Deep Learning Course
Final Project Report
Dataset Analisis Sentimen

Andi Alisha Faiqihah (H071221010)
Dewa Ayu Eka Natalia Pratiwi (H071221021)
Laode Fahmi Hidayat (H071221022)

November 28, 2024

Contents

1 Introduction

Munculnya media sosial telah merevolusi cara kita berkomunikasi secara global, membawa banyak manfaat tetapi juga memunculkan fenomena negatif seperti komentar buruk dan pelecehan online. Komentar negatif di media sosial sering kali berupa label yang merugikan dan dapat memiliki dampak psikiatrik yang serius, jauh lebih besar daripada dampak dari bullying langsung. Kementerian Komunikasi dan Informasi Republik Indonesia telah mengingatkan warganet untuk menghindari komentar negatif demi menjaga reputasi digital mereka.

Dampak dari komentar negatif ini tidak hanya dirasakan oleh individu, tetapi juga oleh masyarakat secara keseluruhan. Komentar tersebut dapat merendahkan martabat korban, menyebabkan stres, dan berkontribusi pada masalah kesehatan mental seperti depresi. Selain itu, anak-anak yang sering menerima komentar negatif cenderung mengembangkan perilaku tidak etis dan bahasa kasar, yang dapat merusak kemampuan mereka dalam berkomunikasi dengan baik.

Untuk memahami kompleksitas dan luasnya fenomena komentar buruk di media sosial, serta untuk mengembangkan metode mitigasi yang efektif, diperlukan dataset analisis sentimen yang representatif. Dataset ini harus mencakup survei yang luas mengenai berbagai topik, platform media sosial, dan konteks yang berbeda. Dengan analisis ini, kita dapat mengidentifikasi pola, intensitas, dan frekuensi komentar negatif serta profil pengguna yang cenderung terlibat dalam perilaku tersebut.

Selain itu, dataset sentimen ini dapat digunakan sebagai dasar untuk pengembangan sistem deteksi otomatis komentar negatif. Sistem ini akan membantu dalam mengidentifikasi dan menghapus komentar yang tidak pantas secara cepat dan efisien. Edukasi publik tentang pentingnya berperilaku positif di internet juga menjadi salah satu tujuan utama dari penelitian ini, dengan harapan dapat meningkatkan kesadaran masyarakat tentang risiko psikologis akibat komentar negatif.

2 Related Works

Berbagai penelitian telah menunjukkan efektivitas penggunaan deep learning dalam analisis sentimen, terutama pada data dari media sosial. Studi oleh Hafidzah et al. (2024) menyimpulkan bahwa model deep learning seperti CNN dan LSTM memberikan hasil yang lebih baik dibandingkan metode tradisional seperti SVM dalam hal akurasi, recall, dan skor F1[1]. Penelitian ini juga menunjukkan bahwa kombinasi model BERT dan CNN memiliki kinerja unggul, dengan BERT sering digunakan karena kemampuannya dalam menangkap konteks yang kompleks melalui arsitektur Transformer[1].

Ahmad et al. (2023) juga mengungkapkan keunggulan CNN dan LSTM dalam analisis sentimen. CNN efektif dalam mengekstrak fitur spasial, sedangkan LSTM mampu menangkap hubungan temporal jangka panjang dalam teks. Penelitian mereka menyoroti pentingnya kombinasi kedua model ini untuk meningkatkan kinerja analisis sentimen, terutama dalam kasus data ulasan yang besar seperti di platform Amazon[2].

Di sisi lain, Kamarula & Rochmawati (2022) menekankan pentingnya metode Word2Vec untuk representasi kata dalam analisis sentimen menggunakan model CNN dan Bi-LSTM. Hasil penelitian mereka menunjukkan bahwa Bi-LSTM unggul dalam akurasi dibandingkan CNN ketika digunakan untuk analisis sentimen dan emosi masyarakat Indonesia di media sosial[3].

Proyek ini melanjutkan fondasi dari penelitian-penelitian sebelumnya dengan mengadopsi pendekatan kombinasi CNN dan LSTM, menggunakan Word2Vec untuk representasi kata. Justifikasi penggunaan CNN dan LSTM terletak pada kemampuannya untuk menangkap pola temporal serta fitur spasial dalam data teks informal. Selain itu, teknik augmentasi data direncanakan untuk memperkaya variasi dataset, sehingga meningkatkan generalisasi model dalam menghadapi tantangan bahasa informal di media sosial.

