SENTIMENT ANALYSIS UNTUK DETEKSI UJARAN KEBENCIAN PADA DOMAIN POLITIK

Farrikh Alzami[1], Nuanza Purinsyira P[2], Ricardus Anggi P[3], Rama Aria Megantara[4], Dwi Puji Prabowo[5]

Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang

Gedung H Lantai 1, Kampus 1 Jl. Imam Bonjol No 207, Semarang

E-mail : alzami@dsn.dinus.ac.id[1]

Abstrak

Twitter merupakan salah satu media social yang banyak digunakan oleh masyarakat Indonesia dalam menyuarakan pendapatnya. Dikarenakan pengguna media social berkisar di segala umur, maka diperlukan sebuah metode untuk mengklasifikasikan bahwa suatu tulisan twitter yang ditulis termasuk dalam kategori kebencian atau tidak. Tujuan klasifikasi ini digunakan sebagai batu pijakan dalam filter ujaran kebencian sehingga para pengguna dapat menggunakan aplikasi media social dengan lebih nyaman. Penelitian ini, menggunakan Sastrawi untuk proses stemming tulisan twitter, kemudian melalui proses pembersihan karakter, unigram dan naïve bayes digunakan untuk tahap klasifikasi, menghasilkan performa dengan recall sebesar 84.8%, precision sebesar 85.4% dan akurasi sebesar 85%. Dengan performa yang cukup menggembirakan, dapat disimpulkan bahwa kombinasi sastrawi, pembersihan karakter, unigram dan naïve bayes dapat digunakan untuk mendeteksi ujaran kebencian pada domain politik.

***Kata Kunci*:** Sentiment Analysis, Ujaran Kebencian, Unigram, Naïve Bayes, Sastrawi

I. PENDAHULUAN (10pt huruf besar,rata kiri/bold)

Penggunaan internet terutama pada world wide web berkembang sangat cepat. Pengguna (user) yang sebelumnya hanya mempunyai posisi sebagai pembaca, sekarang bisa memberikan feedback atau komentar. Salah satu perkembangan world wide web yang paling pesat adalah Social Network (Jejaring Sosial). Jejaring sosial memperkuat kapasitas dan jangkauan komunikasi dan ekspresi pendapat, pada gilirannya, masalah yang sudah ada di masyarakat sekarang mencapai jangkauan yang lebih besar dan konsekuensinya diperkuat. Itu adalah kasus komentar ofensif.

Komentar yang menyinggung (offensive comment), dalam konteks kami, dapat digambarkan sebagai komunikasi yang bertujuan untuk membuat marah satu atau lebih individu. Ini adalah definisi luas yang dapat mencakup ucapan kebencian, kata-kata kotor, penindasan, dan pelecehan.

Banyaknya pengguna aktif di jejaring sosial berarti pada saat yang sama, komentar yang lebih ofensif dihasilkan dan semakin banyak orang yang terkena dampak pelanggaran ini. Baru-baru ini, tindakan hukum telah diambil terhadap perusahaan-perusahaan jaringan sosial seperti Facebook, Twitter, dan YouTube karena mereka memungkinkan pengguna untuk mempublikasikan teks yang dianggap sebagai kebencian. Menangani konten yang ofensif semakin mendapat perhatian. Volume publikasi dalam topik ini telah tumbuh secara signifikan dalam beberapa tahun terakhir(De Pelle & Moreira, 2017)(Pavlopoulos et al.,

2017)(Waseem & Hovy, 2016). Dalam beberapa kasus, para pengguna "jahat" ini, menggunakan platform jejaring sosial untuk melakukan kejahatan seperti peniruan identitas, pencemaran nama baik, polarisasi pendapat, atau penindasan dunia maya(Laorden et al., 2010).

