BAB 1

PENDAHULUAN

1.1 Latar Belakang Masalah

Pemanfatan Teknologi Informasi dan Komunikasi (TIK) kini termasuk internet oleh masyarakat, sudah semakin masif. "Survei Indikator TIK 2015 untuk Rumah Tangga dan Individu" yang dilakukan oleh Kominfo (2015) menyimpulkan bahwa akses rumah tangga terhadap internet mengalami peningkatan pesat. Proporsi rumah tangga yang telah mengakses internet pada tahun 2015 mencapai 35,1% atau setara dengan 22,8 juta rumah tangga. Angka ini jauh lebih tinggi dibanding dua tahun sebelumnya yang baru mencapai 19,6% dan 22,2%. Penelitian ini juga menemukan bahwa aktivitas yang dilakukan sebagian besar pengguna internet adalah membuka situs jejaring sosial, mengirim pesan melalui *instant messaging* (termasuk *chatting*), dan mencari informasi mengenai barang dan jasa (Juditha, 2017).

Pemanfaatan media sosial dan situs berita yang cenderung meningkat dari tahun ke tahun ini menimbulkan fenomena baru. Setiap orang bebas mengungkapkan apa saja melalui akun media sosial, bahkan berita- berita pada situs berita dengan mudah dibagikan ke media sosial dan kemudian dapat dikomentari oleh netizen lainnya. Bahkan kini dalam situs berita online pun disiapkan ruang komentar untuk para pembaca. Berita-berita ini kemudian ditanggapi secara beragam oleh netizen di ruang komentar baik itu positif, negatif, maupun netral. Namun hal ini juga mendatangkan masalah baru di mana praktik *hatespeech* atau ujaran kebencian juga tumbuh pesat melalui medium ini (Juditha, 2017).

Hatespeech adalah segala tindakan dan usaha baik langsung maupun tidak langsung yang didasarkan pada kebencian atas dasar suku, agama, aliran keagamaan, keyakinan/ kepercayaan, ras, antar golongan, warna kulit, etnis, gender, kaum difabel, dan orientasi seksual (HAM, 2015). *Hatespeech* sangat berbahaya karena bisa membuat *stereotyping*/ pelabelan, stigma, pengucilan, diskriminasi dan kekerasan. Pada tingkat yang paling mengerikan bisa menimbulkan kebencian kolektif pembantaian etnis, pembakaran kampung, pengusiran, pembumihangusan kampung atau pemusnahan terhadap kelompok yang menjadi sasaran ujaran kebencian (HAM, 2015).

Upaya mengatasi ujaran kebencian secara otomatis dengan memanfaatkan algoritma pembelajaran mesin telah dilakukan oleh beberapa peneliti yang dibangun saat ini masih dikhususkan untuk bahasa tertentu, seperti bahasa Inggris, Jerman, Belanda. Beberapa kesulitan yang ditemukan dalam menentukan ujaran kebencian adalah perbedaan pendapat antara manusia yang memberikan label sebuah tulisan sebagai ujaran kebencian atau bukan. Hal ini menunjukkan potensi kesalahan klasifikasi pada algoritma pembelajaran mesin yang nantinya akan dilatih berdasarkan pelabelan manusia (Herwanto et al., 2019).

Klasifikasi text merupakan masalah penting yang sulit dari komputasi linguistik dan *Natural Language Processing*. Berbagai jenis *neural network* seperti *deep learning, convolutional, recurrent, Long Short Term Memory* (LSTM), dan yang lainnya biasanya digunakan untuk klasifikasi teks, sering kali mencapai kesuksesan yang signifikan (Zolotov & Kung, 2017). Penelitian terdahulu (Joulin et al., 2016) menunjukan bahwa hasil yang sebanding dapat dicapai dengan

menggunakan model klasifikasi linear sederhana yang dikombinasikan dengan model ekstraksi fitur fastText (Zolotov & Kung, 2017).

