https://jurnaleeccis.ub.ac.id/
p-ISSN : 1978-3345, e-ISSN(Online): 2460-8122

Metode Naive Bayes untuk Menganalisis Akurasi Sentimen Komentar di Youtube

Aditiya Rahman[1], Fadhil Rahmat[2], Muhammad Yoga Fariqi[3], dan Sumarni Adi[4]
[1,2,3,4]Program Studi Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
e-mail: [1]aditiya.rahman@students.amikom.ac.id, [2]fadhil.rahmat@students.amikom.ac.id, [4]sumarni.a@amikom.ac.id

Abstract **- The revolution on social media has attracted users to video sharing sites like YouTube. This site is the most popular social media site where people see, share and interact by commenting on videos. There are various types of videos shared by users such as songs, movie trailers, news, entertainment etc. Some time ago the most trending video was a video about World War III (WWIII / WW3). Analyzing comments from videos about WW3 gives viewers opinions about WW3. Study the sentiments expressed in this commentary whether WW3 gets positive or negative feedback. The machine learning algorithm, Naive Bayes, is used in comments to find out its sentiments. The test results of 1500 data produced 30.3% positive sentiment and 60.6% negative sentiment, with an accuracy of 78.17%.**

Index Terms **- YouTube, World War III, WW3, Naive Bayes**

Abstrak - Revolusi di media sosial telah menarik pengguna ke situs berbagi video seperti YouTube. Situs ini adalah situs media sosial paling populer tempat orang melihat, berbagi, dan berinteraksi dengan mengomentari video. Ada berbagai jenis video yang dibagikan oleh pengguna seperti lagu, trailer film, berita, hiburan dll. Beberapa waktu yang lalu video yang paling trending adalah video tentang World War III (WWIII/WW3) atau perang dunia ke 3. Menganalisis komentar dari video tentang WW3 memberikan pendapat pemirsa terhadap WW3. Mempelajari sentimen yang diungkapkan dalam komentar ini menunjukkan apakah WW3 mendapatkan umpan balik positif atau negatif. Algoritma machine learning, Naive Bayes, digunakan pada komentar untuk mengetahui sentimennya. Hasil pengujian sebanyak 1500 data menghasilkan sentimen positif sebanyak 30,3% dan sentimen negatif sebanyak 60,6%, dengan akurasi sebesar 78.17%.

Kata Kunci - YouTube, World War III, WW3, Naive Bayes

I. PENDAHULUAN

Era kini adalah era informatika. Segala sesuatu di dunia dapat terhubung dengan mudah memanfaatkan teknologi informasi yang berkembang pesat, sehingga kehidupan sehari-hari terasa lebih cepat, jarak terasa lebih pendek, segala hal terasa lebih mudah. *Gadget* dan *smartphone* adalah implementasi era informatika. Masyarakat dapat hidup tanpa uang, namun tidak tanpa *gadget* dan *smartphone*, kira-kira begitu frasa yang sering diungkapkan orang kini[1]. Terlebih lagi penggunaan *social* media yang semakin banyak dipergunakan oleh masyarakat. Maraknya penggunaan *social* media sendiri memiliki dampak positif dan dampak negatif. Dampak positif yang didapat dari penggunaan *social* media antara lain mempermudah komunikasi, mempermudah penyebaran informasi, mempermudah mencari ilmu pengetahuan dan lain sebagainya. Sebaliknya penggunaan *social* media memiliki dampak negatif berupa membuat seseorang kurang waspada terhadap lingkungan sekitarnya, rawannya terjadi kejahatan, namun yang paling sering dijumpai adalah banyaknya tersebar *hate speech* atau ujaran kebencian ke kelompok atau individu tertentu.

Hate Speech (Ucapan Penghinaan/atau kebencian) adalah tindakan komunikasi yang dilakukan oleh suatu individu atau kelompok dalam bentuk provokasi, hasutan, ataupun hinaan kepada individu atau kelompok yang lain dalam hal berbagai aspek seperti ras, warna kulit, etnis, gender, cacat, orientasi seksual, kewarganegaraan, agama, dan lain-lain[2]. Seiring dengan perkembangan zaman dan Teknologi Informasi, perkataan itu bisa diucapkan dengan banyak media. Salah satu media yang digunakan untuk melontarkan *hate speech* atau ujaran kebencian adalah situs penyedia layanan *streaming* video yaitu Youtube.

