BAB 1
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan berkembangnya teknologi informasi, terlihat bagaimana masyarakat mulai memiliki rasa ketergantungan dengan produk yang berhubungan dengan teknologi, salah satunya adalah media sosial. Seakan-akan medial sosial sudah menjadi wadah untuk mengekspresikan semua hal, dari mulai mengekspresikan kesenangan, kekesalan terhadap suatu hal, bahkan tidak segan untuk menghina seseorang. Salah satu media sosial yang sering digunakan untuk sarana berekspresi adalah Twitter. Tak jarang komentar / *tweet* yang diunggah oleh pengguna mengandung *sarcasm*(sindiran). *Sarcasm* merupakan pemakaian kata-kata pedas guna menyakiti hati orang lain, cemoohan maupun ejekan agresif. *Sarcasm* bisa mengganti arti dari suatu kalimat ataupun teks menjadi kebalikan dari isi teks tersebut sehingga salah mengartikan arti teks tersebut[1].

Dari fenomena *sarcasm* tersebut akhirnya banyak peneliti di bidang *Data Science*/*Machine Learning* yang melakukan riset mengenai *sentiment analysis* untuk mendeteksi / mengklasifikasikan apakah sebuah kalimat mengandung sarkasme atau tidak. Sentimen adalah sikap, pemikiran, atau penilaian yang didorong oleh perasaan[2]. Sedangkan klasifikasi teks merupakan pengelompokkan suatu dokumen teks berdasarkan isi dari teks maupun karakteristik dari teks tersebut. Klasifikasi teks dapat dilakukan dengan cara membedakan teks yang telah ditentukan ke dalam teks tertentu atau beberapa kelas tertentu[3]. Tetapi seringkali dalam proses pengembangannya, ditemui kasus ketidakseimbangan data / *imbalance datasets*. *Imbalance Datasets* adalah sebuah kondisi dataset dalam melakukan klasifikasi dimana proporsi dari label / target yang dimiliki sangat timpang jauh

Pada penelitian sebelumnya mengenai *sentiment analysis* menggunakan metode *support vector machine(SVM)* yang diimplementasi pada kasus opini masyarakat terhadap penanganan pemerintah terkait COVID-19 di media sosial

twitter, pada penelitian ini hasil akurasi yang dihasilkan adalah 82%[4]. Kemudian pada penelitian *sentiment analysis* lainnya ada yang menggunakan metode Naïve Bayes yang diimplementasi pada kasus e-sports untuk digunakan pada kurikulum pendidikan, pada penelitian ini hasil akurasi yang dihasilkan adalah 70.32%[5]. Pada penelitian lainnya mengenai *sentiment analysis* untuk mendeteksi sarkasme pada komentar Twitter sudah dilakukan menggunakan metode *Bi-LSTM*[6]. *Bi-LSTM* adalah pengembangan dari model LSTM dimana terdapat dua lapisan yang prosesnya saling berkebalikan arah, model ini sangat baik untuk mengenali pola dalam sebuah kalimat karena setiap kata dalam dokumen diproses secara sekuensial[7]. Penelitian tersebut menggunakan Datasets yang tidak seimbang(*Imbalance Dataset*) dengan jumlah 6.699 data, yang terdiri dari 4.926 label sarkasme dan 1.773 label non-sarkasme. Pada penelitian tersebut menghasilkan akurasi 77% menggunakan arsitektur *CBOW*, dan 76% menggunakan arsitektur *Skip-gram.* Pada penelitian tersebut evaluasi model *Bi-LSTM* belum mampu mencapai kategori baik untuk masalah klasifikasi sarkasme. Salah satu penyebab hal ini terjadi adalah karena penggunaan Datasets yang tidak seimbang(*Imbalanced Datasets*)[6].

Salah satu solusi dari kasus *Imbalanced Datasets* adalah dengan menerapkan metode *Synthetic Minority Oversampling Technique*(SMOTE). SMOTE adalah salah satu turunan dari *oversampling*. Metode ini bekerja dengan membuat replikasi dari data minoritas, replikasi tersebut dikenal dengan data sintetis(*syntetic data*). Pada penelitian yang dilakukan oleh Hermanto pada tahun 2020 mengenai *sentiment analysis* komentar aplikasi Gojek pada *platform* google play, terbukti bahwa sebuah model SVM yang dikombinasikan dengan SMOTE menghasilkan akurasi yang lebih tinggi dibandingkan dengan Naïve Bayes(tanpa SMOTE) dengan hasil 81.09% untuk SVM+SMOTE dan 74.41% untuk Naïve Bayes[8]. Pada penelitian lainnya mengenai implementasi metode SMOTE yang dilakukan oleh Yerik Afrianto pada kasus analisis sentimen terhadap produk dan layanan restoran di labuan bajo, terbukti bahwa model yang dibangun menggunakan SMOTE+KNN memiliki hasil akurasi yang lebih tinggi dibandingkan dengan KNN saja, dengan nilai akurasi 93.23% untuk KNN dan 99.27% untuk