Penelitian sebelumnya oleh Ibrohim & Budi (2019), telah berhasil membuat *dataset* untuk mengindentifikasi bahasa kasar dan *hatespeech* dalam bahasa Indonesia di *platform Twitter*. Dataset dibuat berdasarkan hasil *crawling* pada *platform Twitter.* Dalam penelitian juga dilakukan eksperimen dengan menggunakan metode-metode *unigram words, Random Forest Decision Tree* (RFDT), dan *Label Power-set* (LP) merupakan kombinasi terbaik dari fitur, pengklasifikasi, dan metode transformasi data untuk semua skenario dilakukan. Berdasarkan eksperimen yang dilakukan dihasilkan tingkat akurasi sebesar 77,36% untuk melakukan klasifikasi multi-label untuk mengidentifikasi bahasa kasar dan *hatespeech* tanpa mengindentifkasi target, kategori, dan tingkat ujaran kebencian. Di sisi lain, 66,12% untuk pada saat melakukan klasifikasi multi- label untuk mengindentifkasi bahasa kasar dan ujaran kebencian termasuk mengidentifikasi target, kategori dan tingkat ujaran kebencian (Ibrohim & Budi, 2019).

Penelitian yang serupa tentang klasifikasi *hatespeech* di Indonesia dengan menggunakan algoritma fastText dilakukan oleh Herwanto et al. (2019), menjelaskan bahwa model fastText memiliki performa yang lebih baik dalam permasalahan klasifikasi biner meskipun tidak menunjukan peningkatan yang signifikan dari penelitian sebelumnya yang menggunakan *Long Short Term Memory* (LSTM), *Continuous Bag-of-Words* (CBOW) dan skipgram yang dapat mencapai akurasi kinerja lebih dari 85% (Herwanto et al., 2019).

Pada penelitian lain oleh Polignano dan Basile (2018), telah dilakukan perbandingan metode klasifikasi *Logistic Regression*, *Support Vector Classification*, *K-nearest neighbors*, *Decision Tree, Random Forest*, dan *Multilayer Perceptron classifier* dengan *featured extraction* TF-IDF untuk klasifikasi *Italian hatespeech* (Polignano & Basile, 2018). Hasil akhir penelitian yang dinilai dari F1-Score pada tabel berikut ini.

Tabel 1.1 Hasil perbandingan metode pembelajaran mesin.

Algorithm	Macro F1 Score
LR	0.780444109
SVC-rbf – C =1	0.789384136
SVC-poly 2 – C=1	0.758599844
SVC-poly 3 – C=1	0.667374386
KNN – 3	0.705064332
KNN – 4	0.703990867
KNN – 10	0.687719117
KNN – 20	0.663451598
DT	0.68099986
RF - 50	0.75219596
RF - 100	0.764247578
RF - 300	0.787778421
RF - 500	0.768494151
MLP - 1000	0.766835616
MLP - 2000	0.791230474
MLP - 3000	0.76952709

Merujuk dari data Tabel 1.1 dapat dilihat bahwa *Multilayer Perceptron* meraih akurasi sebesar 0.791230474, ini membuktikan bahwa *Multilayer Perceptron* lebih akurat dibandingkan dengan algoritma *Logistic Regression*, *Support Vector Classification*, *K-nearest neighbors*, *Decision Tree, Random Forest*.

Berdasarkan kajian penelitian yang dilakukan sebelumnya maka dalam penelitian ini metode yang digunakan mengadopsi metode yang digunakan oleh

Polignano dan Basile yaitu dengan *Multilayer Perceptron* yang dikombinasikan dengan fastText sebagai fitur ekstraksinya guna mengidentifikasi ujaran kebencian dalam bahasa Indonesia pada platform Twitter dengan dataset yang telah dibuat oleh Ibrohim & Budi (2019).