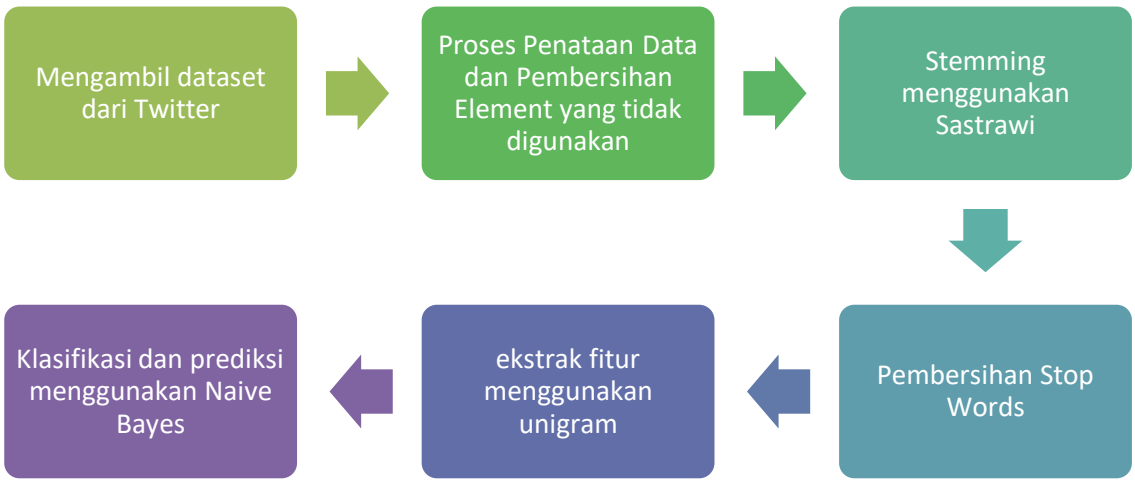
Analisis sentimen bertujuan mengidentifikasi sikap atau suasana hati orang melalui pemrosesan bahasa alami, analisis teks, dan linguistik komputasi. Dalam beberapa tahun terakhir, pembelajaran mesin telah menjadi alat yang sangat kuat untuk mengklasifikasikan sentimen. Dari literatur, analisis sentiment sudah dipakai untuk deteksi cyberbullying(Zhao et al., 2016), deteksi bahasa kasar (Park & Fung, 2017), ulasan film (Jefferson et al., 2017) dan identifikasi cyberhate (Burnap & Williams, 2015).

Menggunakan dataset twitter sebagai data latih dan uji, penelitian ini menghasilkan deteksi ujaran kebencian berdasarkan analisis sentiment analysis menggunakan sastrawi sebagai stemmer, unigram sebagai fitur ekstraksi dan naïve bayes sebagai model pembelajaran.

II. METODOLOGI PENELITIAN

1. Metodologi Penelitian

Alur penelitian merupakan proses dari pembuatan suatu system, dapat dilihat pada gambar berikut:



Gambar 1 Alur Penelitian

Tahapan pada Gambar 1 dapat dijelaskan pada sub-bab sebagai berikut:

a. **Mengambil dataset dari Twitter**

Dataset yang kami ambil, berdasarkan penelitian dari Alfina dkk (Alfina et al., 2017). Dataset tersebut diambil menggunakan Twitter API (Rahutomo et al., 2018), kemudian di label (class) secara manual berupa sentiment positive dan negative. Jumlah class positive dan negative masing-masing adalah 100.

b. **Proses Penataan Dataset dan Pembersihan Element yang tidak digunakan**

Setelah data diambil dari twitter, tahap berikutnya adalah ditata dataset. Pada mulanya dataset yang diambil mempunyai karakter tambahan seperti '$', '?', '!', '#', '\'. Tujuan peneliti membersihkan karakter supaya data dapat diolah menggunakan Stemmer

c. **Stemming Menggunakan Sastrawi**

Stemming merupakan proses pemetaan untuk mendapatkan kata dasar dari berbagai bentuk. Proses ini dilakukan dengan menghilangkan awalan, akhiran, sisipan, atau kombinasi awalan dan akhiran yang terdapat disetiap kata. Disini peneliti menggunakan Sastrawi sebagai tool stemming dikarenakan Sastrawi cocok digunakan untuk teks Bahasa Indonesia. Contoh stemming dapat dideskripsikan sebagai berikut:

Tabel 1. Contoh Stemming menggunakan Sastrawi

Teks Awal	Kamu Sebaiknya Mundur Saja Dari Jabatan
Teks setelah Stemming	Kamu Baik Mundur Saja Dari Jabat

d. **Pembersihan Stop Words**

Setelah dataset destemming, tahap berikutnya adalah pembersihan stop words menggunakan library dari Sastrawi. Contoh pembersihan dapat dideskripsikan sebagai berikut:

