SEMINAR NASIONAL TEKNOKA

Vol. 3, 2018
ISSN No. 2502-8782

Studi Algoritma Liner Support Vector Machine pada Deteksi Ujaran Kebencian Berbahasa Indonesia

Alfi Ramdhani

Universitas Muhammadiyah Prof. DR. HAMKA, Jalan Tanah Merdeka No. 6 B Kp. Rambutan Jakarta Timur 13830
Website:alfiramdhani.com, E-mail: alfiramdhani@uhamka.ac.id

Abstrak *- Ekspresi ujaran kebencian merupakan suata fenomena yang berkembangan di dunia masyarakat era modern ini, banyak dari pengguna media sosial memanfaatkannya untuk mengekspresikan perasaan mereka maupun kehidupannya. Namun dari fenomena ini semua berdampak kepada lingkungan masyarakat yang terkesan sangat bebas mengekspresikan ujaran kebencian dan berujung kepada tindakan kejahatan, entah darimana asal-usul penyebab terjadinya, bisa jadi karena pengaruh provokasi atau hal-hal lainnya yang persuasif. Maka dari itu tujuan penelitian melakukan studi terhadap algoritma Linear Support Vector Machine dalam melakukan deteksi ujaran kebencian berbahasa Indonesia. Metode yang digunakan adalah algoritma Linear Support Vector Machine dengan feature Word N Gram. Dari hasil percobaan, diperoleh hasil evaluasi akurasi sebesar 86.55 % dengan metode 10-fold cross validation*

Kata Kunci : *Kualitas Pepaya California, RGB, Algoritma Min-Max, Java, Android..*

1 Pendahuluan

Perkembangan Teknologi Informasi yang begitu pesat memberikan dampak terhadap masyarakat, sekarang ini pengguna media sosial meningkat drastis. Facebook sebagai media sosial paling populer, pada kuartal ketiga 2018 facebook memiliki 2.27 Milyar user aktif[1]. Ini menunjukkan bahwa media sosial menjadi media komunikasi yang penting hari ini. namun belakangan ini media sosial menjadi media untuk mengutarakan ujaran kebencian yang berdampak perpecahan dan perselisihan.

Ujaran kebencian adalah bentuk komunikasi apapun yang meremehkan seseorang atau kelompok berdasarkan karakteristik seperti ras, etnisitas, jenis kelamin, orientasi seksual, kebangsaan, agama, atau karakteristik lainnya [1]. 11 kriteria ujaran kebencian diantaranya adalah: uses of a sexist or racial slur, attack a minority, promotes hate speech or violent crime, blatantly misrepresents truth, shows support of problematic hashtags, defends xenophobia or sexism, and contains a screen name that is offensive [2].

Ujaran kebencian belakangan ini, terutama mengenai ras dan agama, menjadi bentuk kejahatan *online* yang paling banyak dilaporkan pada tahun 2016, menurut Polisi Indonesia. Petugas kepolisian di Indonesia mengklaim setidaknya 5 kasus dilaporkan setiap hari, yang berarti ada sekitar 150 setiap bulannya [3]. Polisi juga mengatakan bahwa menangani penjahat *cyber* tidaklah mudah sehingga fasilitas dan sumber daya manusia sangat dibutuhkan. Hal ini membuat pendataan ujaran kebencian otomatis perlu dikembangkan untuk bahasa Indonesia sehingga polisi bisa mendeteksi penyebaran ujaran kebencian dengan cepat.

Salah satu media sosial yang populer di Indonesia adalah Twitter, sebagai microblogging web services memberikan kemudahan kepada penggunanya untuk berbagi pendapat, opini atau ekspresinya serta komunikasi sesama pengguna. Twitter memiliki 336 Juta user aktif[2]. Hal ini bisa dapat dianggap positif karena kita dapat berbagi hal-hal kebaikan dan komunikasi dengan seluruh pengguna yang berada di dunia apabila kita menggunakannya secara bijak, namun akan berbeda ceritanya apabila digunakan pada user yang tidak bertanggungjawab sehingga menimbulkan kejahatan.