Youtube adalah sebuah situs web berbagi video yang populer dimana pengguna dapat mengunggah serta menonton berbagai klip video dengan gratis[3]. Berdasarkan CEO Youtube, Susan Wojcicki mengungkapkan bahwa ada 1,8 miliar pengguna Youtube terdaftar yang menyaksikan video di *platform* tersebut setiap bulannya di tahun 2019. Angka ini tidak termasuk penonton yang menyaksikan video di Youtube tanpa membuat akun[4].

Salah satu fitur yang diberikan oleh Youtube untuk penggunanya adalah fitur komentar, di mana pengguna bisa mengomentari sebuah klip video yang mereka buka dengan syarat pengguna harus *login* terlebih dahulu.

Pada kolom komentar di situs Youtube inilah sering terdapat komentar – komentar yang mengandung *hate speech* yang ditujukan kepada pembuat video, orang yang ada di dalam video, dan lain sebagainya.

Saat ini, masih minim penelitian terkait pendeteksian *hate speech* pada komentar pada situs penyedia layanan *streaming* video Youtube. Penelitian pertama kali menyatakan bahwa telah menekankan pada masalah – masalah berikut untuk menentukan polaritas komentar yang diberikan oleh pengguna Youtube. 1) Kamus sentimen saat ini memiliki keterbatasan, 2) Pengguna menggunakan bahasa yang informal, 3) Perkiraan

Manuscript submitted on January 2020, accepted and published on April 2020

sentimen yang dibuat oleh komunitas, 4) Kesulitan menetapkan label yang tepat untuk *events*, 5) Kesulitan mencapai kinerja klasifikasi yang memuaskan[5].

Pada penelitian selanjutnya menunjukkan bahwa penggunaan gabungan kamus *SentiWordNet* yang ada dan daftar yang diperluas untuk mengekspresikan sentimen dan pendapat pengguna memiliki kinerja lebih baik dalam mendeteksi sentimen pengguna dari komentar di Youtube[6].

Metode yang sering digunakan pada saat ini untuk menguji sentimen adalah metode C4.5, Support Vector Machine (SVM) dan Naive bayes. Berdasarkan penelitian yang sudah ada, diantara ketiga algoritma tersebut, algoritma yang paling cocok untuk klasifikasi sentimen adalah algoritma Naive Bayes. Dikarenakan algoritma ini mudah untuk dipahami, lebih cepat dalam hal perhitungan dan hanya memerlukan sedikit data training. Oleh karena itu peneliti memilih untuk menggunakan algoritma Naive Bayes dalam penelitian ini. Beberapa penelitian yang menggunakan metode Naive Bayes untuk mengklasifikasikan Youtube sebagai berikut. Penelitian berjudul 'Sentiment Analysis of Review Datasets using Naïve Bayes and K-NN Classifier' untuk mengklasifikasikan *review* pada *review* hotel dan film[7]. Penelitian tersebut menunjukkan akurasi Naive Bayes sebesar 80% dan lebih besar dari pendekatan k-NN. Penelitian lainnya berjudul 'Aspect-based sentiment analysis to review products using Naïve Bayes' untuk mengklasifikasikan *review* produk[8]. Penelitian tersebut menunjukkan metode Naive Bayes menghasilkan akurasi F1-Measure sebesar 75% dari *review* produk.