Tabel 2. Contoh Pembersihan Stop Words Menggunakan Sastrawi

Teks Stemming	Kamu Baik Mundur Saja Dari Jabat
Teks setelah Stemming	Kamu Baik Mundur Jabat

e. **Ekstrak Fitur Menggunakan Unigram**

Pada pemrosesan dokumen, skema representasi teks biasanya menggunakan vector space model (VSM) yang sering digunakan untuk pembobotan kata (word-weighting). Hasil yang diterima dari VSM merupakan dokumen yang relevan. VSM yang digunakan disini menggunakan kata kunci atau frase, yang secara umum dikenal sebagai unigram, bigram, trigram dan n-gram(Figueiredo et al., 2011; Lee et al., 2012; Xie et al., 2017). Untuk lebih mudahnya, N-gram merupakan urutan dari kata-kata N. misalkan terdapat sebuah kalimat: "makanan ini tidak terlalu enak", maka jika dibuat n-gram didapatkan sebagai berikut:

Kata dasar	Makanan ini tidak terlalu enak
Unigram	'makanan', 'ini', 'tidak', 'terlalu', 'enak'
Bigram	'makanan ini', 'ini tidak', 'tidak terlalu', 'terlalu enak'
trigram	'makanan ini tidak', 'ini tidak terlalu', 'tidak terlalu enak'

metode N-gram membuat keputusan dengan membandingkan nilai ini dengan ratio similaritas, yang didefinisikan sebagai ratio identic N-gram dibandingkan dengan jumlah total N-grams. Rasio similaritas dapat dihitung dengan cara (Gencosman et al., 2014):

$$ratio\ similaritas = \frac{\delta}{\min\ (\alpha, \beta)} \quad (1)$$

Dimana:

δ: jumlah n-gram yang identic

α: jumlah n-gram untuk $kata_A$

β: jumlah n-gram untuk $kata_B$

Disini $kata_A$ merupakan kata pertama dan $kata_B$ merupakan kata kedua yang digunakan sebagai pembanding karakter n-grams.

Saat ini, dikenal beberapa metode dalam pembuatan feature vector untuk tipe data dokumen, antara lain: Bag of Words (BoW) dan Term Frequency and Inverse Document Frequency (TF-IDF). BoW merupakan sebuah pendekatan algoritma yang menghitung seberapa banyak sebuah kata muncul (frekuensi) pada sebuah dokumen. Kelemahan BoW adalah urutan term dan kelangkaan (rareness) term tidak dipertimbangkan. Sedangkan pada TF-IDF, sebuah kata diberikan sebuah weight berupa TF dan IDF score, bukan berupa frekuensi seperti BoW.

Pada penelitian ini, peneliti menggunakan unigram dan TF-IDF sebagai fitur ekstraksi. Alasan penggunaan unigram adalah peneliti menggunakan pendekatan brute force untuk model pembelajaran dan ditemukan unigram lebih baik sebagai fitur ekstraksi.

f. **Klasifikasi dan Prediksi Menggunakan Naïve Bayes**

Setelah data diekstrak menggunakan unigram, tahapan berikutnya adalah menggunakan Naïve Bayes sebagai model pembelajaran. Peneliti memilih Naïve Bayes karena model pembelajaran tersebut bagus untuk kasus text mining, terutama pada fitur yang dihasilkan dari TF-IDF. Disini peneliti membagi dataset dengan ratio 70:30 untuk training dan testing. Kemudian untuk pencarian fitur ekstraksi, parameter naïve bayes dan cross validation, peneliti menggunakan metode brute -force. Dari hasil tersebut, dapat dirangkum sebagai berikut:

Tabel 3. Parameter yang dipakai dalam penelitian

Nama Parameter	**Nilai**
Ekstraksi Fitur	Unigram
TF-IDF	Hanya TF yang digunakan
Norm	L2
Smooth IDF	Digunakan
Sublinear TF	Tidak digunakan

