PENCEGAHAN *CYBERBULLYING* MENGGUNAKAN SISTEM DETEKSI UJARAN KEBENCIAN DALAM BAHASA INDONESIA

Wisnu Nugroho Aji[1], Yudi Tara[2] , Anggit Dwi Hartanto[3]

Program Studi Informatika Universitas AMIKOM
yudi.tara@students.amikom.ac.id, wisnu.1512@students.amikom.ac.id, anggit@amikom.ac.id

Abstrak

Semakin banyaknya pengguna media sosial, maka akan meningkatnya juga tindak kejahatan yang ada di dunia maya. Media sosial sudah mulai banyak digemari oleh semua kalangan umur dan dapat diakses dimana saja. Meningkatnya tindak kejahatan di media sosial berakibat meningkatnya korban tentunya. Seperti yang terjadi pada berita akhir November 2019 ini. Karena beberapa permasalahan tersebut, kami melakukan penelitian dengan objek komentar media sosial Instagram. Penelitian ini menggunakan metode pendekatan K-Nearest Neighbor(KNN) dibantu dengan algoritma pembobotan TF-IDF. Penggunaan metode ini dapat mengklasifikasi objek penelitian dengan jangka waktu yang singkat dengan jumlah dataset yang ditentukan. Dengan hasil penelitian ini diharapkan dapat meminimalisir tindakan kejahatan yang terjadi di dunia maya.

***Kata Kunci* :**

Cyberbullying, ujaran kebencian, KNN, Instagram

Abstract

The more social media users, the more serious crimes that exist in cyberspace. Social Media has started to be widely loved by all ages and can be accessed anywhere. The increasing crimes on social media resulted in increased casualties of course. As happened on the news end of November 2019. Because of some of these issues, we researched with Instagram's social media commentary object. The study uses the K-Nearest Neighbor (KNN) Approach method assisted with the TF-IDF-weighted algorithm. The use of this method can classify a research object with a short period of time with the specified number of datasets. The results of this research is expected to minimize the crimes that occur in cyberspace.

Keywords :

Cyberbullying, *Hatespeech*, KNN, Instagram

Pendahuluan

Bullying menerapkan sebuah tindakan untuk menyakiti secara fisik atau non-fisik dari individu untuk individu lainnya dengan alasan tertentu. Sedangkan kegiatan menyakiti individu atau kelompok lewat media sosial disebut dengan *cyberbullying.* Pada zaman yang sudah modern ini, media sosial memiliki banyak dampak. Mulai dari dampak positif maupun dampak negatif. Dalam waktu belakangan ini sudah ada korban jiwa yang meregang nyawa akibat *cyberbullying* walaupun sebagian besar korban dari kalangan artis atau orang terkenal. Hal itu tidak menutup kemungkinan akan terjadi pada orang terdekat atau ternyata diri kita sendiri [1].

Terdapat 41 persen hingga 50 persen remaja di negara ini pernah mengalami cyberbullying. Data tersebut didapatkan oleh UNICEF pada tahun 2016. Dari macam – macam korban memiliki jenis tindakan berbeda – beda. Contohnya seperti *cyber stalking* atau pengintipan pada dunia maya. Selain itu *revenge pom* yang diartikan sebagai tindakan penyebaran foto atau video dengan tujuan intimidasi daj pemerasan.

Semakin banyak lagi tindakan cyberbullying di Indonesia. Pemerintah mulai melirik dengan mengeluarkan UU ITE. Hal tersebut bertujuan untuk mencegah dan memberikan efek jera pada pelaku kejahatan media social. Selain pemerintah, kami juga akan melakukan upaya pencegahan untuk megurangi maraknya ujaran kebencian di dunia maya khususnya instagram. Dengan banyaknya orang yang peduli akan permasalahan ini, maka masalah mengenai *cyberbullying* akan berkurang. Dengan adanya penelitian ini diharapkan dapat membantu para pengguna agar lebih bijak dalam menulis komentar pada media sosial.

Tinjauan Pustaka

Machine Learning merupakan cabang dari kecerdasan buatan yang memunginkan sistem komputer mempelajari dari data, contoh atau hal yang sudah dialami. [2] ini memungkinkan mesin untuk mengerjakan tugas yang spesifik. Tujuan dasar dari *machine learning* salah satunya adalah memberikan pembelajaran pada computer dengan memanfaatkan data yang diperoleh untuk memecahkan masalah. Kemudian pada penelitian [3] meneliti tentang *Natural Language Processing* yang memiliki arti suatu penelitian dan aplikasi yang mempelajari computer agar dapat memahami teks dengan bahasa alami dan ucapan serta dapat memanipulasinya. Sebagai contoh dari penerapan *natural language processing* ini seperti pengenalan suara dan mesin penerjemah

