HATE SPEECH PREDICTION USING K-MEANS ALGORITHM

[1]Lim, Alexandre N Pratama, [2]Hironimus Leong
[1,2]Program Studi Teknik Informatika Fakultas Ilmu Komputer,
Universitas Katolik Soegijapranata
[2]marlon.leong@unika.ac.id

ABSTRACT

Hate speech in social media nowadays is a common thing to happen. Inspired by the issue, this research utilize data mining algorithm and methods to predict and classify it. By using dataset from twitter, this research will focus to define Hate Speech. Before beginning to use the algorithm, firstly the dataset needs to be cleaned, after that the data will be converted tu numeric values by using TF-IDF. With N-Gram, the final results will be more stable in terms of accuracy. After the preprocessing is done, then the K-Means Algorithm is used. The final results of the research is that by using Tri-Gram, accuracy is better than Bi-Gram and Uni-Gram with highest reach of 80% efficiency.

Keywords: K-Means, Clustering, TF-IDF, N-Gram, Hate Speech

PENDAHULUAN

Di kehidupan sehari-hari, kita tidak bisa lepas dari penggunaan internet. Banyak orang yang menggunakan internet memiliki akun sosial medianya sendiri-sendiri. Seiring berjalannya waktu, mereka suka membagikan momen-momen bahagianya dengan alasan hanya ingin berbagi atau ingin membuat arsip histori mereka. Hari demi hari internet menjadi suatu hal yang menjadi salah satu bagian penting dalam hidup. Hampir semua orang dapat mengoperasikan sosial media. Oleh karena itu juga, tidak jarang ada akun palsu. Sering kali akun palsu ini digunakan untuk melakukan perundungan ataupun menjatuhkan dengan kata-kata kasar dan tidak membangun. Hal ini disebut *Hate Speech* dan tidak jarang terjadi di akun palsu ke orang yang tidak mereka sukai.

Karena sudah sering terjadi, pihak sosial media memberikan solusi berupa fitur laporkan untuk melaporkan komentar atau postingan yang dirasa mengandung *Hate Speech*. Namun laporan ini masih ditinjau secara manual satu per satu. Akan sangat memakan waktu dan sulit untuk mengatasinya. Dengan teknologi saat ini, tidaklah mustahil untuk membuat sistem untuk membantu mengatasi masalah seperti ini.

Sudah banyak riset mengenai analisa sentimen untuk membedakan manakah yang positif dan negatif. Kebanyakan menggunakan algoritma Naive Bayes ataupun Support Vector Machine. Riset ini akan menggunakan algoritma K-Means untuk mengklasifikasikan *Hate Speech.*

LANDASAN TEORI

Riset yang dilakukan oleh Alexander Pak, dan Patrick Paroubek [1] membahas mengenai sentimen analisis dan opinion mining menggunakan algoritma Naive Bayes. Mereka sudah mencoba menggunakan SVM dan CRF sebelumnya, namun hasil akhir yang diberikan oleh Naive Bayes lebih baik dibanding kedua algoritma lainnya. Sebelum melakukan riset, dilakukan tes peforma menggunakan N-Gram. Dari hasil yang dicapai, dengan menggunakan Bi-Gram dapat menghasilkan hasil yang optimal dan akurasi yang tinggi dengan minim decision value. Kesimpulan dari penelitian ini adalah riset ini membuat otomatisasi dalam pengambilan data corpus yang dapat digunakan untuk data training untuk analisa sentimen. Penelitian ini menggunakan TreeTagger untuk melakukan proses POS-Tagging untuk membedakan sentimen positif, negatif, dan netral. Dari data ini, dapat digunakan untuk analisa sentimen menggunakan Naive Bayes.

Penelitian yang dilakukan oleh Try Iryanto Saputra, dan Rini Arianty [2] membahas mengenai implementasi algoritma K-Means untuk mengklasterkan komplain yang didapat dari pengguna Indosat. Data yang digunakan adalah data twitter yang diambil manual dari periode Desember 2018 hingga April 2019 dari konsumer @IndosatCare. Hasil dari pengambilan data manual ini adalah terdapat 300 data sentimen positif dan negatif. Data tersebut ada dalam bentuk file dengan format CSV. Lalu akan dimasukkan ke tahapan pra-proses, TF-IDF, dan K-Means dengan nilai k adalah 2 untuk memproses datanya. Kesimpulan dari penelitian ini ada dalam bentuk 3 wordcloud. Wordcloud pertama mayoritas membahas mengenai buruknya jaringan Indosat, wordcloud kedua membahas mengenai permintaan perbaikan jaringan, dan wordcloud terakhir membahas mengenai buruknya jaringan Indosat di kota Bogor.

