**PROPOSAL PROYEK**

**11S4037 – PEMROSESAN BAHASA ALAMI**

**Multi-label Emotion Classification using Bi-LSTM with GloVe Word Embedding**



**Disusun oleh:**

|  |  |
| --- | --- |
| **12S18004** | **Rosalia Pane** |
| **12S18008** | **Indah Tri Anastasya Manik** |
| **12S18011** | **Nadya Putri Tambunan** |
| **12S18043** | **Roy Gunawan napitupulu** |
| **12S18048** | **Rifka Uli Siregar** |

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2021**

**DAFTAR ISI**

[**BAB I PENDAHULUAN** 5](#_Toc89107066)

[**1.1.** **Latar Belakang** 5](#_Toc89107067)

[**1.2.** **Tujuan** 7](#_Toc89107068)

[**1.3.** **Manfaat** 7](#_Toc89107069)

[**1.4.** **Ruang Lingkup** 7](#_Toc89107070)

[**BAB II ISI** 8](#_Toc89107071)

[**2.1 Analisis** 8](#_Toc89107072)

[**2.1.1 Analisis Data** 8](#_Toc89107073)

[**2.1.2 Analisis Metode** 11](#_Toc89107074)

[**2.2 Desain** 11](#_Toc89107075)

[**2.2.1 *Mapping Label*** 12](#_Toc89107076)

[**2.2.2 *Data Preprocessing*** 12](#_Toc89107077)

[**2.2.3 *Word Embedding Glove*** 15](#_Toc89107078)

[**2.2.4 *Modelling with Bi-LSTM*** 15](#_Toc89107079)

[**2.2.5 *Evaluation and Results*** 16](#_Toc89107080)

[**2.3 Implementasi** 16](#_Toc89107081)

[**2.3.1 *Mapping Label*** 16](#_Toc89107082)

[**2.3.2 *Data Preprocessing*** 16](#_Toc89107083)

[**2.3.3 *Word Embedding Glove*** 22](#_Toc89107084)

[**2.3.3 *Modeling with Bi-LSTM*** 23](#_Toc89107085)

[**2.4 Hasil** 24](#_Toc89107086)

[**2.4.1 Evaluation BI-LSTM Model – Glove** 24](#_Toc89107087)

[**2.4.2 *Accuracy Bi-LSTM Model with GloVe*** 25](#_Toc89107088)

[**BAB III PENUTUP** 26](#_Toc89107089)

[**3.1 Pembagian Tugas dan Tanggung Jawab** 26](#_Toc89107090)

[**3.2** **Kesimpulan** 28](#_Toc89107091)

[**3.2** **Saran** 28](#_Toc89107092)

[**DAFTAR PUSTAKA** 29](#_Toc89107093)

**DAFTAR TABEL**

[**Tabel 1. Atribut pada dataset** 8](#_Toc89098506)

[**Tabel 2. Pembagian tugas dan tanggung jawab** 26](#_Toc89098507)

**DAFTAR GAMBAR**

[**Gambar 1. Distribusi Label pada Dataset** 10](#_Toc89098508)

[**Gambar 2. Flowchart desain** 12](#_Toc89098509)

[**Gambar 3. Contoh Remove words containing numbers** 14](#_Toc89098510)

[**Gambar 4. Contoh clean contractions** 14](#_Toc89098511)

[**Gambar 5.Contoh clean special characters** 14](#_Toc89098512)

[**Gambar 6.Contoh Correct Spelling** 14](#_Toc89098513)

[**Gambar 7.Modeling with Bi-LSTM** 15](#_Toc89098514)

[**Gambar 8. Code program mengecek nilai null pada data** 17](#_Toc89098515)

[**Gambar 9.Output pengecekan nilai null** 17](#_Toc89098516)

[**Gambar 10. Kode Program Clean Text** 18](#_Toc89098517)

[**Gambar 11. Hasil Clean Text** 18](#_Toc89098518)

[**Gambar 12. Kode Program Clean Constractions** 19](#_Toc89098519)

[**Gambar 13. Hasil Clean Constractions** 19](#_Toc89098520)

[**Gambar 14. Kode Program Clean Special Characters** 20](#_Toc89098521)

[**Gambar 15. Hasil Clean Special Character** 20](#_Toc89098522)

[**Gambar 16. Kode Program Correct Spelling** 20](#_Toc89098523)

[**Gambar 17. Hasil Correct Spelling** 21](#_Toc89098524)

[**Gambar 18. Kode Program Remove Space** 21](#_Toc89098525)

[**Gambar 19. Hasil Remove Space** 22](#_Toc89098526)

[**Gambar 20. Kode Program Embedding GloVe** 23](#_Toc89098527)

[**Gambar 21. Output Embedding GloVe** 23](#_Toc89098528)

[**Gambar 22. Kode Program Modelling dengan Bi-LSTM** 23](#_Toc89098529)