Pada beberapa penelitian sebelumnya, deteksi ujaran kebencian cenderung menggunakan bahasa inggris [2], [4]–[6]. Hal ini karena sedang berkembang penelitian deteksi ujaran kebencian di berbagai negara dengan bahasa inggris. Sedangkan penelitian deteksi ujaran kebencian menggunakan bahasa Indonesia masih sangat minim. dan [3] yang melakukan penelitian deteksi ujaran kebencian pada bahasa Indonesia dengan menghasilkan dataset untuk deteksi ujaran kebencian bahasa Indonesia dari Twitter. Karena hal itu lah yang memotivasi penulis untuk ikut mengembangkan penelitian di bidang deteksi ujaran kebencian, dimana ujaran kebencian masih sangat sering terjadi khususnya di media sosial di Indonesia dan berdampak pada perselisihan serta perpecahan.

Pada penelitian ini, penulis melakukan pengujian terhadap algoritma Linear SVM pada studi kasus deteksi ujaran kebencian berbahasa Indonesia. Fokus penelitian ini adalah

performa algoritma dan model dalam klasifikasi ujaran kebencian, sehingga kita dapat mengetahui sejauh mana algoritma mampu mengatasi masalah klasifikasi dan performansi pada deteksi ujaran kebencian berbahasa Indonesia.

2 Dasar Teori

2.1 Natural Language Processing

Natural Language Processing (NLP) merupakan sebuah area penelitian dan aplikasi yang menyelidiki bagaimana komputer dapat digunakan untuk memahami dan memanipulasi teks bahasa alami atau pembicaraan untuk melakukan hal yang berguna. Penerapan NLP termasuk dalam beberapa bidang studi seperti mesin terjemahan, pengolahan teks bahasa alami dan rangkuman, tampilan *user*, pengambilan informasi lintas bahasa dan multibahasa, pengenalan suara, kecerdasan buatan, dan sistem pakar. [7]

2.2 Machine Learning

Pembelajaran mesin atau Machine Learning adalah bidang ilmu komputer yang memberikan komputer kemampuan untuk belajar tanpa secara eksplisit diprogram. Pembelajaran mesin digunakan dalam berbagai tugas komputasi di mana merancang dan memprogram algoritma eksplisit dengan kinerja yang baik. Aplikasi yang termasuk menggunakan machine learning: email filtering, recognition of network intruder atau orang yang tidak bertanggungjawab melakukan pelanggaran data. Salah satu tujuan dasar pembelajaran mesin adalah melatih komputer untuk memanfaatkan data memecahkan masalah yang ditentukan. Sejumlah besar aplikasi pembelajaran mesin seperti pelatihan pengelompokan pada pesan email untuk membedakan antara pesan spam dan non-spam, deteksi penipuan, dll. [8]

2.3 Support Vector Machine

Support vector machine atau dikenal dengan istilah SVM, dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992[9]. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space*. *Hyperplane* pemisah terbaik antara kedua class dapat ditemukan dengan mengukur *margin hyperplane* tersebut. dan mencari titik maksimalnya. [10] [11] [12].

Permasalahan klasifikasi pada teknik *machine learning* SVM dapat diforumulasikan sebagai berikut:

Perlu diperhatikan, bahwa dapat dimasukan

$$\min_{\vec{w}} \tau(\vec{w},\xi) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l}\xi_i \qquad (1)$$
$$y_i(\vec{x_i}.\vec{w}+b) \geq 1-\xi_i, \forall i$$

paramater bias dengan menambah semua vektor data x_n dengan nilai skalar 1 [12]. Masalah optimasi tak dibatasi yang terhubung sebagai berikut:

$$L = \min_w \frac{1}{2}w^T w + C\sum_{n=1}^{N}\max\{1-w^T x_n t_n, 0\} \qquad (2)$$

Persamaan diatas merupakan masalah utama dari L1-SVM, dengan kerugian standar. Karena L1-SVM tidak terdefferiansi maka versi yang popular dikenal dengan L2-SVM dengan minimalisir kerugian standar kuadrat. [12]

$$L = \min_w \frac{1}{2}w^T w + C\sum_{n=1}^{N}\max\{1-w^T x_n t_n, 0\}^2 \qquad (3)$$

