Pemodelan Mesin Pendeteksi Ujaran Kebencian di Sosial Media Indonesia

Made Krisnanda[1]

1Jurusan Pendidikan Teknologi Informasi dan Komunikasi UNIMA)
Email:madekrisnanda@unima.ac.id

***Abstract*— This research discusses the development of a model that can be used to detect hate speech in Indonesia's popular social media. Several methods are collected, analyzed, and implemented to build a framework that is practical and can be used to detect hate speech. The research methodology starts from the study of literature, analysis and design of the model. From the analysis results, N-Gram, Word2vec, LSTM and emotion classification are used as part of the process of the model. This model consists of ten processes that are carried out sequentially so that comments collected through Facebook and Whatsapp social media can be identified when they contain hate speech. The model also considers the applicable law in Indonesia to facilitate legal handling of perpetrators.**

***Keywords*—** Social Media, Deep Learning, LSTM, Indonesia, N-Gram

***Intisari*— Penelitian ini membahas mengenai pengembangan model yang dapat digunakan untuk mendeteksi ujaran kebencian di sosial media populer Indonesia. Beberapa metode dikumpulkan, dianalisa, dan diimplementasikan untuk membangun kerangka kerja yang secara praktis dan dapat digunakan untuk mendeteksi ujaran kebencian. Metodologi penelitian dimulai dari studi pustaka, analisa dan desain model. dari hasil analisa, N-Gram, Word2vec, LSTM dan klasifikasi emosi digunakan sebagai bagian dari pemodelan. Model ini terdiri dari sepuluh proses yang dilakukan secara berurutan sehingga komentar yang dikumpulkan melalui sosial media Facebook dan Whatsapp dapat diidentifikasi bilamana mengandung ujaran kebencian. Model juga mempertimbangkan hukum yang berlaku di Indonesia untuk memudahkan penanganan hukum terhadap pelaku.**

***Kata Kunci*—** Social Media, Deep Learning, LSTM, Indonesia, N-Gram.

I. PENDAHULUAN

Indonesia adalah negara multi etnis yang terdiri dari banyak suku dan ras. Keragaman ras, suku, bahasa, budaya, dan agama merupakan ciri khas serta kelebihan bangsa Indonesia yang membedakannya dengan bangsa lain. Dengan kondisi geografis yang terdiri atas ribuan pulau dan luas laut dengan keragaman masyarakat suku bangsa dan kebudayaannya, Negara Kesatuan Republik Indonesia (NKRI) dihadapkan pada masalah integrasi nasional yang berat dan rumit [13]. Ujaran kebencian adalah salah satu sikap yang dapat mengancam NKRI, dimana sikap ini sering dialami oleh kaum minoritas. Ujaran kebencian seringkali dinyatakan melalui aplikasi sosial media seperti Facebook, Twitter, maupun Whatsapp. Hal ini terjadi karena melalui sosial media pengguna dapat dengan mudah memposting status atau komentar yang dapat disebarkan dengan cepat. Kejadian terakhir terjadi dimana Pemerintah Indonesia harus memutuskan akses internet di Pulau Papua untuk mengurangi dampak dari sosial media terhadap kerusuhan. Facebook sendiri harus menutup 69 akun dan 42 halaman yang terlibat dengan perilaku tidak otentik di Papua barat[3].

Berdasarkan masalah tersebut, maka diperlukan suatu mekanisme untuk mendeteksi komentar ujaran kebencian di sosial media Indonesia yang melanggar hukum. Penelitian ini akan menggunakan *Deep Learning* sebagai model penyelesaian dan bahasa Indonesia sebagai set data karena bahasa dan cara penggunaannya merupakan tantangan besar dalam mengembangkan dan mengklasifikasikan komentar ujaran kebencian[7]. Penelitiaan ini akan mengkaji beberapa metode terbaru di bidang kecerdasan buatan yang dapat menghasilkan sebuah model yang dapat diuji.

