KLASIFIKASI KOMENTAR SPAM PADA SITUS YOUTUBE DENGAN METODE NAÏVE BAYES

Spam Comment Classification on YouTube using Naïve Bayes Method

Rya Meyvriska (G64164008), Ismail Adima(G64164053), Fafilia Masrofin (G64164056), Arie Cipta Ramadhan (G64164063)

Abstrak

YouTube merupakan salah satu penyedia video digital yang memiliki milyaran pengguna dan sempat menjadi nomor 1 sebagai penyedia yang paling tidak bisa ditinggalkan. Sebagai penyedia video digital ternama, YouTube masih memiliki masalah dengan spam. Terdapat komentar-komentar spam YouTube yang bahkan berpeluang mengandung URL malware yang berbahaya. Oleh karena itu, dilakukan penelitian pengklasifikasian komentar spam situs YouTube menggunakan methode Naïve Bayes. Penelitian menggunakan data Eminem yang diperoleh dari *repository* UCI yaitu YouTube Spam Collection. Metode menggunakan 2 cara yaitu tanpa *laplace* dan menggunakan *laplace*. Hasil akurasi pengklasifikasian Naïve Bayes tanpa laplace menunjukan akurasi 96%, *recall* 95%, dan presisi 96%. Sedangkan pengklasifikasian Naïve Bayes dengan laplace menunjukan akurasi 98%, *recall* 87%, dan presisi 93%.

Kata kunci:
Eminem, Klasifikasi, *laplace*, Naïve Bayes, Spam, UCI, YouTube

PENDAHULUAN

Perkembangan koneksi internet yang semakin cepat mendorong kemajuan perkembangan *video host* dan layanan berbagi video semakin tumbuh cepat dan popular di kalangan masyarakat. Hal inilah yang coba dimanfaatkan oleh banyak industri musik untuk mengenalkan karya mereka. Salah satu platform *video sharing* yang digunakan adalah YouTube. YouTube adalah platform publikasi konten video dengan fitur pencarian dan jejaring sosial sebagai dukungan untuk posting komentar dalam memberikan interaksi antara pemilik saluran dan pemirsa. Statistika penggunaan YouTube pada tahun 2015 menunjukkan bahwa platform video ini memiliki lebih dari 1 miliyar pengguna, 300 video diunggah setiap menitnya dan ditonton oleh milyaran pengguna (YouTube, 2017). YouTube juga menjadi konten pertama yang paling tidak bisa ditinggalkan anak remaja versi Defi Medya (Kompas.com, 2016).

Baru-baru ini YouTube mengeluarkan sebuah kebijakan monetisasi yang memberikan hadiah bagi creator dengan konten yang asli dan berkualitas sehingga meningkatkan jumlah penayangan di YouTube. Setelah sistem tersebut dikembangkan, mulai muncul konten-konten yang tidak diinginkan dan mengandung informasi berkualitas rendah atau yang lebih dikenal dengan istilah spam. Dari sekian banyak konten spam tersebut, menjadi masalah terbesar bagi YouTube untuk menangani komentar-komentar dalam jumlah yang sangat besar yang diposting oleh pengguna dengan tujuan untuk mempromosikan diri atau menyebarkan tautan yang berbahaya. Menurut Proofpoint, salah satu perusahaan keamanan komputer, 1 dari 200 media sosial ditemukan pesan spam dan 15% diantaranya

berisi tautan URL yang mengarah pada malware (Proofpoint, 2017).

Penelitian yang akan dilakukan ini membahas tentang proses untuk mengklasifikasi konten komentar yang ada pada YouTube. Ide dari penelitian ini adalah mencari kata kunci yang penting sebuah komentar yang ada dalam video YouTube dan memberikan label apakah komentar tersebut merupakan spam atau bukan spam. Penelitian ini menggunakan data set *YouTube Spam Collection* yang diperoleh dari *repository* UCI. Sebagai pendukung proses klasifikasi, pada tahap seleksi fitur digunakan teknik IDF dan pada tahap klasifikasi konten digunakan algoritma Naïve Bayes.

Rumusan Masalah

Berdasarkan latar belakang, beberapa masalah yang akan dibahas pada penelitian ini yaitu :

1. Bagaimana cara membedakan komentar yang termasuk kategori spam dan tidak dengan metode klasifikasi ?
2. Seberapa efektifkah metode klasifikasi dengan algoritma Naive Bayes untuk membedakan komentar spam dan bukan spam pada situs YouTube ?