III. HASIL DAN PEMBAHASAN

Setelah melakukan klasifikasi dan prediksi. Tahap berikutnya adalah menampilkan performa dari model pembelajaran yang peneliti buat. Rangkuman performa dapat ditampilkan sebagai berikut:

Tabel 4. Rangkuman Performa

Nama Performa	**Nilai**
Akurasi	85%
Precision	85.4%
Recall	84.8%

Dari tabel 4, dapat disimpulkan, bahwa model pembelajaran yang peneliti usulkan mendapatkan nilai yang cukup memuaskan. Namun dengan kisaran angka sebesar 85% untuk akurasi precision dan recall, masih diperlukan untuk peningkatan performa yang lebih lanjut

IV. KESIMPULAN

Deteksi Ujaran Kebencian menggunakan sentiment analysis berdasarkan Unigram dan Naïve Bayes berhasil dikembangkan dengan performa yang cukup memuaskan yaitu 85% untuk akurasi precision dan recall. Tahapan berikutnya adalah penggunaan ekstraksi fitur selain TF-IDF, seperti BOW, glove, dll.

V. UCAPAN TERIMA KASIH (Jika ada)

Penelitian ini didanai oleh LPPM Udinus pada Skema Penelitian IPTEKS.

VI. REFERENSI

[1] Alfina, I., Sigmawaty, D., Nurhidayati, F., & Hidayanto, A. N. (2017). Utilizing Hashtags for Sentiment Analysis of Tweets in The Political Domain. *ACM International Conference Proceeding Series*, *Part F1283*, 43–47.

[2] Burnap, P., & Williams, M. L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

[3] De Pelle, R. P., & Moreira, V. P. (2017, July 6). Offensive Comments in the Brazilian Web: a dataset and baseline results. *Anais Do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. https://doi.org/10.5753/brasnam.2017.3260

[4] Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira Jr., W. (2011). Word co-occurrence features for text classification. *Information Systems*, *36*(5), 843–858. https://doi.org/10.1016/j.is.2011.02.002

[5] Gencosman, B. C., Ozmutlu, H. C., & Ozmutlu, S. (2014). Character n-gram application for automatic new topic identification. *Information Processing & Management*, *50*(6), 821–856. https://doi.org/10.1016/j.ipm.2014.06.005

[6] Jefferson, C., Liu, H., & Cocea, M. (2017). Fuzzy approach for sentiment analysis. *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015577

[7] Laorden, C., Sanz, B., Alvarez, G., & Bringas, P. G. (2010). A threat model approach to threats and vulnerabilities in on-line social networks. In *Advances in Intelligent and Soft Computing* (Vol. 85, pp. 135–142). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16626-6_15

[8] Lee, L. H., Isa, D., Choo, W. O., & Chue, W. Y. (2012). High Relevance Keyword Extraction facility for Bayesian text classification on different domains of varying characteristic. *Expert Systems with Applications*, *39*(1), 1147–1155. https://doi.org/10.1016/j.eswa.2011.07.116

[9] Park, J. H., & Fung, P. (2017). *One-step and Two-step Classification for Abusive Language Detection on Twitter*. 41–45. https://doi.org/10.18653/v1/w17-3006

[10] Pavlopoulos, J., Malakasiotis, P., & Androutsopoulos, I. (2017). *Deep Learning for User Comment Moderation*. 25–35. https://doi.org/10.18653/v1/w17-3004

[11] Rahutomo, F., Saputra, P. Y., & Fidyawan, M. A. (2018). Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine. *Jurnal*

Informatika Polinema, *4*(2), 93–100.

[12] Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of the NAACL Student Research Workshop*, 88–93. https://doi.org/10.18653/v1/N16-2013

[13] Xie, F., Wu, X., & Zhu, X. (2017). Efficient sequential pattern mining with wildcards for keyphrase extraction. *Knowledge-Based Systems*, *115*, 27–39. https://doi.org/10.1016/j.knosys.2016.10.011

[14] Zhao, R., Zhou, A., & Mao, K. (2016). Automatic detection of cyberbullying on social networks based on bullying features. *Proceedings of the 17th International Conference on Distributed Computing and Networking - ICDCN '16*, 1–6. https://doi.org/10.1145/2833312.2849567