SMOTE+KNN[9]. Kemudian pada penelitian lain yang dilakukan oleh Suja Alex pada tahun 2022 memberikan hasil bahwa metode SMOTE+LSTM memiliki hasil akurasi yang lebih tinggi dibandingkan dengan LSTM saja, dengan rincian 0,85 untuk SMOTE+LSTM dan 0,80 untuk LSTM saja[10].

Berdasarkan latar belakang yg telah disampaikan sebelumnya, maka dalam penelitian ini akan menggunakan metode Bi-LSTM untuk membuat model analisis sentiment sarkasme. Metode ini dipilih karena Bi-LSTM dapat membantu mengatasi masalah *vanishing gradient* yang sering terjadi pada arsitektur LSTM. Masalah *vanishing gradient* terjadi ketika gradien yang mengindikasikan seberapa besar pengaruh suatu neuron terhadap kesalahan total model menyebar melalui jaringan. Selain itu pada penelitian ini juga menggunakan metode SMOTE untuk membuat jumlah sampel di kelas minoritas sama dengan kelas mayoritas.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas, maka dapat diidentifikasi permasalahan yaitu Datasets yang tidak seimbang(*Imbalance Dataset*) dapat mempengaruhi akurasi suatu model pada kasus sentiment analysis *sarcasm*, atau dengan kata lain Datasets yang seimbang berpotensi untuk meningkatkan akurasi dari suatu model dibandingkan dengan dataset yang *imbalance*.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk mengukur nilai akurasi yang dihasilkan dari model pendeteksian *sarcasm* pada komentar masyarakat di sosial media Twitter menggunakan metode BiLSTM dan metode SMOTE untuk membuat jumlah *sample* pada setiap *class* menjadi seimbang.

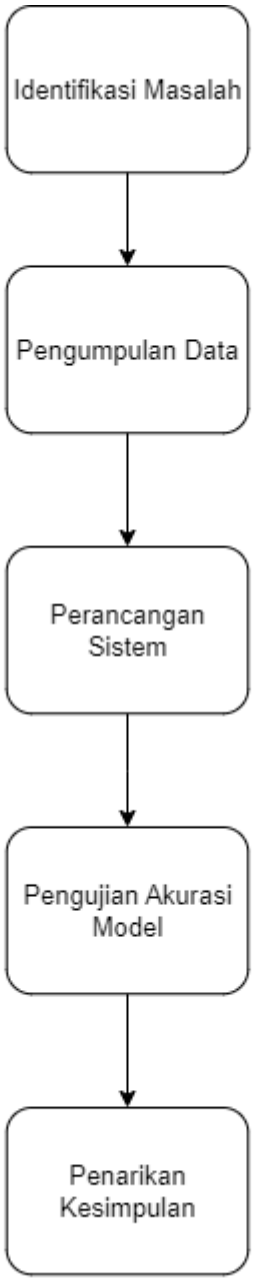
1.4 Batasan Masalah

Dengan mengidentifikasi terhadap masalah-masalah yang ada, agar lebih terarah dan dapat dipahami dengan mudah, maka perlu dilakukan pembatasan masalah. Pembatasan masalah pada penelitian ini antara lain:

1. Kalimat/Text yang dimasukkan pada sistem ini menggunakan Bahasa Indonesia baku sesuai dengan Kamus Besar Bahasa Indonesia(KBBI) yang didapat dari hasil *Crawling* pada media sosial twitter periode 28-09-2022 sampai 25-03-2023. *Keywords* yang digunakannya adalah “AD”, “CN”, “AB”, “PM”, “LP”, “GP”, dan “RP”.
2. Jika terdapat bahasa daerah atau bahasa asing, maka perlu dilakukan transformasi terlebih dahulu ke dalam bentuk bahasa Indonesia baku sesuai dengan Kamus Besar Bahasa Indonesia(KBBI).
3. Komentar yang di *crawling* dari twitter merupakan komentar masyarakat terhadap politikus yang ada di Indonesia.
4. Menggunakan *Confusion Matrix* untuk melakukan evaluasi terhadap model yang telah dikembangkan.