II. STUDI LITERATUR

Menurut hukum di Indonesia, seseorang yang membuat tulisan atau gambar untuk ditempatkan, ditempelkan, atau disebarluaskan di tempat umum dengan tujuan menunjukkan rasa benci terhadap etnis atau ras tertentu dapat dikenakan pidana penjara dan/atau denda. Undang-undang nomor 11 tahun 2008 tentang Informasi dan Transaksi Elektronik membahas hukum pidana mengenai penyebaran diskriminasi dan ujaran kebencian melalui media elektronik[10].

Rasa benci harus mengandung diskriminasi dan pernyataan atau opini emosional sehingga dapat dibedakan dengan aksi lainnya[11]. Penelitian ini juga mengelompokkan emosi menjadi dua yaitu positif (senang) dan negatif (marah, jijik, takut dan sedih). Pernyataan kebencian terhadap ras tertentu dapat dibagi menjadi dua domain. Domain pertama yaitu implisit atau eksplisit, dan domain kedua secara langsung atau tidak langsung[7].

Sebuah penelitian mengusulkan *Deep Learning* sebagai solusi dengan menggunakan Convolutional Neural Networks (CNN) dan *Long Short-Term Memory* (LSTM) sebagai Jaringan Syaraf Tiruan dan algoritma *Support Vector Machine* untuk mengidentifikasi komentar diskriminasi [16]. Kehandalan CNN dan LSTM juga didukung oleh penelitian dalam mengidentifikasi ujaran kebencian di sosial media Twitter[5]. Sedangkan penelitian lain menggunakan lima

tahapan dan empat belas dimensi untuk mempermudah identifikasi komentar negatif [11].

III. KOMENTAR UJARAN KEBENCIAN DI SOSIAL MEDIA INDONESIA

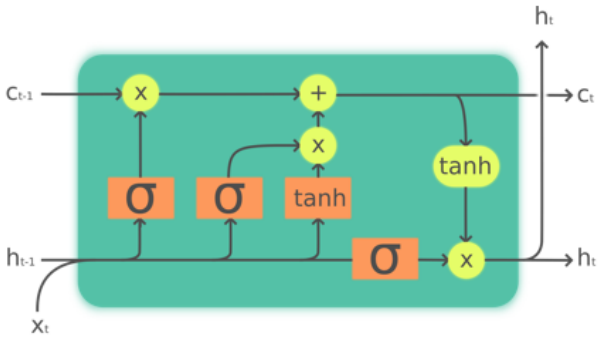
Komentar kebencian di Indonesia seringkali menyebabkan kerugian yang sangat besar dan mengancam keutuhan kesatuan negara. Salah satu contoh adalah kerusuhan di Wamena [8] dan Jayapura [9]. Kerusuhan ini berawal dari komentar rasis di Surabaya [1], yang menyebar dengan cepat sehingga menimbulkan kerusuhan dan menyebabkan baik penduduk asli Papua maupun pendatang banyak yang kembali ke daerah asalnya. Hal ini sangat disayangkan mengingat komentar kebencian tersebut hanya dilakukan beberapa orang dan tidak mewakili penduduk suatu daerah tertentu. Polisi sendiri sedang memburu para pelaku yang terlibat dalam penyebaran konten kebencian melalui sosial media yang masih banyak ditemukan. Konten kebencian sendiri seringkali hanya dilakukan sebagian orang yang disebut "*buzzer"* untuk kepentingan politik [4].

Seandainya komentar kebencian dapat dideteksi lebih dini tentu saja dampak yang ditimbulkan dapat diminimalisir. Sosial media yang menjadi saluran untuk menyebarkan komentar kebencian juga dapat segera bertindak sebelum komentar tersebut dibaca lebih banyak orang. Selain itu polisi juga dapat dengan mudah mengidentifikasi dan menindak pelaku ujaran kebencian sesuai hukum yang berlaku.

IV. METODOLOGI PENELITIAN

Metodologi penelitian yang digunakan dimulai dari studi literatur, dimana akan dipelajari beberapa metode yang dapat digunakan untuk membuat model pendeteksi ujaran kebencian. Selanjutnya dilakukan analisa terhadap metode – metode yang telah dipelajari. Salah satu arsitektur yang menarik untuk mendeteksi ujaran kebencian adalah Long Short Term Memory (LSTM) (Gbr 1). Konversi teks yang tidak terstruktur seperti (Kantung N-Gram) juga dapat digunakan untuk membagi kalimat menjadi urutan kata dan menimbang bobotnya menggunakan algoritma TF-IDF. Beberapa metode ini dianalisa kehandalannya dalam model yang diusulkan.