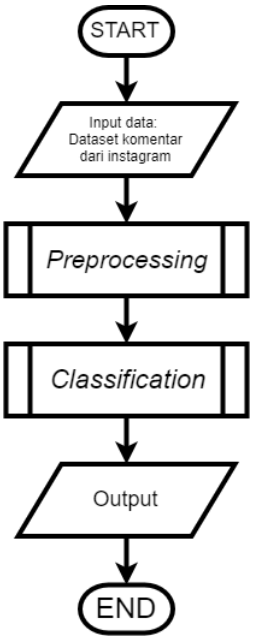
Sebelumnya sudah terdapat penelitian serupa mengenai ujaran kebencian di berbagai media sosial dengan berbagai metode yang berbeda, seperti [4]. Pada penelitian tersebut menggunakan tahapan *case folding*, *tokenizing*, *filtering*, dan *stemming*. Sedangkan penelitian yang kami buat menggunakan metode k-Nearest Neighbor(KNN) seperti yang dilakukan pada [5] namun dengan tahapan yang sedikit berbeda dengan penelitian [4]

Berdasarkan penelitian yang sudah dilakukan [6], untuk mengklasifikan teks berukuran pendek dengan metode k-NN, Naïve Bayes dan Algoritma SVM menjunjukkan hasil dengan akurasi yang terbaik diperoleh metode k-NN. Klasifikasi dengan teks singkat ini memakan waktu lebih sedikit dibandingkan penggunaan teks lengkap.

Metode Penelitian

K-Nearest Neighbors (KNN) Merupakan salah satu bentuk dari klasifikasi yang sudah banyak digunakan, karena kefektifannya dan mudah dimplementasikan serta non-parametrik [6]. K-Nearest Neighbors ini merupakan suatu medtode yang menggunakan algoritma supervised. Singkatnya cara kerja dari k-nearest neighbor adalah dengan membandingkan k objek dalam data training apakah mendekati atau mirip dengan data baru atau data testing [5].

Berikut adalah flowchart penelitian menggunakan metode KNN:

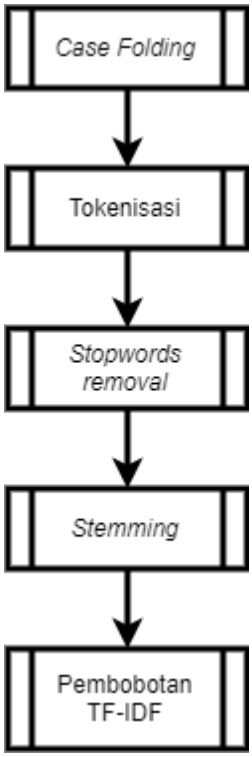


Gambar 1. Alur Penelitian

A. Pengumpulan Data
Pada pengumpulan data dalam penelitian ini adalah sebuah komentar dalam Bahasa Indonesia dalam media sosial *Instagram.* Pengambilan data komentar dari akun media sosial yang berpotensi mengalami ujaran kebencian. Akun media sosial yang berpotensi adalah milik *public figure* atau orang yang sudah dikenal masyarakat umum melalui internet. Data yang diambil tersebut dibagi menjadi dua variable, yakni variable positif dan variable negatif.

B. Preprocessing
Dalam metode ini terdapat langkah-langkah dalam *preprocessing*, yang dilakukan terlebih dahulu adalah *casefolding*, tokenisasi, *stop words removing* dan *stemming*. Proses – proses ini diperlukan untuk memilah data yang berguna dengan data yang tidak berguna untuk proses selanjutnya. Setelah itu yaitu menentukan bobot setiap kata pada data set menggunakan *TF-IDF method*.

Berikut adalah alur dari proses *preprocessing*:



Gambar 2. Preprocessing Data

1) Case folding

Menggubah setiap huruf pada data komentar menjadi huruf kecil dan menghilangkan karakter selain huruf "a" sampai huruf "z" karena dianggap delimiter atau melebihi batas. Contoh :

Tabel 1. Proses *Case Folding*

Komentar	Setelah *Case Folding*
Katanya Bangsawan tp kok bego	katanya bangsawan tapi kok bego
Bang boy makin ganteng aja sih	bang boy semakin ganteng saja sih

2) Tokenisasi

Proses pemotongan tiap kata penyusun pada kalimat pada data komentar. Prosesnya ditampilkan dengan contoh pada Tabel 2

Tabel 2. Proses *Tokenisasi*

Case Folding	Setelah Tokenisasi
katanya bangsawan tapi kok bego	katanya bangsawan tapi kok bego
bang boy semakin ganteng saja sih	bang boy semakin ganteng saja sih

3) Stopwords removal

Pembuangan kata atau term yang tidak relevan atau tidak memiliki arti, contoh dari stopword seperti : itu, ke, kepada, yang, dan seterusnya. Prosesnya ditampilkan dengan contoh pada Tabel 3

Tabel 3. Proses Stopwords Removal

Tokenisasi	Setelah Stopwords removal
katanya bangsawan kok tapi bego	katanya bangsawan bego
bang boy semakin ganteng saja sih	boy semakin ganteng

