BAB 1

PENDAHULUAN

1.1 Latar Belakang

Cyberbullying merupakan perilaku menyimpang yang dilakukan dengan sengaja untuk menjatuhkan seseorang melalui media internet. Dampak negatif akibat *Cyberbullying* adalah gangguan mental, traumatik, depresi hingga bunuh diri pada korbannya tanpa melihat golongan, umur, dan latar belakang. Diperlukan langkah nyata untuk memutus mata rantai tindakan *Cyberbullying* sebagai *Cyber Crime* sudah tentu pemerintah selaku pemangku konstitusi di negara ini melalui aksi nyata dapat memberikan edukasi literasi media kepada semua lapisan masyarakat serta memberikan peringatan didalamnya melalui penanganan secara tegas dengan atas nama keadilan dan hukum perundang-undangan yang mengatur didalamnya, melalui undang-undang nomor 11 Tahun 2008 telah diatur tentang informasi dan transaksi elektronik (UU ITE) dalam pasal 27 ayat (3) dan (4)[1]. Instagram adalah salah satu dari yang platform media social yang saat ini digunakan oleh masyarakat.

Namun dengan adanya kemajuan teknologi dalam media sosial membawa dampak negatif akibat kurang bijaknya netizen dalam berkomentar dan terjadinya *Cyberbullying*. Telah dilakukan penelitian-penelitian mengenai *Cyberbullying* pada media sosial. penelitian yang dilakukan oleh Heri Santoso dalam mendeteksi komentar *Cyberbullying* pada platform media sosial Instagram[2]. Selain itu, penelitian yang dilakukan oleh Fauzan Baehaqi dan Nuri Cahyono dalam analisis Sentimen Terhadap Cyberbullying Pada Komentar di Instagram [3]. Penelitian yang dilakukan oleh Rizky Dhian Syarif dkk pada klasifikasi komentar Instagram dalam kasus pemilihan presiden pada tahun 2019 dengan menggunakan metode *Lexicon Based* dan *Naïve Bayes Classifier*[4].

Pada penelitian sebelumnya dilakukan oleh Rizky Dhian Syarif dkk memperoleh hasil pengujian *Lexicon-Based* dengan akurasi 58% tidak lebih baik dibandingkan metode *Naïve Bayes Classifier* dengan nilai akurasi 97% dikarenakan penggunaan kata yang tidak baku yang tidak terdeteksi oleh sistem walaupun pada penelitian ini kamus bahasa gaul sudah diterapkan. Sehingga keterbatasan kamus sentimen yang digunakan sebagai identifier bisa mempengaruhi hasil klasifikasi yang kurang baik[4]. Padahal penelitian yang dilakukan oleh Muhammad Wisnu Prayuda dkk dalam penggunaan metode *Lexicon-Based* dalam menganalisis sentimen terkait dengan fenomena mudik pada saat perayaan Lebaran di Indonesia memiliki nilai akurasi sebesar 85%[5]. Penelitian yang dilakukan oleh Solagratia Saron Tandiapa dan Gladly Caren Rorimpandey membahas tentang analisis sentimen terhadap ulasan pengguna pada aplikasi Threads menggunakan dua metode, yaitu *Lexicon Based* dan *Naive Bayes Classifier* dan perbandingan akurasi dari kedua metode menghasilkan 55% untuk metode *Lexicon Based* dan 51% untuk metode *Naive Bayes Classifier*. Dari hasil tersebut dapat menunjukkan bahwa hasil akurasi dari metode *Lexicon-Based* memiliki nilai yang lebih tinggi dari pada metode *Naive Bayes Classifier*[6]. Penelitian yang dilakukan oleh Galuh Etha Pratiwi dkk yang melakukan Analisis Sentimen pada twitter mengenai Penggunaan artis Korea Selatan sebagai brand ambassador produk Indonesia menggunakan metode *Lexicon* yang mana hasil pengujian menunjukkan akurasi 78%[7].