Dengan pendekatan ini, proyek ini diharapkan dapat berkontribusi signifikan terhadap pengembangan metodologi analisis sentimen yang lebih efektif dan adaptif, terutama dalam mengatasi tantangan teks informal yang sering ditemukan di platform media sosial.

3 Dataset and Material

3.1 Sumber Dataset

Dataset yang digunakan dalam proyek ini dikumpulkan dari berbagai platform media sosial dan aplikasi menggunakan teknik scraping. Scraping sendiri merupakan cara pengambilan suatu data atau informasi tertentu dengan jumlah besar untuk nantinya digunakan dalam berbagai keperluan seperti riset, analisis dan lainnya[4].

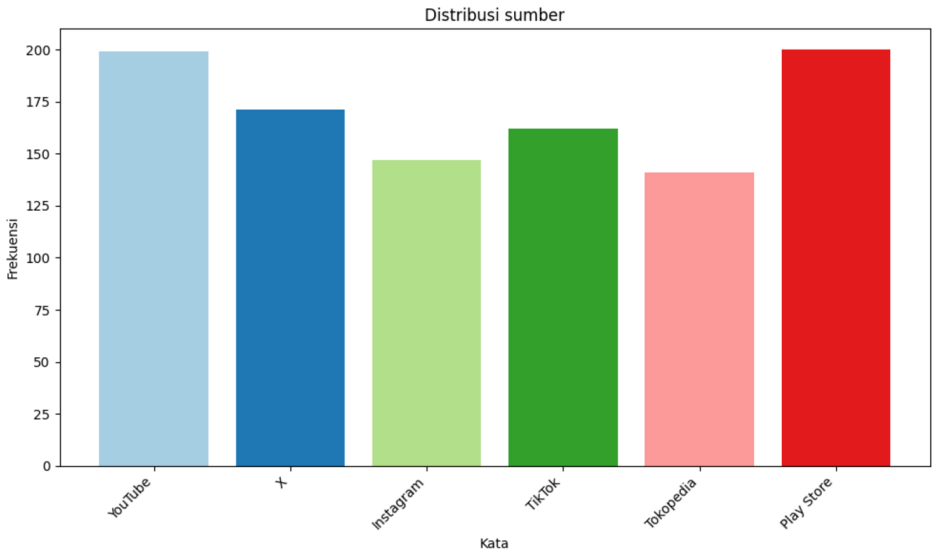

Figure 1: Distibusi Sumber Dataset

Tujuan pengumpulan data ini adalah untuk mendapatkan variasi komentar dan opini yang mencakup berbagai topik, emosi, serta konteks yang berbeda. Berikut merupakan sumber-sumber dari dataset ini:

- YouTube
- Instagram
- Play Store
- Tokopedia

- X (sebelumnya Twitter)
- Tiktok

3.2 Pra-Processing

Figure 2: Word Cloud Dataset

Dataset yang dikumpulkan melalui scraping membutuhkan beberapa langkah pra-processing untuk memastikan data yang diperoleh siap untuk analisis dan pembuatan model. Berikut adalah langkah-langkah pra-pemrosesan yang dilakukan:

1. **Pengumpulan Data:** Data dikumpulkan melalui teknik web scraping dari platform yang telah disebutkan, menggunakan alat seperti BeautifulSoup dan Selenium.
2. **Penghapusan Duplikasi:** Menghapus data yang duplikat untuk memastikan setiap entri unik.
3. **Pembersihan Teks:** Menghapus karakter khusus, tautan, dan tanda baca yang tidak diperlukan dari teks komentar.
4. **Normalisasi Teks:** Mengubah semua teks menjadi huruf kecil untuk konsistensi.
5. **Tokenisasi:** Memecah teks menjadi kata-kata atau token.
6. **Penghapusan Stopwords:** Menghilangkan kata-kata umum yang tidak memiliki makna signifikan dalam analisis (seperti ”dan”, ”atau”, ”yang”).
7. **Labeling:** Menandai setiap komentar dengan label positif atau negatif berdasarkan analisis sentimen awal.