J. LSTM



Gbr. 1 Arsitektur LSTM

LSTM memiliki bentuk rangkaian modul jaringan saraf dimana terdapat empat lapisan jaringan saraf yang saling berinteraksi. Lapisan pertama dalam LSTM adalah memutuskan informasi apa yang akan dibuang. Keputusan ini dibuat oleh lapisan sigmoid yang disebut "***forget gate layer***."Hal ini diwakili oleh simbol (h{t-1}) dan (x_t), untuk menghasilkan angka antara (0) dan (1) untuk setiap angka dalam keadaan sel (C{t-1}). Nilai (1) mewakili keputusan "simpan sepenuhnya" sementara nilai (0) mewakili keputusan "hapus sepenuhnya". Lapisan selanjutnya memutuskan informasi baru apa yang akan disimpan. Lapisan ini memiliki dua bagian. Pertama, lapisan sigmoid yang disebut "lapisan gerbang input" memutuskan nilai mana yang akan diperbarui. Bagian kedua adalah lapisan **tanh** untuk membuat sebuah vektor nilai kandidat baru (~{C}t) yang bisa ditambahkan dalam kondisi sel. Selanjutnya keduanya akan digabungkan untuk membuat pembaruan ke kondisi sel. Untuk mendapatkan keadaan sel yang baru (Ct), kondisi sel yang lama (C{t-1}) dikalikan dengan (f_t) untuk membuang hal yang tidak perlu dan menambahkan (i_t) dikalikan dengan (~Ct). Sehingga menghasilkan kandidat keadaan sel yang baru. Pada lapisan akhir, diputuskan keluaran yang dihasilkan. Keluaran akan didasarkan pada status sel terakhir dengan tambahan penyaringan. Penyaringan dilakukan untuk menambahkan informasi yang relevan dan lebih sesuai dengan kasus yang ingin diselesaikan, dalam hal ini mendeteksi ujaran kebencian.

K. Kantung N-Gram

Kantung N-gram adalah perpanjangan alami kantung kata-kata. N-gram hanyalah urutan n token (kata-kata). Sebuah kalimat dapat dibagi menjadi 1,2, atau 3 kesatuan arti (1,2,3-gram). Kantung n-gram menangkap konteks di sekitar setiap kata.

L. Algoritma TF-IDF

Term Frequency — Inverse Document Frequency (TF–IDF) adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Secara sederhana, metode TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen. Rumus algoritma ini dapat dilihat sebagai berikut:

$$W_{dt} = tf_{dt} * IDF_t \qquad (1)$$

Dimana:
d = dokumen ke-d
t = kata ke-t dari kata kunci
W = bobot dokumen ke-d terhadap kata ke-t
tf = banyaknya kata yang dicari pada sebuah dokumen
IDF = *Inversed Document Frequency*
IDF = $\log_{10}(D/df)$
D = total dokumen
df = banyak dokumen yang mengandung kata yang dicari

V. PEMODELAN

Dari hasil studi literatur didapatkan bahwa setidaknya diperlukan *Dataset, Feature Extraction, Word2vec,* dan klasifikasi komentar. Penjelasan dari beberapa konsep ini adalah sebagai berikut:

A. Dataset

Dataset yang digunakan untuk pelatihan dan pengujian diekstraksi dari Facebook dan Whatsapp. Dataset yang tersedia terdiri dari komentar yang berlabel Kebencian dan Netral. Selain itu dataset juga memerlukan pra-proses untuk mengubah karakter alfabet menjadi huruf kecil, dan menghapus simbol tanda baca, URL, karakter non-Indonesia, karakter berulang dan angka. *Emoticon* dikonversi ke teks yang setara, dan masalah kesalahan ejaan ditangani menggunakan pemeriksa ejaan.