L2-SVM dapat diturunkan dan menyebabkan sebuah kerugian besar untuk titik-titik yang melanggar batas. Pada penelitian ini kelas yang digunakanya hanya 2 buah sehingga untuk mendapatkan skor ujaran kebencian dapat dirumuskan sebagai berikut:

$$f(z) = W^t z + b_0 \qquad (4)$$

y merupakan kelas label dari ujaran kebencian. Sehingga ketika $f(z) \geq 0$, maka nilai $y = +1$ dan ketika $f(z) < 0$, maka nilai $y = -1$. Berdasarkan hal tersebut dapat dituliskan sebagai berikut:

$$f(z_i)\begin{cases}\geq 0 & y = +1\\ < 0 & y = -1\end{cases} \qquad (5)$$

Sehingga dapat diketahui keberhasilan klasifikasi jika $y_i f(z_i) > 0$. Pada penelitian ini inisialisasi awal nilai W dilakukan dengan membangkitkan nilai secara acak antara $[-1, +1]$.

3 Metodologi Penelitian

3.1 Pengumpulan Data

Data pada penelitian ini menggunakan dataset Twitter berbahasa Indonesia hasil penelitian sebelumnya [3]. Dataset dapat diunduh secara offline via Github[3]. Pada dataset tersebut terdapat sebanyak 713 baris *tweet* yang berisi ujaran kebencian berbahasa Indonesia dengan kelas label "HS" untuk hate speech dan "Non_HS" untuk bukan hate speech.

Data yang diperoleh merupakan ujaran kebencian yang dikoleksi dari media sosial Twitter dan diberi anotasi secara manual oleh 30 sukarelawan yang memiliki latar belakang beragam dari usia, agama dan ras. Distribusi jumlah kelas label pada dataset ditunjukkan pada Tabel 1

Tabel 1 Distribusi jumlah kelas label

Label Tweet	Jumlah Data
Non Hatespeech	453
Hatespeech	260
Total	713

Dataset dibiarkan tidak seimbang agar data bersifat realistis.

3.2 Data Praproses

Selanjutnya data yang sudah didapatkan, terlebih dahulu dilakukan pengolahan data awal atau data preprocessing. Tahap ini bertujuan untuk mendapatkan bentuk data yang diinginkan sebelum masuk ketahap implentasi. Terdapat beberapa proses pada tahap tersebut meliputi antara lain:

1. Praproses umum, mengubah semua huruf menjadi huruf kecil, menggantikan 2 atau lebih titik (.) dengan spasi, menghapus spasi dan tanda kutip (' dan ") di akhir *tweet*, menggantikan 2 spasi atau lebih dengan 1 spasi.

[3] https://github.com/ialfina/id-hatespeech-detection

ISSN No. 2502-8782

2. Special Twitter Feature Handles, mengubah link URL menjadi "URL", menghapus tanda *hashtag* seperti "#" lalu merubahnya menjadi kata menghapus *Retweet* RT, mengubah simbol *emoji* menjadi kata "EMO_POS" untuk *emoji* positif dan "EMO_NEG" untuk *emoji* negatif, mengahpus *spaces* berlebih dan mengubah *user mention* menjadi "USER_MENTION".
3. Penyaringan Kata, menghapus segala tanda baca ['"?!,.():;], menghapus penggunaan huruf ganda, menghapus tanda - dan ', validasi kata dengan pengecekan kata dan karakter secara alfabetis

3.3 Fitur

Pada penelitian ini digunakan dua fitur Word N-Gram yaitu unigram dan bigram. kedua fitur tersebut digunakan pada masing-masing percobaan.

3.4 Klasifikasi dan Evaluasi

Proses klasifikasi data ujaran kebencian berbahasa Indonesia menggunakan pendeketan machine learning. Algoritma yang digunakan adalah Linear Support Vector Machine. Untuk memudahkan percobaan, penulis menggunakan library scikit-learn [13] dengan bahasa pemrograman Python.
Evaluasi algoritma menggunakan metode 10-fold cross validation dan mencari nilai akurasi dari hasil evaluasi pada penelitian ini. dari hasil percobaan dengan menggunakan kedua fitur akan dibandingkan mana yang memiliki nilai terbaik.