3.3 Fitur dan Label

Dataset yang dihasilkan mencakup beberapa fitur penting yang digunakan untuk analisis dan pembuatan model:

- **Komentar:** Teks ulasan atau komentar dari pengguna.
- **Timestamp:** Waktu saat komentar dibuat.
- **Sumber:** Platform dari mana komentar tersebut diambil (YouTube, Instagram, Play Store, Tokopedia, X, TikTok).
- **Label:** Kategori sentimen dari komentar, yang bisa berupa positif atau negatif.

3.4 Alat, Perpustakaan, dan Kerangka Kerja yang Digunakan

Untuk mengimplementasikan model dan analisis, berikut beberapa alat, perpustakaan, dan kerangka kerja yang digunakan:

- **Python:** Bahasa pemrograman utama yang digunakan untuk pengumpulan data, pra-pemrosesan, dan pembuatan model.
- **BeautifulSoup dan Selenium:** Alat scraping web yang digunakan untuk mengumpulkan data dari berbagai platform.
- **Pandas:** Perpustakaan untuk manipulasi dan analisis data.
- **NumPy:** Perpustakaan untuk operasi numerik yang efisien.
- **NLTK (Natural Language Toolkit):** Digunakan untuk tugas pemrosesan bahasa alami seperti tokenisasi dan penghapusan stopwords.
- **Scikit-learn:** Perpustakaan yang digunakan untuk pra-pemrosesan data dan pembuatan model machine learning.
- **TensorFlow dan Keras:** Kerangka kerja deep learning yang digunakan untuk membangun dan melatih model neural network, termasuk model BERT dan MLP.
- **Matplotlib dan Seaborn:** Perpustakaan visualisasi data yang digunakan untuk membuat grafik dan plot untuk analisis data eksploratif.

Dataset ini juga akan digunakan untuk melatih model analisis sentimen berbasis deep learning seperti Word2Vec-CBOW dan LSTM. Model-model ini dipilih karena kemampuannya dalam menangkap konteks dan nuansa emosional dalam teks dengan lebih baik dibandingkan metode tradisional. Dengan memanfaatkan dataset yang kaya dan beragam ini, diharapkan model yang dihasilkan dapat memberikan akurasi tinggi dalam mengklasifikasikan sentimen di media sosial, serta mampu menangani tantangan yang muncul dari penggunaan bahasa informal dan kebisingan dalam data.

4 Result and Discussion

Hasil penelitian menunjukkan bahwa model berbasis Word2Vec dengan metode Continuous Bag of Words (CBOW) dan Long Short-Term Memory (LSTM) memberikan performa yang baik dalam klasifikasi sentimen, dengan keunggulan masing-masing sesuai kebutuhan analisis.