Tujuan

Tujuan dari penelitian ini adalah untuk mengidentifikasi komentar yang termasuk spam dan bukan pada YouTube Spam Collection Dataset menggunakan metode klasifikasi Naïve Bayes.

Ruang Lingkup

Ruang lingkup dari penilitian ini fokus pada proses klasifikasi dengan metode Naïve Bayes dengan menggunakan data Eminem yang diperoleh dari *repository* UCI yaitu YouTube Spam Collection dan menghitung keakuratan yang dihasilkan.

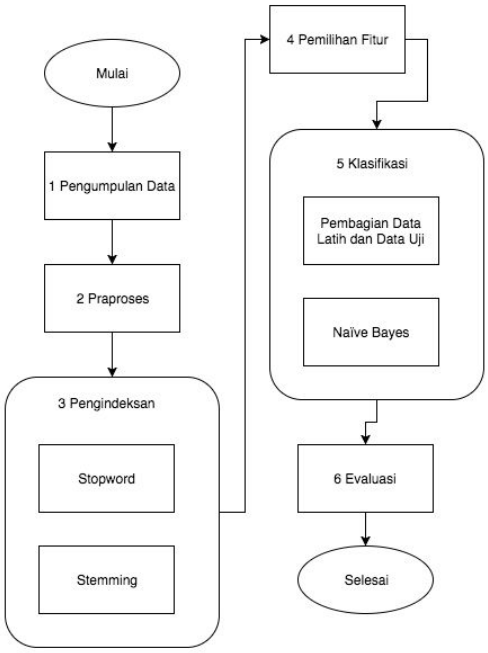
METODE

Tahapan-tahapan yang dilakukan pada penelitian ini diawali dengan pengumpulan data hingga evaluasi, dimana alurnya diperlihatkan pada Gambar 1. Penjelasan dari tahapan metode yang dilakukan adalah sebagai berikut:

Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data yang didapat dari UCI dengan set data berjudul YouTube Spam. Set data berisi 5 kumpulan komentar YouTube yang videonya masuk ke dalam 10 video paling banyak dilihat pada YouTube API V3. Lima kumpulan video pada YouTube Spam diperlihatkan pada Tabel 1.

Setiap kumpulan video berisi id komentar, nama pemilik komentar, tanggal, isi komentar,



Gambar 1 Tahapan Penelitian

karena memiliki jumlah komentar yang paling banyak, yaitu 448 komentar. Setiap data pada komentar di Eminem akan digunakan kecuali data id komentar, karena id komentar tidak memiliki hubungan apapun untuk klasifikasi.

Tabel 1 Kumpulan Video pada YouTube Spam

Judul	Jumlah Komentar
Sakira	370
Eminem	448

LMFAO	438
Katy Perry	350
Psy	350

Praproses

Hal yang dilakukan pada tahap praproses adalah toLowerCase, yaitu mengubah semua karakter huruf menjadi huruf kecil, dan menghilangkan *delimiters* seperti tanda titik (.), koma (,), tab maupun spasi berlebih (\t,\s,\n).

Pengindeksan

Tahap awal yang dilakukan pada pengindeksan adalah menghilangkan *stopwords,* dan *stemming*. *Stopword* adalah kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen (Dragut, 2009). Misalnya "*the*", "*an*", "*a*", dan sebagainya. *Stemming* merupakan suatu proses pemetaan dan penguraian berbagai bentuk (variants) dari suatu kata menjadi bentuk kata dasarnya (stem) (Tala, 2003). Tujuan dari proses *stemming* adalah menghilangkan imbuhan baik itu berupa prefiks, sufiks, maupun konfiks yang ada pada setiap kata.

Pemilihan Fitur

Pemilihan fitur merupakan proses pemilihan *term* yang mewakili informasi penting dari suatu dokumen atau teks yang akan digunakan sebagai fitur pada klasifikasi dokumen. Subset kata unik yang terpilih disebut dengan penciri. Seleksi fitur memiliki dua tujuan, yaitu mengurangi jumlah kata yang digunakan dan meningkatkan akurasi hasil klasifikasi (Manning et al. 2009).

Pada penelitian ini, pemilihan fitur yang akan digunakan menggunakan fungsi freqWord yang ada pada software R. Fungsi freqWord akan melakukan pembobotan dan menyeleksi fitur yang cocok digunakan untuk proses klasifikasi berikutnya.

Klasifikasi

Sebelum dilakukan klasifikasi, tahapan yang dilakukan adalah pembagian data sebagai data uji dan data latih. Pembagian data uji dan data latih sebesar 75% untuk data latih dan 25% untuk data uji.