B. Feature Extraction

Feature extraction adalah representasi teks dalam bentuk vektor untuk pengklasifikasi pembelajaran mesin. Untuk mendeteksi teks ofensif dalam komentar. Fitur ini menggunakan n-gram (bigram, unigram, dan trigram). Setiap fitur ditimbang oleh TF-IDF-nya. Dari fitur-fitur ini, dapat dilakukan klasifikasi untuk mengkategorikan komentar bermuatan ujaran kebencian.

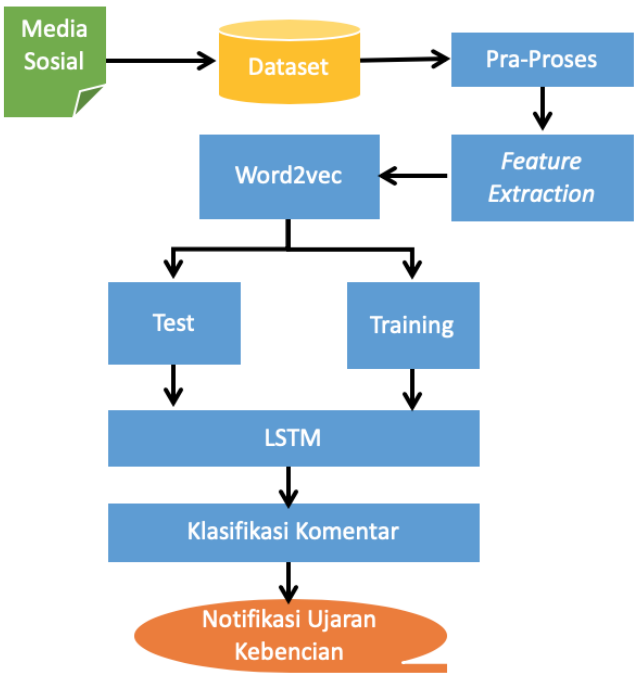
C. Word2vec

Lapisan *Embedding* adalah representasi input ke regresi logistik dan LSTM. Lapisan ini mengkodekan setiap kata dalam kosakata untuk digunakan dalam model penyelesaian. Model Word2vec digunakan untuk menghasilkan vektor untuk setiap kata dalam ruang dimensi. Arsitektur kode model Word2vec digunakan untuk pelatihan dari kumpulan besar komentar.

D. Klasifikasi Komentar

Klasifikasi teks diperlukan untuk mengkategorikan bilamana komentar bersifat positif, negatif, atau netral. Hasil klasifikasi ini diperlukan sebagai pemberi notifikasi kepada aparat penegak hukum atau pihak lain yang berkepentingan untuk menjaga perdamaian.

Berdasarkan analisa beberapa konsep diatas maka model mesin pendeteksi ujaran kebencian dapat dilihat pada Gbr 2.



Gbr. 2. Model Mesin Pendeteksi Ujaran Kebencian

Model dimulai dari pengumpulan komentar dari sosial media yang populer seperti Facebook dan Whatsapp. Komentar ini kemudian dipilih untuk disimpan dalam dataset. Praproses digunakan untuk menghilangkan karakter yang tidak baku, sehingga dapat mudah diklasifikasikan melalui *Feature Extraction.* Word2vec digunakan untuk menghasilkan vektor untuk setiap kata dalam ruang dimensi, sehingga dapat digunakan untuk *test* dan *training* pada arsitektur LSTM. Keluaran dari LSTM diklasifikasikan agar dapat memberi notifikasi ke pihak berwenang bilamana sebuah komentar mengandung ujaran kebencian atau tidak.

VI. KESIMPULAN

Model mesin pendeteksi ujaran kebencian dihasilkan dari studi pustaka beberapa jurnal yang menggunakan metode yang telah diuji. bagaimanapun juga model ini perlu diuji lebih lanjut untuk ujaran kebencian, khususnya yang menggunakan Bahasa Indonesia. Pengembangan lebih lanjut juga diperlukan untuk menggunakan atau mengkombinasikan beberapa model prediksi teks diluar LSTM.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Program Studi Teknik Informatika Universitas Negeri Manado untuk fasilitas yang diberikan sehingga penelitian ini dapat diselesaikan.