Berdasarkan permasalahan pada penelitian Rizky Dhian Syarif dkk yang

penggunaan kata yang tidak baku yang tidak terdeteksi oleh sistem maka solusi yang dapat diterapkan pada penelitian ini adalah adanya tahapan *Word Normalization* dan *Typo-Checking* agar membantu mengatasi masalah *Lexicon Based* dengan perbaikan kata tidak baku dan perbaikan kesalahan ejaan. Hal ini diperkuat oleh penelitian yang dilakukan oleh Novrido Charibaldi dkk yang melakukan membahas pengaruh normalisasi kata terhadap dua metode klasifikasi yang umum digunakan dalam analisis sentimen, yaitu klasifikasi *Naïve Bayes* dan metode *K-Nearest Neighbor*

(KNN) dimana hasil pengujian *K-Nearest Neighbor* dengan *Word Normalization* memiliki akurasi 80.48% sedangkan *K-Nearest Neighbor* tanpa *Word Normalization* memiliki akurasi 77.14%[8]. Selain itu, penelitian sebelumnya oleh Katarzyna Marszalek-Kowalewsk membahas dampak *Word Normalization* dalam bahasa Persia dimana menunjukkan bahwa *Word Normalization* secara signifikan meningkatkan akurasi yaitu *F1-Score* 26%[9]. Penelitian yang dilakukan oleh Muhammad Javed dkk yang melakukan *Word Normalization* pada analisis sentimen dalam penanganan teks yang tidak terstruktur dan informal yang menunjukkan adanya peningkatan akurasi dari 75.42% menjadi 82.357%[10]. Kemudian, penelitian yang dilakukan oleh Prananda Antinasari dkk yang melakukan analisis sentimen opini Film Pada Dokumen *Twitter* Berbahasa Indonesia Menggunakan *Naive Bayes* Dengan Perbaikan Kata Tidak Baku.pada penelitian ini menggunakan normalisasi *Levenshtein Distance* yang menunjukan hasil peningkatan akurasi dengan nilai akurasi sebesar 98.33% dari 91,67%.[11].

Pada Penelitian yang dilakukan oleh M. Adnan Nur membahas perbandingan antara dua metode pengukuran kesamaan string, yaitu *Levenshtein Distance* dan *Jaro-Winkler Distance*, dalam konteks koreksi kata sebagai bagian dari pra-pemrosesan analisis sentimen pada teks dari pengguna Twitter dimana hasil pengujian menunjukkan bahwa Metode *Levenshtein Distance* menghasilkan nilai tertinggi *accuracy* 72,40%, *recall* 72.07% dan *f1-score* 80,46% sedangkan *Jaro-Winkler Distance* hanya memiliki nilai *accuracy* 70%, *recall* 69,87% dan f1score 79,11%. Hal ini menunjukkan bahwa metode Levenshtein Distance lebih optimal digunakan sebagai koreksi kata dalam *Pre-Processing*[12]. Penelitian yang dilakukan oleh Fahmi Reza Prasastio dkk yang Menerapkan perbaikan kata *Levenshtein Distance* untuk *Pre-Processing* dan algoritma *Naïve Bayes* dalam melakukan analisis sentimen komentar masyarakat tentang vaksin Covid-19 dimana Akurasi pengujian menggunakan data uji lama yang berjumlah 479 data meningkat dari 61% menjadi 71% dan pengujian dengan data uji baru yang berjumlah 100 data akurasi meningkat dari 59% menjadi 66%[13].

Penelitian ini akan dilakukan analisis sentimen menggunakan metode *Lexicon-Based* pada komentar instagram untuk mengklasifikasikan sentimen *Cyberbullying* atau *nonCyberbullying* dengan penggunaan *word normalization* dan algoritma *Levenshtein Distance* untuk melakukan *Typo Checking*.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan maka Bagaimana pengaruh *Word Normalization* dan *Typo Checking* menggunakan algoritma *Levenshtein Distance* pada perbaikan kata tidak baku terhadap *Cyberbullying* di komentar instagram.