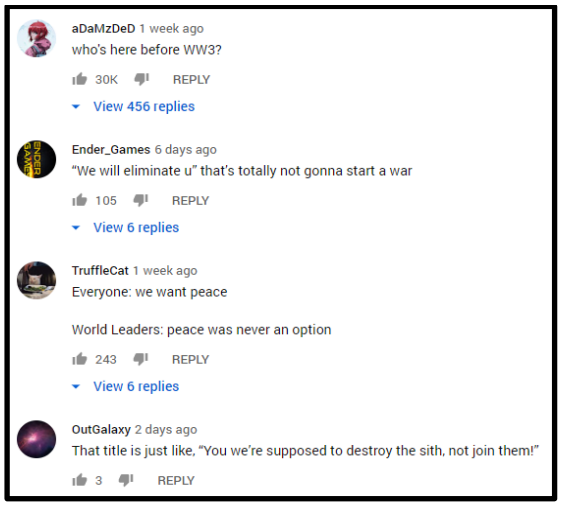
Pada penelitian ini, digunakan metode Naive Bayes untuk mengklasifikasikan apakah suatu komentar pada video di Youtube termasuk *sentimen positif atau negatif*. Hasil klasifikasi akan dievaluasi untuk mengetahui tingkat akurasinya. Kemudian akan diketahui berapa persentase dari sentimen positif dan negatif yang di dapat dari komentar di Youtube. Hal ini merupakan langkah awal untuk memahami sentimen dari komentar di Youtube.

II. METODE PENELITIAN

Pada tahap ini langkah – langkah penelitian yang dilakukan sesuai dengan alur penelitian adalah sebagai berikut :

1. Data komentar

Mengumpulkan data komentar dari Youtube. Komentar yang diambil adalah komentar yang menggunakan Bahasa Inggris. Jumlah komentar yang



Gambar 1. Dataset tes komentar youtube

diambil adalah sebanyak 1.500 komentar dari video channel **The Sun** dengan judul "*'We'll eliminate you': Donald Trump says 'we terminated Iran general to stop a war not start one*'" tentang World War 3. Data set ini di ambil dari live record comment youtube dengan menggunakan YouTube Data API v3.

2. Tahap preprocessing

Mencakup delapan tahapan, yaitu tahap pembersihan garis miring, memperbaiki kata singkatan, POS-*tagging,* tokenisasi, tupling, dan pemilihan kelas. Penghapusan garis miring dilakukan untuk mengatasi kelemahan dari POS *tagger.* Karakter garis miring dihapus dari komen, kemudian hasil dari POS *tagger* akan valid. Kemudian memperbaiki kata – kata singkatan yang sering digunakan seperti 'tdk' diubah menjadi 'tidak'.

3. Proses tokenisasi

Memecah komentar Youtube menjadi beberapa kata atau kumpulan kata yang berdiri sendiri. Penelitian ini menggunakan metode tokenisasi unigram dengan *negation-tag*.

4. Proses Bigrams Collocation

Meningkatkan ketepatan hipotesa kalimat. Hipotesisnya adalah komen orang yang mengatakan hal-hal seperti "Tidak hebat", yang merupakan ekspresi negatif kemudian Model kata menafsirkan sebagai ekspresi positif karena kata "Hebat" sebagai kata yang terpisah.

5. Proses Labelling Kalimat

Melakukan pelabelan kalimat atau emoticon positif, dan negatif sebagai dataset latih. Setelah mengetahui pola kata positif dan negatif, kemudian menerapkan dataset latih ke dataset uji dari live record di komentar youtube.

6. Youtube Data Api v3

Data Komentar youtube di peroleh dari Youtube API yang disediakan oleh Youtube yang terhubung dengan program sentimen analisis, kemudian data dari API tersebut di download, lalu di analisis menggunakan metode klasifikasi naive bayes. Pada saat pengumpulan data penelitian ini peneliti menentukan data komen youtube apa yang ingin dianalisis, hanya dengan memasukkan videoId, dan jumlah dataset yang ingin dianalisis, maka dari hasil pemrosesan itu menghasilkan hasil sentimen analisis.