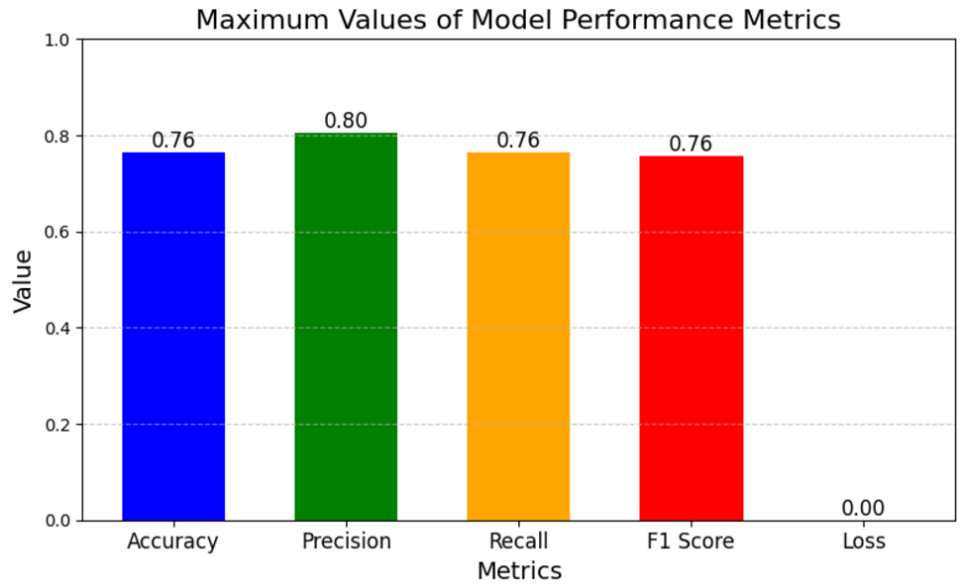

Figure 3: Visualisasi Word2Vec-CBOW

Pada model Word2Vec-CBOW, performa terbaik tercapai dengan test accuracy sebesar 76%, serta metrik precision 80%, recall 76%, dan F1 score 76%. Meskipun akurasi ini tergolong baik, proses pelatihan membutuhkan waktu lebih lama untuk mencapai stabilitas, terutama pada dataset dengan tingkat keragaman yang tinggi. Keunggulan utama dari Word2Vec-CBOW terletak pada kemampuannya menghasilkan representasi kata yang efisien, memungkinkan model menangkap pola-pola fitur penting dari teks. Namun, model ini cenderung sensitif terhadap noise, sehingga preprocessing data yang cermat diperlukan untuk memastikan embedding kata optimal.

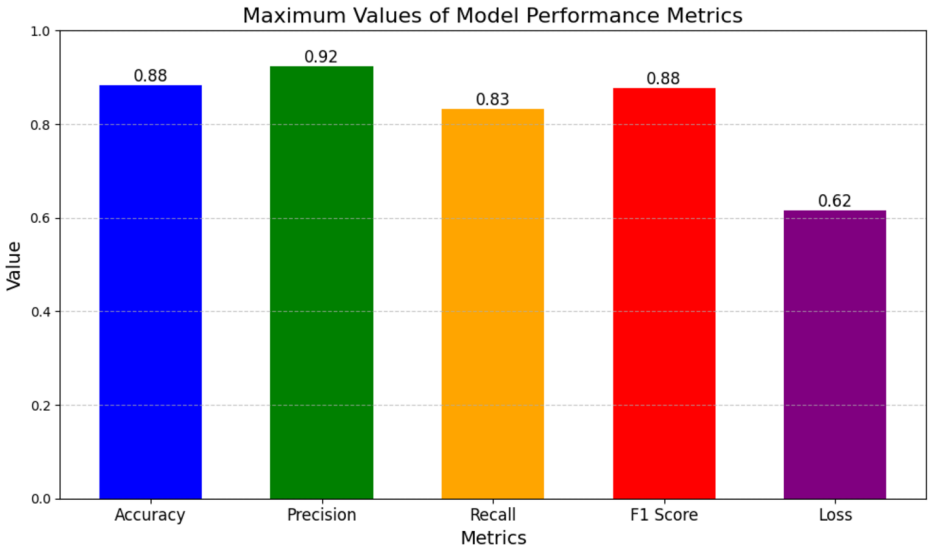

Figure 4: Visualisasi LSTM

Sebaliknya, model LSTM menunjukkan performa yang lebih kompetitif dengan test accuracy sebesar 88,23%, precision 92,39%, recall 83,33%, dan F1 score 87,62%. Kemampuan utama LSTM adalah menangkap pola ketergantungan antar kata dalam data

berurutan, sehingga menghasilkan performa yang lebih stabil. Distribusi metrik performanya yang lebih seimbang menunjukkan kemampuan generalisasi yang baik terhadap data yang belum pernah dilihat. Namun, waktu pelatihan yang lebih lama dibandingkan Word2Vec-CBOW menjadi salah satu kekurangannya.

Referensi dari jurnal oleh Ahmad et al. (2023) mendukung hasil ini. Pada dataset ulasan produk Amazon, model Word2Vec-CBOW dalam penelitian tersebut mencapai akurasi 87%, sementara model LSTM pada dataset serupa berhasil mencapai akurasi 93,66%. Hal ini menunjukkan bahwa kedua model dapat memberikan performa optimal tergantung pada karakteristik dataset. Secara keseluruhan:

- **Word2Vec-CBOW** lebih cocok untuk skenario dengan data bersih di mana efisiensi waktu pelatihan menjadi prioritas, meskipun akurasi relatif lebih terbatas.
- **LSTM** lebih unggul dalam menangani data berurutan atau konteks semantik yang kompleks, memberikan hasil yang lebih stabil dengan generalisasi lebih baik, meskipun memerlukan waktu pelatihan lebih panjang.