Klasifikasi yang digunakan pada penelitian ini menggunakan metode berbasis peluang, yaitu *Multinomial Naïve Bayes*. Perhitungan peluang didapatkan dengan formula:

$$P(c|d) \propto P(c)\Pi_{\text{aksn}_d} P(t_k|c) \quad (2)$$

dengan parameter $P(c)$ merupakan peluang kelas c, $P(tk\ |c)$ adalah peluang token *tk* muncul pada dokumen c, dan *nd* adalah jumlah token unik pada dokumen (Manning et al. 2008). Kemudian nilai P(c) dan $P(tk\ |c)$ didapatkan dari formula:

$$P(c) = \frac{N_c}{N}\ ,\ P(t|c) = \frac{Tct}{\sum_{t' \in V} Tct'} \quad (3)$$

yaitu Nc adalah jumlah dokumen yang terdapat pada kelas c, N adalah total dokumen, *Tct* adalah banyaknya token t dalam dokumen latih dari kelas c dan $\sum$ Tct' t'$\in$V adalah jumlah seluruh token yang terdapat pada seluruh dokumen di kelas c. Nilai peluang P(t|c) dapat bernilai nol jika suatu kata dalam data uji tidak ada pada data latih. Untuk itu, digunakan *Laplace Smoothing*, dengan formula:

$$P(t|c) = \frac{Tct+1}{(\sum_{t' \in V} Tct') + B} \quad (4)$$

dengan B adalah jumlah *vocabulary* atau kata unik yang didapat dari data latih. Penambahan 1 dilakukan untuk mencegah terjadinya nilai peluang nol pada suatu kata. (Manning et al. 2008)

Evaluasi

Penelitian ini akan menggunakan *precision*, *recall* dan akurasi untuk menghitung kinerja dari hasil klasifikasi. Menurut Pantouw (2017) yang mengutip pada Yang et al. (2014) recall adalah

proporsi kalimat yang ditemukan kembali sebagai ringkasan, dan precision adalah proporsi jumlah kalimat yang ditemukan dan dianggap relevan. Tabel pendukung yaitu *Confusion Matrix* digunakan untuk memudahkan penghitungan evaluasi. *Confusion matrix* merupakan suatu tabel yang mengandung informasi tentang hasil aktual dan prediksi dari proses klasifikasi yang dilakukan oleh sistem (Han et al. 2012). Tabel 2 memperlihatkan bentuk *confusion matrix* dengan 2 kelas.

Tabel 2 *Confusion Matrix*

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Nilai *precision* dapat dihitung dengan formula:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

dengan TP (*True positive*) merupakan jumlah data hasil klasifikasi prediksi yang benar terhadap kelas aktual positif, FP (*False Positive*) merupakan jumlah data hasil prediksi positif terhadap kelas aktual negatif. Untuk nilai recall dapat diperoleh dengan formula:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

dengan FN (*False Negative*) merupakan jumlah data hasil prediksi negatif terhadap kelas aktual positif, selanjutnya akurasi dapat diperoleh dengan formula:

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \quad (6)$$

dengan TN (*True Negative)* merupakan jumlah data hasil klasifikasi prediksi yang negatif dan aktualnya juga negatif.

HASIL DAN PEMBAHASAN

Berdasarkan pengujian yang telah dilakukan, maka diperoleh hasil sebagai berikut:

1. Data

Data YouTube Spam pada video berjudul Eminem terdapat 3 atribut yang digunakan yaitu *comment*, *author*, dan *content.* Selain ketiga atribut tersebut, terdapat 1 atribut yaitu *class* sebagai target kelas spam atau bukan spam. Atribut *date* dihilangkan karena tidak ada hubungannya dengan penelitian yang dilakukan dan tidak digunakan untuk perhitungan klasifikasi.

2. Membuat TDM dan memilih fitur

Tahap ini menghasilkan *matrix term documents* yang terdiri dari 448 dokumen dan 2215 *terms*. Hasil ini di dapat setelah dilakukan pembuangan tanda baca, *stopwords*, dan spasi tambahan. Setelah itu, dilakukan pembobotan dengan library freqword dan dihasilkan term yang bagus menurut library tersebut untuk fitur klasifikasi.

3. Memisahkan data latih dan data uji

Pemisahan data latih dan data uji perlu dilakukan untuk melakukan klasifikasi. Pada tahap ini dilakukan pemisahan data dan dihasilkan data sebanyak 336 sebagai data latih dan 112 sebagai data uji yang ditentukan secara random.