4 Temuan dan Pembahasan

Percobaan yang dilakukan adalah melatih algoritma klasfikasi dengan masukan dataset yang sudah diproses. Lalu didapat keluaran nilai akurasi. Selanjutnya melakukan evaluasi dengan metode 10-fold cross validation. Percobaan dilakukan sebanyak dua kali dengan fitur unigram dan bigram lalu dilakukan perbandingan hasil antara kedua fitur tersebut. Hasil percobaan bisa dilihat pada Tabel 2

Tabel 2. Hasil evaluasi 10 fold cross validation

Fitur	Akurasi
Unigram	86.55 %
Bigram	76,67 %

Dari hasil tabel berikut, klasifikasi data ujaran kebencian dengan fitur unigram terlihat unggul dengan hasil 86.55 % sedangkan fitur bigram 76.67 %. Hal ini dikarenakan tiap baris data pada dataset cenderung berukuran satu kata sehingga fitur unigram unggul dalam klasifikasi dataset tersebut. Dengan begitu fitur Word Unigram dikombinasikan dengan algoritma Linear Support Vector Machine dinilai cukup baik dalam mengatasi masalah klasifikasi pada deteksi ujaran kebencian berbahasa Indonesia dengan hasil akurasi 86.55 %..
Hasil penelitian ini terhadap Klasifikasi menggunakan Linear Support Vector Machine pada tidak sama dengan penelitian sebelumnya [14]. Hal ini disebabkan karena fitur yang berbeda dan perhatian terhadap susunan kata, berbeda dengan penelitian ini fitur yang digunakan tidak memperhatikan susunan kata dan berdasarkan sebaran vector.

5 Simpulan dan Saran

Dari hasil studi ini, algoritma Linear Support Vector Machine dengan fitur Word Unigram mampu menghasilkan akurasi dengan metode evaluasi 10-fold cross validation sebesar 86.55 % dan memiliki performa cukup baik.

Untuk penelitian selanjutnya harus diperhatikan fitur yang digunakan karena sangat berpengaruh pada proses pelatihan algoritma. Fitur yang mungkin bisa digunakan seperti word2vec, Paragraph2vec, BOW dsb.

Algoritma yang digunakan bisa dikembangkan lebih lanjut kepada ensamble method atau kombinasi algoritma machine learning dengan deep learning. Karena dengan berbagai konfigurasi, kombinasi dan perbedaan metode yang digunakan bisa dilihat hasil terbaik dari berbagai metode tersebut

Kepustakaan

[1] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," *Proceeding LSM '12 Proc. Second Work. Lang. Soc. Media*, no. Lsm, pp. 19–26, 2012.

[2] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," *Proc. NAACL Student Res. Work.*, pp. 88–93, 2016.

[3] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "A Dataset and Preliminary Study," *Adv. Comput. Sci. Inf. Syst. (ICACSIS), 2017 Int. Conf. 2017*, pp. 1–5, 2017.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *Proc. 11th Int. AAAI Conf. Web Soc. Media*, no. Icwsm, pp. 512–515, 2017.

[5] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," *Proc. First Work. Abus. Lang. Online*, no. 7491, pp. 85–90, 2017.

[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," vol. 2017, no. April, pp. 1–3, 2017.

[7] G. Chowdhury, "Natural language processing," *Annu. Rev. Inf. Sci. Technol.*, vol. 37, pp. 51–83, 2003.

[8] N. Kumar, "A Review on Machine Learning Algorithms , Tasks and Applications," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 6, no. 10, 2017.

[9] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support vector machine: Teori dan Aplikasinya dalam Bioinformatika," *IlmuKomputer.Com.*, 2003.

[10] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 297, no. 20, pp. 273–297, 1995.

[11] D. Fradkin and I. Muchnik, "Support Vector Machines for Classification," *DIMACS Ser. Discret. Math. Theor. Comput. Sci.*, pp. 1–9, 2000.

[12] Y. Tang, "Deep Learning using Linear Support Vector Machines," *ICML 2013 Challenges Represent. Learn. Work.*, 2013.

[13] F. Pedregosa, R. Weiss, and M. Brucher, "Scikit-learn : Machine Learning in Python," *J. ofMachine Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language ∗," 2013.

ISSN No. 2502-8782