Dengan demikian, pemilihan antara kedua model ini harus disesuaikan dengan kebutuhan spesifik proyek, seperti efisiensi waktu pelatihan, tingkat akurasi, atau kemampuan generalisasi. Penelitian ini mengonfirmasi bahwa baik Word2Vec-CBOW maupun LSTM adalah pilihan yang sangat kompetitif untuk tugas klasifikasi sentimen.

5 Conclusion

Penelitian ini bertujuan untuk mengevaluasi performa model berbasis Word2Vec-CBOW dan LSTM dalam tugas klasifikasi sentimen, dengan fokus pada akurasi, stabilitas, dan kemampuan generalisasi. Tujuan ini berhasil dicapai dengan membandingkan kedua model menggunakan metrik kinerja utama, termasuk akurasi, precision, recall, dan F1 score. Hasil menunjukkan bahwa Word2Vec-CBOW memberikan akurasi 76% dengan keunggulan dalam efisiensi waktu pelatihan, sementara LSTM mencapai akurasi yang lebih tinggi sebesar 88,23%, didukung oleh stabilitas performa dan kemampuan menangkap pola ketergantungan antar kata dalam data berurutan.

Beberapa wawasan kunci yang diperoleh dari penelitian ini adalah bahwa Word2Vec-CBOW lebih cocok untuk dataset yang bersih dan membutuhkan pelatihan cepat, sedangkan LSTM lebih unggul dalam menangani konteks semantik yang kompleks dengan generalisasi yang lebih baik pada data yang belum terlihat. Namun, sensitivitas Word2Vec-CBOW terhadap noise dan waktu pelatihan yang lebih lama pada LSTM menjadi pertimbangan dalam pemilihan model.

Untuk penelitian di masa depan, disarankan untuk mengeksplorasi pengaruh metode preprocessing data terhadap performa Word2Vec-CBOW, serta menguji arsitektur LSTM yang lebih ringan untuk mengurangi waktu pelatihan. Selain itu, penelitian dapat diperluas ke dataset lain dengan karakteristik berbeda untuk mengonfirmasi generalisasi hasil ini. Kombinasi Word2Vec-CBOW dan LSTM dalam model hybrid juga dapat menjadi arah yang menjanjikan untuk mencapai hasil yang lebih optimal.

References

[1] Hafidzah, P., Maryani, S., Ihsani, B. Y., Nurmiwati, N., Erwin, E., Niswariyana, A. K. Penerapan Deep Learning dalam Menganalisis Sentimen di Media Sosial. *Seminar Nasional Paedagoria*, 2024. Available online: https://journal.ummat.ac.id/index.php/fkip/article/view/25651 (accessed on November 27, 2024).

[2] Ahmad, S., Ridwan, A. M., Setiawan, G. D. Analisis Sentimen Product Tools Home Menggunakan Metode CNN dan LSTM. *TEKNOKOM: Jurnal Teknologi dan Rekayasa Sistem Komputer*, 2023, 6(2), 133-140. Available online: https://www.researchgate.net/publication/373078716_ANALISIS_SENTIMEN_PRODUCT_TOOLS_HOME_MENGGUNAKAN_METODE_CNN_DAN_LSTM (accessed on November 27, 2024).

[3] Kamarula, M. R. F., Rochmawati, N. Perbandingan CNN dan Bi-LSTM pada Analisis Sentimen dan Emosi Masyarakat Indonesia di Media Sosial Twitter Selama Pandemik Covid-19 yang Menggunakan Metode Word2Vec. *JINACS: Journal of Informatics and Computer Science*, 2022, 4(2). Available online: https://ejournal.unesa.ac.id/index.php/jinacs/article/view/50063/42702 (accessed on November 27, 2024).

[4] Telkom University. Web Scraping: Pengertian dan Fungsinya dalam Pengambilan Data. Available online: https://it.telkomuniversity.ac.id/web-scraping-pengertian-dan-fungsinya-dalam-pengambilan-data/ (accessed on November 27, 2024).

[5] Muhammad, P. F., Kusumaningrum, R., Wibowo, A. Sentiment analysis using Word2VEC and Long Short-Term Memory (LSTM) for Indonesian hotel reviews. *Procedia Computer Science*, 2021, 179, 728–735. Available online: https://doi.org/10.1016/j.procs.2021.01.061 (accessed on November 27, 2024).