1.3 Maksud dan Tujuan

Berdasarkan rumusan masalah maka maksud dari penelitian ini adalah mengimplementasikan Word Normalization dan *Typo Checking* menggunakan algoritma *Levenshtein Distance* dalam perbaikan kata tidak baku terhadap deteksi *Cyberbullying* di komentar instagram.

Berdasarkan rumusan masalah maka tujuan utama pada penelitian ini adalah mengukur pengaruh *Pre-Processing Word Normalization* dan *Typo Checking* menggunakan algoritma *Levenshtein Distance* terhadap Perbaikan kata tidak baku terhadap *Cyberbullying* di komentar instagram.

1.4 Batasan Masalah

1. Penelitian menggunakan Dataset dari Kaggle *Cyberbullying* Bahasa Indonesia dari Cita Tiara Hanni berisi field nama instagram, komentar, Kategori, Tanggal Posting dan nama akun Artis/Selebgram[14].
2. Dataset yang akan digunakan adalah 650 dalam bahasa indonesia yang diambil dari *Kaggle*.
3. Penelitian menggunakan kamus *Indonesian sentiment* (Inset) lexicon.
4. Data diolah menggunakan *Word Normalization* dan *Levenshtein Distance*

5. *Inset Lexicon* menggunakan Dataset dari Github Analysis Sentimet PTM Terbatas with Inset Lexicon and Edit Distance dari I Kadek Arya Budi Artana[15].
6. Kamus KBBI menggunakan Dataset dari Github Muhammad Choirul Anwar Python. Skripsi.[27]
7. Kamus Singkatan menggunakan Dataset dari Github Muhammad Choirul Anwar Python. Skripsi[27]
8. Metode yang digunakan pada pengklasifikasian menggunakan *Lexicon-Based*
9. Pengklasifikasian komentar berdasarkan dua kategori yaitu *Bullying* dan *non-Bullying*
10. Koreksi salah ejaan dibatasi hanya memperbaiki kata non-word
11. Perbaikan kata tidak baku dibatasi hanya melakukan bahasa Indonesia

1.5 Metodologi Penelitian

Metodologi Penelitian menjelaskan rancangan dan prosedur penelitian yang akan dilaksanakan dan Penelitian ini akan menggunakan pendekatan eksperimental dengan memanfaatkan metode utama, yaitu *Lexicon-Based*, untuk analisis sentimen *Cyberbullying* pada komentar Instagram seperti pada Gambar 1.1 Metode Penelitian dibawah ini.

Gambar 1. 1 Metode Penelitian

Pada Gambar 1.1 menjelaskan langkah langkah untuk melakukan metode penelitian sebagai berikut

1.5.1 Pengumpulan Data

Penelitian ini menggunakan dataset yang didapat dari website kaggle dataset yang digunakan dalam penelitian ini sebanyak 650 komentar di instagram[12]

1.5.2 Analisa

a. Analisis Data

   Pada tahap ini mengamati kata yang tidak baku pada data yang didapat dari website *kaggle* . Data tersebut merupakan komentar instagram yang mengandung *Bullying* dan *Non-bullying*.

b. Analisis Metode

Analisis metode akan mencakup penggunaan *Word Normalization*, *Levenshtein distance* dan *Confusion Matrix* dalam kebutuhan metode *Lexicon-Based* pada penelitian ini.

1 *Word Normalization*

Dalam metode *Lexicon-Based*, *Word Normalization* dapat membantu dalam memperbaiki tidak baku dalam komentar dengan mengubahkan menjadi kata baku atau bentuk standar dalam kamus.

2 *Levenshtein Distance*

Levenshtein Distance dapat memperbaiki kesalahan pengejaan kata dengan mengambil jarak terkecil dari perbandingan kata-kata pada komentar terhadap setiap kata dalam KBBI.