1.5 Metode Penelitian

Metode penelitian yang digunakan adalah metode eksperimental. Metode ini dipilih karena sesuai dengan lingkup penelitian yang membutuhkan percobaan berbagai macam nilai parameter untuk membantu membuat model yang lebih baik. Proses penelitian ini dimulai dari Identifikasi masalah yang terdapat pada penelitian sebelumnya, kemudian pengumpulan Datasets yang di *crawling* dari media sosial Twitter, dilanjutkan dengan perancangan sistem, pengujian akurasi terhadap data baru, dan yang terakhir adalah penarikan kesimpulan. Alur penelitian yang akan digunakan dalam penelitian ini dapat dilihat pada Gambar 1.1.



Gambar 1.1 Alur Penelitian

1.5.1 Identifikasi Masalah

Identifiaksi masalah pada penelitian ini adalah dengan cara menganalisis penelitian sebelumnya yang sudah dilakukan, dimana salah satu hasil analisisnya adalah belum di implementasinya metode *BiLSTM* pada Datasets yang seimbang

dan apakah dengan diimplementasinya metode SMOTE untuk membuat datasets menjadi seimbang bisa meningkatkan nilai akurasinya atau tidak.

1.5.2 Pengumpulan Data

Pengumpulan data pada penelitian ini adalah dengan cara melakukan *crawling* komentar / tweet pada media sosial Twitter mengenai komentar masyarakat terhadap politikus di Indonesia, contoh dari hasil *crawling* data dapat dilihat pada tabel 1.1.

Tabel 1.1 Hasil Crawling Data

	Data
1	Bukankah hanya pada saat mencemooh, putus asa, marah, dan sejenis itu kita menekankan suku kata terakhir pada kata-kata yang terdir #CakNun 😳
2	Terdengar suara-suara tak indah dalam perjuangan menuntut hak juga tidak mengapa. kalau goverment hanya pengin mendengar suara-suar #CakNun 😳
3	Wkwkwkkwkwk @caknundotcom kesambet nya udah dari lama deh kayaknya… wokwokwokwok 😳
4	@aniesbaswedan prettttt,gagasan mulu yg dibahas nyatanya zonk,eksekusinya apalagi...haha ga malu kah semua jualanmu ity berbanng terbalik dari kenyataan. 😠
5	@Wiji_kedua @aniesbaswedan Pemborosan Bangunan tingkat kampung Sebentar bangun, sebentar bongkar... 😠
6	@tvOneNews Alhamdulillah.... Segala Puji bagiMu ya Allah...\nBarakallah Bpk Anies Baswedan... \nAnies Baswedan kereeeeeeeeeeeeennnn....!!! 😠

7	Kemarin ada yg lempar - lempar kaos sambil masem mukanya, ini lempar - lempar uang sambil tersenyum.\nSambo\n#PuanMaharani https://t.co/jt2NGli9SV 💯
8	@KompasTV Ga perduli penjelasanmuuuu puan, memang kau itu orangnya begitu. Kau sembunyi2kan pun, memang kau ini tak suka pada rakyat kecil.\nMengatasnamakan rakyat kecil yg mana partaimu. dompetmu kah yg kecil ????????\n@puanmaharani_ri 0l 💯
9	@RositaPratiwi9 @puanmaharani_ri Presiden PDIP 🤣🤣 💯
10	Ingat kata pak luhut binsar panjaitan LBP\n\nMENJADI ABDI/ PAHLAWAN NEGARA ITU TIDAK HANYA MENJADI SEORANG PRESIDEN ATAU MENTRI ORANG BIASA JUGA BISA ASALKAN NIATNYA UNTUK NEGARA 🥹

1.5.3 Perancangan Sistem

Prototype dari model/sistem yang akan dibangun untuk mengklasifikasikan apakah sebuah teks yang berasal dari twitter mengandung unsur sarkasme atau tidak dapat dilihat pada gambar 1.2.

Gambar 1.2 Prototype Perancangan Sistem

Dari gambar 1.2 di atas, dapat dilihat bahwa dalam proses Bi-LSTM pertama-tama *Datasets* yang digunakan akan dilakukan *preprocessing* terlebih dahulu, seperti *data cleaning*, *stopwords removal*, dan *data transformation*. Setelah *Data Preprocessing* langkah selanjutnya adalah melakukan proses *Data Labelling* yang bertujuan untuk melabeli setiap data dengan label sarkasme(1) dan

