout for subsurface uncertainty model accuracy. *Journal of Petroleum Science and Engineering*, *205*. https://doi.org/10.1016/j.petrol.2021.108975

Mouriño García, M. A., Pérez Rodríguez, R., & Anido Rifón, L. (2017). Wikipedia-based cross-language text classification. *Information Sciences*, *406–407*, 12–28. https://doi.org/10.1016/j.ins.2017.04.024

Ofer, D., Brandes, N., & Linial, M. (2021). The language of proteins: NLP, machine learning & protein sequences. In *Computational and Structural Biotechnology Journal* (Vol. 19, pp. 1750–1758). Elsevier B.V. https://doi.org/10.1016/j.csbj.2021.03.022

Shrivastava, M., & Kumar, S. (2021). A pragmatic and intelligent model for sarcasm detection in social media text. *Technology in Society*, *64*. https://doi.org/10.1016/j.techsoc.2020.101489

Silva, F. B., Werneck, R. de O., Goldenstein, S., Tabbone, S., & Torres, R. da S. (2018). Graph-based bag-of-words for classification. *Pattern Recognition*, *74*, 266–285. https://doi.org/10.1016/j.patcog.2017.09.018

Verberne, S., Boves, L., & Oostdijk, N. (2010). *What Is Not in the Bag of Words for Why-QA?* http://direct.mit.edu/coli/article-pdf/36/2/229/1810264/coli.09-032-r1-08-034.pdf

Yan, D., Li, K., Gu, S., & Yang, L. (2020). Network-Based Bag-of-Words Model for Text Classification. *IEEE Access*, *8*, 82641–82652. https://doi.org/10.1109/ACCESS.2020.2991074

Zaccone, G. (2016). *Getting started with tensorflow : get up and running with the latest numerical computing library by Google and dive deeper into your data!* Packt Publishing.

Zhou, Y., Liu, Y., Wang, D., & Liu, X. (2021). Comparison of machine-learning models for predicting short-term building heating load using operational parameters. *Energy and Buildings*, *253*. https://doi.org/10.1016/j.enbuild.2021.111505