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan sebelumnya, maka dapat dirumuskan permasalahan yang diteliti yaitu :

1. Bagaimana cara mengimplementasi metode *Multilayer perceptron* dengan fastText sebagai fitur ekstraksinya untuk mengidentifikasi ujaran kebencian dalam bahasa Indonesia pada platform Twitter?
2. Bagaimana performa metode *Multilayer perceptron* dengan fastText sebagai fitur ekstraksinya untuk mengidentifikasi ujaran kebencian dalam bahasa Indonesia pada platform Twitter?

1.3 Batasan Masalah

Pada bagian ini dijelaskan mengenai berbagai batasan masalah yang telah ditentukan untuk penelitian ini. Berbagai batasan yang telah ditentukan untuk penelitian ini adalah :

1. Dataset yang digunakan adalah dataset yang telah dibuat oleh Ibrohim & Budi (2019) termasuk identifikasi target, kategori, dan tingkat ujaran kebencian dengan menggunakan pedoman anotasi.

2. Dalam dataset acuan, tersedia informasi multilabel untuk mengidentifikasi bahasa kasar dan *hatespeech* tanpa mengindentifkasi target, kategori, dan tingkat ujaran kebencian. Dalam penelitian yang dilakukan proses klasifikasi hanya dilakukan tanpa mengidentifikasi target, kategori, dan tingkat ujaran kebencian.
3. Bahasa yang digunakan untuk klasifikasi ujaran kebencian adalah Bahasa Indonesia pada platform media sosial Twitter.
4. Performa metode diukur dengan metrik perhitungan F1 Score.

1.4 Tujuan Penelitian

Berdasarkan latar belakang masalah serta rumusan masalah yang telah dijelaskan, maka tujuan penelitian ini adalah :

1. Mengimplementasikan metode *Multilayer Perceptron* dengan fastText *word embedding* untuk klasifikasi ujaran kebencian atau *hatespeech* dalam bahasa Indonesia pada platform Twitter
2. Mengevaluasi performa metode *Multilayer Perceptron* dengan fastText *word embedding* pada platform Twitter dengan menggunakan *F1 Score*.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah untuk pengembangan ilmu mengenai klasifikasi *text* untuk mengidentifikasi ujaran kebencian dalam bahasa Indonesia dengan metode *Multilayer Perceptron* dengan fastText *word* embedding dan juga dapat dikembangan lebih lanjut untuk mengatasi ujuran kebencian yang ada di sosial media di Indonesia.

1.6 Sistematika Penulisan

Sistematika penulisan laporan penelitian disusun dan dibagi atas 5 (lima) bab sebagai berikut.

1. BAB 1 PENDAHULUAN

   Bab pertama ini menjelaskan latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

2. BAB 2 LANDASAN TEORI

   Bab kedua membahas teori-teori dan konsep dasar yang mendukung dalam penelitian ini, seperti *Text Classification* dan *Text Processing*, FastText, *Multilayer Perceptron*, Evaluasi Klasifikasi.

3. BAB 3 METODOLOGI DAN PERANCANGAN APLIKASI

   Bab ketiga menjelaskan metode penelitian yang digunakan dan perancangan aplikasi meliputi *Flowchart* Umum Proses *Training* dan *Evaluation*, *Flowchart Modul Data Preparation*, *Flowchart Modul Case Folding*, *Flowchart Modul Transform Alay Sentence, Flowchart Modul Stopwords Removal, Flowchart Modul Tokenizing, Flowchart Modul Stemming, Flowchart Modul FastText Training Model, Flowchart Modul Vectorized Word Embedding, Flowchart Modul Multilayer Perceptron, Flowchart Modul* Aplikasi *Hatespeech Classification*.

4. BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM

   Bab keempat berisi implementasi system yang diikuti oleh data hasil penelitian yang dilakukan.

5. BAB 5 SIMPULAN DAN SARAN

Bab kelima merupakan bab terakhir yang berisi simpulan dari hasil pengujian aplikasi dan juga saran untuk pengembangan aplikasi di masa mendatang.

.