non-sarkasme(0). Setelah seluruh data yang terdapat pada Datasets sudah bersih dan sudah terlabeli seluruhnya, langkah selanjutnya adalah melakukan proses *feature extractions*, proses ini dilakukan untuk merubah seluruh text yang merupakan data kategorik menjadi numerik, pada penelitian ini metode yang akan digunakan untuk melakukan *feature extractions* adalah GloVe(Global Vectors for Word Representations). Setelah merubah seluruh text dalam kalimat menjadi representasi numerik, langkah selanjutnya adalah melakukan proses *Oversampling* menggunakan metode SMOTE(*Synthetic Minority Oversampling Technique*), *oversampling* ini bertujuan untuk membuat data sintetik dari kelas minoritas agar jumlah sample pada kelas minoritas tersebut memiliki jumlah yang sama dengan kelas mayoritas. Selanjutnya dilakukan pembentukan arsitektur model Bi-LSTM, model ini terdiri dari dua lapisan LSTM, yaitu LSTM maju(*forward*) dan LSTM mundur(backward). LSTM maju memproses text dari awal hingga akhir, sementara LSTM mundur memproses text dari akhir hingga awal. Output dari kedua lapisan LSTM ini digabungkan untuk menghasilkan representasi text yang lebih kaya dan dipahami konteksnya. Setelah model Bi-LSTM dibangun, tahap selanjutnya adalah pelatihan model. Model dilatih menggunakan data latih yang telah diberi label sebelumnya. Selama proses pelatihan, bobot dan parameter model disesuaikan menggunakan algoritma backpropagation. Setelah melalui proses pelatihan, model Bi-LSTM dievaluasi menggunakan data uji yang belum pernah dilihat sebelumnya. Evaluasi dilakukan dengan menggunakan *metrics* evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-Score* untuk mengukur performa model dalam mengklasifikasikan sarkasme.

1.5.4 Pengujian Akurasi Model

Pada tahap ini akan menggunakan *Confusion Matrix* untuk melakukan proses perhitungan akurasi dengan cara mengetahui jumlah data uji yang terklasifikasi dengan benar maupun salah dalam pengklasifikasiannya. Hasil yang diambil adalah nilai akurasi model ketika diberikan data uji.

1.5.5 Penarikan Kesimpulan

Jika akurasi di rentang 80% - 100% berarti metode Bi-LSTM yang dikombinasikan dengan SMOTE dapat meningkatkan akurasi dari model. Yang berarti dapat disimpulkan bahwa *Datasets* yang seimbang dapat meningkatkan hasil akurasi dari ML Model yang dibangun.

1.6 Sistematika Penulisan

Untuk mempermudah melihat dan mengetahui pembahasan yang ada pada skripsi ini secara menyeluruh, maka perlu dikemukakan sistematika yang merupakan kerangka dan pedoman penulisan skripsi. Adapun sistematika penulisannya adalah sebagai berikut:

Penyajian laporan skripsi ini menggunakan sistematika penulisan sebagai berikut:

1. Bagian Awal Skripsi

   Bagian awal memuat halaman sampul depan, halaman judul, halaman persetujuan dosen pembimbing, halaman pengesahan, halaman motto dan persembahan, halaman kata pengantar, halaman daftar isi, halaman daftar tabel, halaman daftar gambar, halaman daftar lampiran, arti lambang dan singkatan abstraksi.

2. Bagian Utama Skripsi

   Bagian Utama terbagi atas bab dan sub bab yaitu sebagai berikut:

   BAB I PENDAHULUAN

   Bab ini terdiri dari latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan skripsi.

BAB II LANDASAN TEORI

Bab ini meliputi pembahasan pengertian Analisis Sentimen, Sarkasme, Politikus, Data Preprocessing, Bi-LSTM, SMOTE, Confusion Matrix.

BAB III ANALISIS DAN PERANCANGAN

Bab ini menjelaskan secara teknis mengenai tahap-tahap yang dilakukan pada pengembangan model analisis sentiment deteksi sarkasme, yang meliputi *Data Preprocessing*, *Data Labelling*, *Word Embedding*, SMOTE, Bi-LSTM, *Testing*, dan analisis kebutuhan non-fungsional.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Bab ini menjelaskan mengenai implementasi dari proses analisis dan perancangan model/sistem analisis sentiment untuk mendeteksi sarkasme, yang meliputi implementasi *Data Preprocessing*, Implementasi metode GloVe untuk membuat *Word Embedding*, implementasi metode SMOTE, implementasi model *Bi-LSTM*, pengujian parameter, pengujian performansi, dan kesimpulan dari hasil evaluasi model.

BAB V KESIMPULAN DAN SARAN

Bab ini menjelaskan hasil dari penelitian berdasarkan tahap pengujian. Hasil tersebut ditarik menjadi suatu kesimpulan dan saran untuk penelitian-penelitian berikutnya.

3. Bagian Akhir Sripsi

   Bagian akhir dari skripsi ini berisi tentang daftar pustaka dan daftar lampiran.