4) Stemming

Pencarian kata dasar dari tiap kata hasil *stopwords removal*. Mengembalikan bentukan kata ke dalam bentuk reperesntatif yang sama. Proses ini berguna untuk mereduksi bentuk term – term yang berbeda namun memiliki arti dasar yang sama menjadi satu bentuk.. Proses ditampilkan dengan contoh pada Tabel 4

Tabel 4. Proses *Stemming*

Stopword removal	Setelah Stemming
katanya bangsawan bego	kata bangsawan bego
boy semakin ganteng	boy makin ganteng

5) Pembobotan TF-IDF
TF atau *Term Frequency* yang berarti menghitung jumlah frekuensi atau seringnya frasa atau kata tersebut muncul. Pembobotan yang didapat dari frekuensi kemunculan suatu *term* dalam dokumen, semakin besar jumlah kemunculan maka semakin besar pula bobot dan nilai kesesuaiannya. Dalam pembobotan ini maka setiap dokumen dirubah menjadi per kata. Dengan tujuan untuk menghitung frekuensi setiap kata yang muncul dalam dokumen.

Setelah itu kami mencari kemiripan setiap *term* dengan menggunakan rumus *cosine similarity:*

$$\cos(\theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}}$$

Keterangan :

d_{ik} : bobot *term* ik
d_{ik} : bobot *term* jk
k: banyaknya *term* muncul

C. Klasifikasi
KNN merupakan metode yang mengklasifikasi dengan pendekatan dari kategori k tetangga yang terdekat. Tujuannya ialah mengklasifikasi obyek data berdasarkan *training sample* [7]. Jadi setelah menemukan tetangga terdekat maka akan ditemukan hasil dari klasifikasi.

Proses dari tahapan klasifikasi yaitu dengan mengolah data yang sudah diterima dari proses – proses sebelumya yang sudah melewati proses *preprocessing* dan pembebobotan dengan TF-IDF. Kemudian membuat dataset sesuai kelas yang sudah ditentukan yaitu negatif dan positif. Setelah itu menentukan data ujinya..Yang terakhir mencari kedekatan dari data uji dengan dataset yang sudah dikelaskan apakan nilai k mendekati kelas negatif atau kelas positif

Hasil dan Pembahasan

Dari pengujian data tersebut maka didapatkan hasil klasifikasi dari metode yang kami gunakan. Penguji membagi menjadi dua variabel, yakni variabel positif dan negatif. Hal itu memudahkan dan akan mempercepat proses klasifikasi. Maka komentar yang diinputkan dapat memiliki output atau hasil yang tergolong dalam positif atau negatif. Dengan begitu setiap individu dapat dengan mudah membedakan komentar yang termasuk dalam golongan positif atau malah yang negatif.

Kesimpulan

Proses klasifikasi atau pengelompokan menggunakan metode KNN guna mengidentifikasi ujaran kebencian yang terdapat dalam komentar *Instagram* efektif untuk dicoba. Karena waktu yang dibutuhkan dalam mengklasifikasi *data training* terbilang singkat. Kami melakukan pengujian menggunakan metode tersebut dengan bantuan perhitungan *cosine similarity*. Terdapat beberapa faktor yang dapat mempengaruhi berjalannya proses klasifikasi seperti jumlah frasa yang banyak.

Pada penelitian selanjutnya diharapkan dapat memperbaiki kekurangan tersebut. Maka dengan banyaknya data yang akan digunakan tidak akan mempengaruhi proses kalsifikasi. Selain itu juga dapat menyederhanakan pada proses pembobotan TF-IDF agar lebih efektif dibandingkan dengan penelitian sebelumnya.

Daftar Pustaka

[1] K. Oktaviani, "Apa itu "Cyber Bullying"?," Kompasiana, 25 September 2013. [Online]. Available: https://www.kompasiana.com/kiranaoktaviani/552ff83a6ea8344b778b45d4/apa-itu-cyber-bullying. [Accessed 15 November 2019].

[2] D. Sharma and N. Kumar, "A Review on Machine Learning Algorithms, Task and Aplications," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. VI, no. 10, pp. 1-2, 2017.

[3] G. G. Chowdhury, "Natural Language Processing," *Annual Review of Information Science and Technology*, p. 1, 2003.

[4] M. B. Ismiati, "DETEKSI KOMENTAR NEGATIF DI INSTAGRAM MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER," *Prosiding SNST Fakultas Teknik*, pp. 1-6, 2018.

[5] H. Leidiyana, "PENERAPAN ALGORITMA K-NEAREST NEIGHBOR UNTUK PENENTUAN RESIKO," *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*, 2013.

[6] K. Khamar, "Short Text Classification Using kNN Based on," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 4, 2013.

[7] O. S. S. S. E. Syahfitri Kartika Lidya, "SENTIMENT ANALYSIS PADA TEKS BAHASA INDONESIA," *SENTIKA 2015*, 2015.