7. Algoritma Naive Bayes

Naive Bayes classifier merupakan metode classifier yang berdasarkan probabilitas dan Teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas. [9]

A. Teori Bayesian
   a. X adalah data sampel dengan kelas (label) yang tidak diketahui.
   b. H merupakan hipotesa bahwa X adalah data dengan kelas (label) C. P(H) adalah peluang dari hipotesa H.
   c. P(X) adalah peluang data sampel yang diamati.
   d. P(X|H) adalah peluang data sampel X, bila diasumsikan bahwa hipotesa benar (valid). Techno.COM, Vol. 14, No. 4, November 2015: 299-314 302
   e. Untuk masalah klasifikasi, yang dihitung adalah P(H|X), yaitu peluang bahwa hipotesa benar (valid) untuk data sample X yang diamati:
      1) Naïve Bayesian Classifier mengasumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan keberadaan atribut (variabel) yang lain. Karena atribut tidak saling terkait maka :
      2) Bila P(X|Ci) dapat diketahui melalui perhitungan diatas maka label dari data sampel X adalah label yang memiliki P(X|Ci) * P(Ci) maksimum.

B. Kelebihan Naive Bayes classifier
   a. Mudah diimplementasi
   b. Memberikan hasil yang baik untuk banyak kasus

C. Kekurangan Naive Bayes classifier
   a. Harus mengasumsi bahwa antar fitur tidak terkait (independen) Dalam realita, keterkaitan itu ada
   b. Keterkaitan tersebut tidak dapat dimodelkan oleh Naïve Bayesian Classifier

III. HASIL DAN PEMBAHASAN

Bagian ini berisikan mengenai hasil-hasil yang didapatkan dengan menggunakan metode yang telah diterangkan dalam bagian sebelumnya.

Jumlah data yang digunakan adalah sebanyak 1500 data. Dimana data tersebut didapatkan secara dinamis dengan jumlah data yang diminta. untuk mendapatkan hasil sentimen analisis, ada beberapa proses yang dilakukan selama pengujian, diantaranya :

A. Classification

Klasifikasi Naive bayes menggunakan machine learning dapat dibagi dalam dua langkah :

a. Mempelajari model klasifikasi data training

Sebelum melakukan pengklasifikasian data tes, Klasifikasi Naive Bayes terlebih dahulu mempelajari model data training menjadi kelas positif dan negatif pada sebuah kalimat dan emoji.

TABLE I
DATA TRAINING KALIMAT

Positif	Negative
a feel-good picture in the best sense of the term.	on its own, it's not very interesting. as a remake, it's a pale imitation.
thoroughly enjoyable.	i didn't laugh. I didn't smile. I survived.
I admired this work a lot.	everything is off.
daring , mesmerizing and exceedingly hard to forget.	plodding , peevish and gimmicky.
…(5331 data)	…(5331 data).

TABLE II
DATA TRAINING EMOJI

Positif	Negative
(^.^)	:'(
(:	:\
\o/	:-X
;-)	@.@
…(200 data)	…(271 data)

(Data training : http://bit.ly/2u802Pe)

b. Menerapkan model klasifikasi data training ke data tes.

Setelah mengenali pola data training, klasifikasi naive bayes menerapkan data uji yang diambil melalui komentar youtube, dengan kata lain proses ini mentransformasikan pola data training ke data tes.

TABLE III
DATA UJI

Data Set	Hasil Sentiment
I don't know why people are making memes of world war 3 when it happens they will be scared not happy	-1
This guy is literally my times hitler. At the beginning of every century there's a war. Y'all say peace, but it'll get broken again and again	-1
He did the right thing! That guy was horrible he was responsible for killing so many.	-1
I feel like if the leaders of this country want war so bad why don't they fight and risk their life. Just saying □	-1
Does trump think we're in creative mode or something	1
… (1500 data)	

Keterangan :
Positif = 1
Negatif = -1
Netral = 0

Tabel 3 merupakan tabel yang telah dilakukan klasifikasi berdasarkan pola data training yang diterapkan pada data tes, beberapa data diatas merupakan hasil klasifikasi dari 1500 data yang tidak dicantumkan satu per satu.

c. Hasil Data Uji

Klasifikasi naive bayes telah berhasil menganalisis sentimen perihal WWIII dari 1500 komentar youtube. dengan beberapa fitur tambahan dalam mendukung keakuratan analisis.

```
Enter VideoId : O7Yj9pnOUP8
Enter no. of comments to extract : 1500
Comments downloading
[============================================================] 100.0%
Positive sentiment :  39.333333333333336
Negative sentiment :  60.666666666666664
```

Gambar 2. Hasil Analisis

Pada gambar diatas menunjukkan bahwa 30,3% dari 1500 komentar youtube merupakan sentimen positif dan 60.6% selebihnya merupakan sentimen negatif.