[**Gambar 23. Output Modelling dengan Bi-LSTM** 24](#_Toc89098530)

[**Gambar 24. Kode Evaluasi Bi-LSTM - Glove** 24](#_Toc89098531)

[**Gambar 25. Output Evaluasi Bi-LSTM - Glove** 24](#_Toc89098532)

[**Gambar 26. Kode Program Akurasi Bi-LSTM – Glove** 25](#_Toc89098533)

[**Gambar 27. Output Akurasi Bi-LSTM - Glove** 25](#_Toc89098534)

# **BAB I PENDAHULUAN**

Bagian ini menyajikan latar belakang, tujuan, manfaat dan ruang lingkup pengerjaan proyek.

## **Latar Belakang**

Emosi adalah keadaan pikiran yang berlangsung terus-menerus, yang ditandai dengan gejala mental, fisik dan perilaku. Emosi seseorang dapat diidentifikasi secara langsung melalui ekspresi wajah dan ucapannya. Mendeteksi emosi secara otomatis sangat penting karena dapat diterapkan di berbagai bidang. Misalnya dalam dunia pendidikan, analisis emosi dapat dimanfaatkan untuk lingkungan *e-learning*. Selain itu, dalam bisnis yang digunakan untuk mengidentifikasi keluhan pelanggan [1].

Dari pengalaman sehari-hari, beberapa emosi tampaknya berbeda dan terjadi secara independen. Emosi yang secara inheren kontradiktif, seperti love dan hate mungkin memerlukan serangkaian kelas yang terpisah untuk mengakomodasikan aspek dari setiap kelas. Disisi lain, emosi yang identik biasanya berada dibawah valensi emosional yang sama dan sering muncul bersamaan dalam situasi tertentu. Oleh karena itu, berbagai emosi ini dapat dikelompokkan bersama. Deteksi emosi, yang berperan sebagai masalah klasifikasi *multi-label* dapat membantu menjelaskan sifat kompleks dari emosi yang terjadi bersamaan, sehingga memberikan pemahaman tentang karakteristik setiap emosi [2].

Pendeteksian emosi merupakan salah satu masalah yang muncul di bidang *Natural Language Processing* (NLP). NLP digunakan untuk memproses data seperti teks yang terstruktur maupun tidak menjadi pengetahuan bermakna untuk berbagai masalah bisnis. NLP telah banyak digunakan untuk solusi masalah seperti klasifikasi, pemodelan topik, *text generation*, QA *system*, rekomendasi, dan lainnya[3].

Baru-baru ini, masalah klasifikasi *multi-label* telah menarik minat yang cukup besar karena penerapannya ke berbagai *domain*, termasuk klasifikasi teks, klasifikasi adegan dan video, dan bioinformatika [4]. Berbeda dengan masalah klasifikasi tradisional label tunggal (yaitu, multi-kelas atau biner), di mana sebuah *instance* dikaitkan dengan hanya satu label dari satu set label yang terbatas, dalam masalah klasifikasi *multi-label*, sebuah *instance* dikaitkan dengan *subset* dari label [5].

Pendeteksian emosi menggunakan klasifikasi *multi-label* menjadi masalah karena suatu kalimat cenderung melibatkan lebih dari satu kategori emosi. Sehingga, tantangan utama yang muncul adalah bagaimana memodelkan ketergantungan antar label menggunakan pendekatan klasifikasi. Misalnya, emosi dengan label “*angry*” dan “*disgust*” memiliki ketergantungan daripada emosi “*sad*” dan “*joy*” yang saling bertentangan [6]. Analisis emosi melalui sebuah teks tampaknya juga menjadi tantangan karena faktanya bahwa ekspresi tekstual tidak selalu secara langsung melibatkan kata-kata yang berhubungan dengan emosi, tetapi seringkali suatu kalimat perlu dipahami untuk memberikan sebuah makna [7].

Salah satu pendekatan yang dapat digunakan untuk mengatasi masalah klasifikasi multi-label adalah metode *Bidirectional Long Short Term Memory* (Bi-LSTM). Metode Bi-LSTM merupakan perkembangan dari model LSTM dengan dua lapisan, dimana lapisan pertama bergerak maju (*forward*) untuk memahami dan memproses dari kata pertama menuju kata terakhir, demikian sebaliknya lapisan atasnya bergerak mundur (*backward*) untuk memahami dan memproses dari kata terakhir menuju kata pertama. Oleh karena itu, Bi-LSTM sangat baik digunakan untuk mengenali pola dalam kalimat, dikarenakan setiap kata dalam kalimat diproses secara sekuensial [8].