DAFTAR PUSTAKA

[1]. Ayuningtias, Rita *Polri Buru Penyebar Konten Rasis Pemicu Kerusuhan Manokwari.* Aug.2019. Accessed on : October 2019 [Online]. Available

:https://www.liputan6.com/news/read/4041247/polri-buru-penyebar-konten-rasis-pemicu-kerusuhan-manokwari

[2]. Chang, J. *et al.* (2010) "EPluribus : Ethnicity on social networks," *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pp. 18–25.

[3]. CNN Indonesia *Facebook dan Instagram Hapus Akun Terkait Papua Barat.*, Oct 2019, Accessed on : October 2019. Available: https://www.cnnindonesia.com/teknologi/20191004180933-185-436840/facebook-dan-instagram-hapus-akun-terkait-papua-barat

[4]. CNN Indonesia *Riset Oxford : Buzzer Indonesia Dibayar Rp.1-50 Juta Giring Isu*, Oct 2019, Accessed on : October 2019. Available: https://www.cnnindonesia.com/nasional/20191006170414-12-437255/riset-oxford-buzzer-indonesia-dibayar-rp1-50-juta-giring-isu

[5]. Dias, D. S., Welikala, M. D. and Dias, N. G. J. (2019) "Identifying racist social media comments in Sinhala language using text analytics models with machine learning," in *18th International Conference on Advances in ICT for Emerging Regions, ICTer 2018 - Proceedings*. doi: 10.1109/ICTER.8615492.

[6]. Hayat, M. K. *et al.* (2019) "Towards Deep Learning Prospects: Insights for Social Media Analytics," *IEEE Access*. IEEE, 7(May), pp. 36958–36979. doi: 10.1109/ACCESS.2019.2905101.

[7]. Hemker, K. and Schuller, B. (2018) "Data Augmentation and Deep Learning for Hate Speech Detection."

[8]. *Ini Daftar Kerusakan Akibat Kerusuhan di Wamena-Papua* (no date). Available at: https://regional.kompas.com/read/2019/09/25/11482551/ini-daftar-kerusakan-akibat-kerusuhan-di-wamena-papua (Accessed: October 10, 2019).

[9]. Isidorus, Robert *Kerugian Akibat Kerusuhan di Kota Jayapura sekitar 19 M*, Sept 2019. Accessed on : October 2019. Available:https://www.beritasatu.com/nasional/573436/kerugian-akibat-kerusuhan-di-kota-jayapura-sekitar-rp-19-m

[10]. Kementrian Keuangan Undang-undang Republik Indonesia Nomor 11 Tahun 2008. April 2008. Accessed on : October 2019. Available: http://www.jdih.kemenkeu.go.id/fullText/2008/11TAHUN2008UU.htm

[11]. Martins, R. *et al.* (2018) "Hate speech classification in social media using emotional analysis," *Proceedings - 2018 Brazilian Conference on Intelligent Systems, BRACIS 2018*, (April 2019), pp. 61–66. doi: 10.1109/BRACIS.2018.00019.

[12]. M. Wegmuller, J. P. von der Weid, P. Oberson, dan N. Gisin, "Highresolution fiber distributed measurements with coherent OFDR," *Proc. ECOC'00*, 2000, paper 11.3.4, hal. 109.

[13]. Susilowati, E. and Masruroh, N. N. (2018) "Merawat Kebhinekaan Menjaga Keindonesiaan: Belajar Keberagaman dan Kebersatuan dari Masyarakat Pulau," *Jurnal Sejarah Citra Lekha*. doi: 10.14710/jscl.v3i1.17856.

[14]. Suwandi, Dhias *Ini Daftar Kerusakan Akibat Kerusuhan di Wamena Papua*, Sept 2019. Accessed on : October 2019. Available : https://regional.kompas.com/read/2019/09/25/11482551/ini-daftar-kerusakan-akibat-kerusuhan-di-wamena-papua

[15]. Tulkens, S. *et al.* (2016) "The Automated Detection of Racist Discourse in Dutch Social Media," in *Computational Linguistics in the Netherlands Journal*.

[16]. Zhang, Z., Robinson, D. and Tepper, J. (2018) "Hate Speech Detection Using a Convolution-LSTM Based Deep Neural Network," *ESWC 2018: The semantic web*.