METODE PENELITIAN

Penelitian ini menggunakan dataset yang didapat dari kaggle, dengan menggunakan python sebagai bahasa pemrogramannya. Tahap pra-proses terbagi menjadi beberapa bagian yaitu case folding untuk merubah data menjadi huruf kecil semua, menghilangkan stopwords, stemming, N-Gram, dan TF-IDF. Tahap pra-proses menggunakan bantuan dari sastrawi untuk bagian stemming. Untuk tahap TF-ID, riset ini menggunakan bobot TF-IDF untuk nantinya masuk ke algoritma K-Means. Setelah tahap pra-proses selesai, barulah masuk ke algoritma K-Means untuk mengklusterkan Hate Speech.

Pada algoritma K-Means ini, penelitian ini menggunakan nilai k adalah 2. Centroid pertama di set untuk mengindikasikan Hate Speech dan centroid kedua bukan Hate Speech.

HASIL DAN PEMBAHASAN

Pertama-tama data diolah terlebih dahulu pada tahap pra-proses. Contoh perbandingan data yang belum diolah dan telah diolah sebagai berikut.

Tabel 1. Pra-proses data

Sebelum	Sesudah
USER USER Kaum cebong kapir udah keliatan dongoknya dari awal tambah dongok lagi hahahah'	kaum cebong kafir sudah lihat dongok awal tambah dungu haha
menurutku pintu sorga ada yaitu pintu sorga yang asli dan pintu hatimu modusbanget	turut pintu sorga yaitu pintu sorga asli pintu hati modus banget

Setelah melalui tahap pra-proses, lanjut ke tahap N-Gram dan TF-IDF. Dalam tahap TF-IDF yang diambil untuk dimasukkan ke dalam algoritma adalah bobot dari TF-IDF per dokumen. Untuk data dummy yang dicoba, tahap N-Gram belum diterapkan. Untuk hasil dari bobot TF-IDF adalah sebagai berikut.

Tabel 2. Bobot TF-IDF

Document	Weight
1	3.15625288
2	2.50755042

Dari data diatas, dapat mulai masuk ke dalam algoritma. Dengan nilai centroid yang sudah ditentukan, algoritma ini menggunakan rumus Euclidean Distance untuk menghitung jarak tiap titik ke masing-masing centroid. Berikut adalah rumus Euclidean Distance.

$$d(p,q) = \sqrt{\sum_{i=1}^{N}(p_i - q_i)^2} \quad (1)$$

Dengan data dummy yang digunakan, hasil akhirnya terdapat 2 iterasi dan tabel iterasi terakhir nya adalah sebagai berikut.

Tabel 3. Tabel Iterasi Terakhir Data Dummy

No	C1	C2	Closest Cluster
1	0.444641718	0.753871598	C1
2	1.093344178	0.105169138	C2
3	0.930487368	0.268025948	C2
4	0.139838008	1.058675308	C1
5	0.414940378	0.783572938	C1
6	0.410927678	0.787585638	C1
7	1.614805648	0.416292333	C2
8	1.665747272	2.864260588	C1
9	0.255399488	0.943113828	C1
10	1.155416068	0.043097248	C2

Cara mengetahui akurasi dari data di atas adalah dengan menghitung berapa yang termasuk dalam kluster centroid 1 dan centroid 2. Centroid 1 adalah yang terindikasi Hate Speech, dan centroid 2 adalah yang bukan Hate Speech. Lalu setelah dapat datanya, baru dibandingkan dengan parameter aslinya. Dari 10 data dummy didapatkan akurasi 70%.

Setelah mendapatkan hasil dari data dummy, dilakukan penelitian ke data asli dengan bermacam-macam range datanya. Berikut adalah hasil dari penerapan ke 10 sampai 100 data.

Tabel 4. Tabel Akurasi

Total Data	Accuracy
10	70%
20	60%
30	66,7%
40	55%
50	54%
60	53,33%
70	58,57%
80	53,75%
90	54,44%
100	56%

Dari data diatas, dilakukan analisa lebih lanjut untuk meningkatkan akurasi yang ada. Dengan membandingkan tiap titik centroid satu dengan yang lainnya dapat mempengaruhi akurasi yang ada. Berikut hasil dari akurasi yang menggunakan titik centroid efektif.

Tabel 5. Tabel Akurasi Efektif

Total Data	Accuracy
10	70%
20	60%
30	66,7%
40	55%
50	54%
60	53,33%
70	58,57%
80	55%
90	55,55%
100	56%

Lalu untuk membuat akurasinya lebih baik lagi, diterapkan metode N-Gram. Hasil yang didapat dari N-Gram dapat meningkatkan akurasi. Yang digunakan dari metode N-Gram ini adalah Bi-Gram dan Tri-Gram. Hasil dari Tri-Gram terlihat lebih stabil dibandingkan dengan metode normal dan Bi-Gram. Akurasi tertinggi yang dicapai adalah 80% di 10 data.