3 *Confusion Matrix*

Confusion Matrix membantu dalam mengukur pengaruh *Word Normalization* dan *Levenshtein Distance* terhadap kinerja sistem klasifikasi *Lexicon-Based*. Dengan menggunakan *Confusion Matrix*, dapat diketahui bagaimana perbaikan kata tidak baku menggunakan Word normalization dan *Levenshtein Distance* mempengaruhi tingkat akurasi model.

Dengan mengintegrasikan Word normalization dan *Levenshtein Distance* dalam metode *Lexicon-Based* dapat meningkatkan akurasi. *Word normalization* membantu dalam mengurangi perbaikan kata tidak baku, sementara *Levenshtein Distance* membantu perbaiki salah ejaan.

1.5.3 Implementasi dan Pengujian

Pada tahap ini dilakukan untuk mensimulasikan kinerja dari algoritma sesuai dengan konsep dan scenario yang telah ditentukan sebelumnya. Simulasi dilakukan dengan membangun antarmuka dengan dimulai input komentar Instagram pada Tahap *Pre-Processing* termasuk *Word Normalization* dan *Levenshtein Distance*. *Word*

Normalization dilakukan memperbaiki kata tidak baku pada komentar kemudian Typo *Checking* menggunakan Algoritma *Levenshtein Distance* dengan membandingkan pada setiap kata pada komentar dengan setiap kata di KBBI berdasarkan *Threshold*. Kemudian nanti keluar kata-kata rekomendasi dan pengguna dapat memilih kata dari rekomendasi tersebut untuk diproses pada *Pre-Processing* selanjutnya. Hasil *Pre-Processing* tersebut digunakan pada klasifikasi *Lexicon-Based* yang nanti akan memberikan informasi skor polaritas kata sesuai *Inset Lexicon* dan hasil sentimen *Bullying* atau *Non-Bullying*. *Confusion Matrix* untuk menghitung *Presicion, Recall*, dan *Accuracy*.

1.5.4 Evaluasi Performa

Tahap ini dilakukan evaluasi perbandingan dari Pengujian menggunakan Tahap *Word Normalization* dan *Leveishtein Distance* dengan Pengujian Tanpa *Word Normalization* dan *Leveishtein Distance*. Hal ini untuk menganalisa seberapa pengaruh *Word Normalization* dan *Leveishtein Distance* untuk Klasifikasi Lex*icon-Based* dalam tingkat akurasi.

1.6 Sistematika Penulisan

BAB 1 PENDAHULUAN

Bab ini berisi penjelasan mengenai latar belakang masalah, rumusan masalah, maksud dan tujuan, batasan masalah, metodologi penelitian serta sistematika penulisan yang dimaksudkan agar dapat memberikan gambaran umum pada bab ini.

BAB 2 TINJAUAN PUSTAKA

Bab ini membahas mengenai landasan teori yang digunakan untuk menganalisis teori yang dipakai dalam data penelitian yaitu *Cyberbullying*, Analisis Sentimen, Perbaikan Kata tidak Baku, *Typo Checking, Pre-Processing, Case Folding, Data Cleaning, Tokenizing, Word Normalization, Typo-Checking, Stopword Removal*,dan *Levenshtein Distance*, Klasifikasi *Lexicon-Based, Confusion Matrix*.

BAB 3 ANALISIS DAN PERANCANGAN

Bab ini membahas tentang analisa seperti analisis masalah, analisis data masukkan, gambaran umum sistem, analisis metode, *Pre-Processing* termasuk *Word Normalization* dan *Typo Checking* menggunakan *Levenshtein Distance*, klasifikasi *Lexicon-Based, Confusion Matrix.*

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini membahas tentang implementasi sistem pada perangkat lunak, pengujian sistem, pengujian akurasi serta kesimpulan dari pengujian sistem .

BAB 5 KESIMPULAN DAN SARAN

Bab ini membahas tentang kesimpulan yang didapatkan dari hasil penelitian serta saran yang dapat diimplementasikan untuk penelitian selanjutnya.