TABLE IV
HASIL

Positif	Negative	Akurasi (Naive Bayes)
39,3%	60,6%	78.17%

Dari hasil yang didapatkan menggunakan metode klasifikasi naive bayes, serta features extraction yang digunakan ini menghasilkan sentimen positif sebesar 39,3%, dan negatif sebesar 60,6%, dengan tingkat akurasi sebesar 78,17%.

IV. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, sentimen dari komentar di YouTube dapat diklasifikasikan menjadi dua yaitu positif dan negatif. Dari hasil pengujian 1500 data menunjukkan jumlah sentimen negatif lebih besar dari sentimen positif, yaitu sentimen negatif sebesar 60,6% dan sentimen positif sebesar 39,3%. Hasil akurasi dari pengujian tersebut dengan menggunakan algoritma Naive Bayes adalah 78.17%.

Namun dalam penelitian ini masih banyak kekurangan, diantaranya :

1. Dari total 43,464 komentar YouTube, hanya 1500 yang bisa dianalisa di karenakan cakupan program yang dibuat hanya dapat mendownload data sampai 1500 data saja.
2. Program Sentiment analysis terhadap komentar youtube ini masih berbasis CL.
3. Tingkat akurasi klasifikasi naive bayes 78,17% tergolong masih rendah.

REFERENSI

[1] Proposal-bisnis-bamboo-speaker.pdf https://karinov.co.id/contoh-proposal-bisnis-plan/#contoh-bisnis-plan

[2] Definisi Hate Speech https://www.bulelengkab.go.id/detail/artikel/hate-speech-definisi-hate-speech-66

[3] Apa itu Youtube? Pengertian, Karakteristik, dan Manfaat https://www.pahlevi.net/apa-itu-youtube/

[4] Jumlah Pengguna YouTube per Bulan Capai 1,8 Miliar https://kumparan.com/kumparantech/jumlah-pengguna-youtube-per-bulan-capai-1-8-miliar

[5] Muhammad Zubair Asghari, Shakeel Ahmad, Afsana Marwat, Fazal Masud Kundi, "Sentiment Analysis on YouTube : A Brief Survey," MAGNT Research Report Social and Information Networks, vol. 3, no. 1, pp. 1250-1257, 2015. [Online]. Available: Cornell University, https://arxiv.org/abs/1511.09142

[6] Choudhury, Smitashree and Breslin, John G. (2010). User sentiment detection: a YouTube use case. In: The 21st National Conference on Artificial Intelligence and Cognitive Science, 30 Aug - 1 Sep 2010, Galway, Ireland. [Online]. Available: http://oro.open.ac.uk/32459/

[7] Dey, Lopamudra et al. "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier." International Journal of Information Engineering and Electronic Business 8.4 (2016): 54–62. Crossref. Web.

[8] Mohamad Syahrul Mubarok, Adiwijaya, Muhammad Dwi Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes" AIP Conference Proceedings, vol. 6, pp. 180-185, August 2017.

[9] M. Trupthi, S. Pabboju and G. Narasimha, "Improved feature extraction and classification — Sentiment analysis," 2016 International Conference on Advances in Human Machine Interaction (HMI), Doddaballapur, 2016, pp. 1-6.

[10] S. M. Dr. Taufik Fuadi Abidin, "Naive Bayesian Classifier," FMIPA Universitas Syiah Kuala, Banda Aceh, 2013.

[11] PAMUNGKAS, Dyarsa Singgih; Setiyanto, Noor Ageng; Dolphina, Erlin. "Analisis Sentiment Pada Sosial Media Twitter Menggunakan Naive Bayes Classifier Terhadap Kata Kunci 'kurikulum 2013'". Techno.Com, [S.l.], v. 14, n. 4, p. 299-314, nov. 2015.