Tabel 6. Tabel Akurasi N-Gram

Total Data	Bi-Gram	Tri-Gram
10	80%	80%
20	50%	50%
30	43,33%	56,66%
40	62,5%	62,5%
50	46%	62%
60	51,67%	58,33%
70	54,28%	58,57%
80	56,25%	60%
90	54,44%	60%
100	57%	60%

Tabel 7. Precision Recall Table

	Prediksi: Hate Speech	**Prediksi: Not Hate Speech**
Aktual: Hate Speech	TP=37	FN=21
Aktual: Not Hate Speech	FP=23	TN=19

KESIMPULAN

Dari tahap testing yang dilakukan, dapat disimpulkan bahwa jumlah data dapat mempengaruhi akurasi, begitu juga dengan titik centroid yang ditentukan. Dari rumusan masalah yang ada, didapatkannya kesimpulan sebagai berikut :

1. K-Means dapat membedakan Hate Speech dan Bukan Hate Speech dari dataset yang digunakan. Hal ini dapat terlihat dari perbandingan data yang diprediksi dan data aktual. Dapat juga terlihat dari tabel precision recallnya. Hasil yang valid adalah yang terbaca sebagai TP (True Positive) dan TN (True Negative).
2. K-Means dapat dengan efektif membedakan Hate Speech, tergantung dari banyaknya dataset dan titik centroid yang ada dengan akurasi tertinggi yang diraih yaitu 80%. Akurasi ini didapatkan dari centroid yang paling efektif.
3. Dengan menggunakan N-Gram, hasil akurasi yang ada terbukti lebih stabil terutama pada Tri-gram. Ini terjadi karena distribusi frekuensi yang lebih efisien dibandingkan tokenisasi biasa yang diterapkan dalam TF-IDF dan Tri-gram yang diterapkan dalam TF-IDF.

DAFTAR PUSTAKA

[1] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010, 1320–1326. https://doi.org/10.17148/ijarcce.2016.51274

[2] Saputra, T. I., & Arianty, R. (2019). Implementasi Algoritma K-Means Clustering Pada Analisis Sentimen Keluhan Pengguna Indosat. Jurnal Ilmiah Informatika Komputer, 24(3), 191–198. https://doi.org/10.35760/ik.2019.v24i3.2361

[3] Lutfi, A. A., Permanasari, A. E., & Fauziati, S. (2018). Corrigendum: Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine. Journal of Information Systems Engineering and Business Intelligence, 4(2), 169. https://doi.org/10.20473/jisebi.4.2.169

[4] Wangsanegara, N. K., & Subaeki, B. (2015). Implementasi Natural Language Processing Dalam Pengukuran Ketepatan Ejaan Yang Disempurnakan (Eyd) Pada Abstrak Skripsi Menggunakan Algoritma Fuzzy Logic. Jurnal Teknik Informatika, 8(2). https://doi.org/10.15408/jti.v8i2.3185

[5] Parveen, H., & Pandey, S. (2017). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, ICATccT 2016, 416–419. https://doi.org/10.1109/ICATCCT.2016.7912034

[6] Rezwanul, M., Ali, A., & Rahman, A. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. International Journal of Advanced Computer Science and Applications, 8(6), 19–25. https://doi.org/10.14569/ijacsa.2017.080603
[7] Riaz, S., Fatima, M., Kamran, M., & Nisar, M. W. (2019). Opinion mining on large scale data using sentiment analysis and k-means clustering. Cluster Computing, 22, 7149–7164. https://doi.org/10.1007/s10586-017-1077-z
[8] Windarto, A. P. (2017). Penerapan Datamining Pada Ekspor Buah-Buahan Menurut Negara Tujuan Menggunakan K-Means Clustering Method. Techno.Com, 16(4), 348–357. https://doi.org/10.33633/tc.v16i4.1447
[9] Alkhairi, P., & Windarto, A. P. (2019). Penerapan K-Means Cluster pada Daerah Potensi Pertanian Karet Produktif di Sumatera Utara. Seminar Nasional Teknologi Komputer & Sains, 762–767. http://seminar-id.com/prosiding/index.php/sainteks/article/download/228/223
[10] Dewi, S. M., Windarto, A. P., Damanik, I. S., & Satria, H. (2019). Analisa Metode K-Means pada Pengelompokan Kriminalitas Menurut Wilayah. Seminar Nasional Sains & Teknologi Informasi (SENSASI), 620–625. http://prosiding.seminar-id.com/index.php/sensasi/article/download/376/368