Kemudian dalam melakukan klasifikasi, metode Bi-LSTM akan digabungkan dengan salah satu pendekatan pembobotan kata yaitu *pre-trained word embeddings GloVe*. Dimana, *GloVe* merupakan salah satu pendekatan yang memiliki akurasi yang baik untuk memproses pembobotan kata dalam data dibanding model *word embeddings* lain seperti CBOW dan *skip-grams*. Secara keseluruhan, *GloVe* mengungguli model lain dalam hal analogi kata, kemiripan kata dan tugas *named entity recognition* [9]

Oleh karena itu, berdasarkan uraian dari permasalahan sebelumnya, penulis berfokus pada pengklasifikasian emosi *multi-label*, yang bertujuan untuk mengembangkan sistem otomatis untuk mengkategorikan kalimat ke dalam netral dan 27 emosi seperti *admiration*, *amusement*, *anger* dan emosi lainnya. Penggunaan pendekatan Bi-LSTM dan *word embedding GloVe* akan membantu dalam membentuk matriks *embedding* pada masalah klasifikasi *multi-label* dengan menggunakan *dataset GoEmotions* yang diperoleh dari *Hugging Face.*

## **Tujuan**

Tujuan dari proyek *multi-label emotion classification* ini, antara lain:

1. Menerapkan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dengan *GloVe* sebagai *word embedding* dalam menganalisis teks apakah teks tersebut termasuk ke dalam teks dengan beberapa label (*multi-label*).
2. Untuk mengetahui bagaimana tingkat akurasi menggunakan metode *Bidirectional Long Short Term Memory* (Bi-LSTM) dengan *GloVe* sebagai *word embedding* dalam melakukan pengklasifkasian *multi-label*.

## **Manfaat**

Berikut adalah manfaat dari pembuatan *multi-label emotion classification* menggunakan metode Bi-LSTM dan *GloVe* antara lain:

1. Mengetahui cara dan proses klasifikasi *multi-label emotion* menggunakan metode Bi-LSTM dan *GloVe* sebagai *word embedding*.
2. Mengetahui tingkat akurasi menggunakan metode Bi-LSTM dan *GloVe* sebagai *word embedding* dalam melakukan klasifikasi *multi-label emotion*.

## **Ruang Lingkup**

Ruang lingkup dalam pengerjaan proyek ini yaitu menggunakan metode Bi-LSTM dan *word embedding* *GloVe* dengan menggunakan *dataset GoEmotions* yang diperoleh dari *Hugging Face* [10]*.*

# **BAB II ISI**

Pada bab ini mencakup analisis yaitu analisis terhadap data dan analisis terhadap metode.

# **DAFTAR PUSTAKA**

|  |  |
| --- | --- |
| [1] | M. S. a. M. r. a. A. M. Saputri, "Emotion Classification on indonesian twitter dataset," in *2018 International Conference on Asian Language Processing (IALP)*, 2018, pp. 90--95. |
| [2] | V. R. Berhitoe, Multi-label emotion detection in Twitter, 2017. |
| [3] | A. Das, "Multi-Label Emotion Classification with PyTorch + HuggingFace's Transformers and W&B for Tracking," [Online]. Available: https://towardsdatascience.com/multi-label-emotion-classification-with-pytorch-huggingfaces-transformers-and-w-b-for-tracking-a060d817923. [Accessed 10 November 2021]. |
| [4] | J. a. P. B. a. H. G. a. F. E. Read, "Classifier chains for multi-label classification," *Machine learning,* vol. 85, no. 3, pp. 333--359, 2011. |
| [5] | M. a. M. A. Jabreel, "A deep learning-based approach for multi-label emotion classification in tweets," *Applied Sciences,* vol. 9, no. 6, p. 1123, 2019. |
| [6] | D. a. J. X. a. L. J. a. L. S. a. Z. Q. a. Z. G. Zhang, "Multi-modal Multi-label Emotion Detection with Modality and Label Dependence," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3584--3593. |
| [7] | A. R. a. K. K. A. Murthy, "A Review of Different Approaches for Detecting Emotion from Text," in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 012009. |
| [8] | H. F. a. H. A. F. Fadli, "Identifikasi Cyberbullying pada Media Sosial Twitter Menggunakan Metode LSTM dan BiLSTM," *AUTOMATA,* vol. 2, no. 1, 2021. |
| [9] | J. a. S. R. a. M. C. D. Pennington, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532--1543. |
| [10] | D. a. M.-A. D. a. K. J. a. C. A. a. N. G. a. R. S. Demszky, "GoEmotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547,* 2020. |
| [11] | J. a. P. J. a. K. M. Han, Data mining: concepts and techniques, Elseiver, 2011. |
| [12] | A. O. a. J. S. Salau, "Feature extraction: a survey of the types, techniques, applications," in *2019 International Conference on Signal Processing and Communication (ICSC)},*, 2019. |
| [13] | M. Eklund, Comparing Feature Extraction Methods and Effects of Pre-Processing Methods for Multi-Label Classification of Textual Data, 2018. |