4. Menghitung Peluang Prior

Nilai peluang prior dari setiap kelas dapat dilihat pada Tabel 3 sebagai berikut:

Tabel 3 Peluang prior

	Jumlah	P(c)
Spam	245	0.56
Tidak Spam	203	0.43

Setelah menghitung nilai peluang prior secara keseluruhan, kemudian menghitung nilai peluang *conditional* dari setiap *term* dengan menggunakan metode Naive Bayes (baik dengan nilai laplace = 0, dan laplace = 1). Contoh hasil perhitungan peluang *conditional* untuk setiap term dapat dilihat pada Tabel 4.

Tabel 4 Peluang conditional NB (laplace =0)

Term	P(t|c)	
	Tidak Spam	Spam
also (no)	1	0.97
also (yes)	0	0.03
amazing (no)	0.98	0.99
amazing (yes)	0.02	0.01
channel (no)	0.99	0.85
channel (yes)	0.01	0.15

4. Evaluasi

Tahap terakhir dari penelitian ini adalah melakukan evaluasi terhadap 112 data uji yang telah dipilih. Evauasi dari penelitian ini digambarkan dalam tabel confusion matrix dengan Naive Bayes menggunakan nilai laplace = 0 pada Tabel 5 dan nilai laplace =1 pada Tabel 6.

Tabel 5 *Confussion matrix* NB (laplace = 0)

Aktual	Prediksi	
	Spam	Tidak Spam
Spam	52	3
Tidak Spam	2	55

Berdasarkan Tabel 5, dari 112 data uji menunjukkan bahwa 52 data diprediksi spam hasilnya spam, sedangkan 55 sesuai dengan prediksi bahwa data tersebut tidak spam. Sebaliknya, terdapat 3 data yang diprediksi tidak spam padahal spam dan 2 data yang diprediksi spam padahal tidak spam. Evaluasi yang dilakukan adalah dengan menghitung nilai presisi, recall, dan akurasi dari *confussion matrix* pada Tabel 5 dan diperoleh hasil sebagai berikut:

$$Precision = \frac{52}{52+2} = 0.96$$

$$Recall = \frac{52}{52+3} = 0.95$$

$$Akurasi = \frac{52+55}{52+55+2+3} = 0.96$$

Tabel 6 *Confussion matrix* NB (laplace = 1)

Aktual	Prediksi	
	Spam	Tidak Spam
Spam	48	7
Tidak Spam	1	56

Berdasarkan Tabel 6, dari 112 data uji menunjukkan bahwa 48 data diprediksi spam hasilnya spam, sedangkan 56 sesuai dengan prediksi bahwa data tersebut tidak spam. Sebaliknya, terdapat 7 data yang diprediksi tidak spam padahal spam dan 1 data yang diprediksi spam padahal tidak spam. Evaluasi yang dilakukan adalah dengan menghitung nilai presisi, recall, dan akurasi dari *confussion matrix* pada Tabel 5 dan diperoleh hasil sebagai berikut:

$$Precision = \frac{48}{48+1} = 0.98$$

$$Recall = \frac{48}{48+7} = 0.87$$

$$Akurasi = \frac{48+56}{48+56+1+7} = 0.93$$

Dari hasil evaluasi dari kedua tabel, ternyata presisi dari Tabel 6 menunjukkan hasil yang lebih baik daripada presisi dari Tabel 5. Hasil ini menunjukkan bahwa tingkat ketepatan antara data yang ada dengan perhitungan yang dilakukan oleh sistem menggunakan NB laplace 1 lebih baik daripada NB laplace 0. Hal ini terjadi karena pada data set yang besar dan pemilihan data latih secara random menyebabkan kemungkinan adanya nilai 0 dalam model peluang. Oleh sebab itu, NB laplace 0 tidak dapat mengklasifikasi sebuah data inputan dengan benar sehingga nilai presisi yang dihasilkan lebih rendah dibanding dengan NB laplace 1. Meskipun memiliki nilai presisi yang lebih rendah NB laplace 0 ternyata memiliki nilai recall yang lebih besar dari NB laplace 1. Ini artinya tingkat keberhasilan sistem dalam mengelompokkan data lebih baik menggunakan NB laplace 0. Sedangkan untuk nilai akurasi, NB laplace 0 memiliki nilai akurasi yang lebih tinggi daripada NB laplace 1. Hal ini menunjukkan bahwa sistem telah mengklasifikasikan data dengan sangat baik

karena tingkat kedekatan antara nilai prediksi dan nilai aktual pada NB laplace 0 menunjukkan angka yang sangat tinggi. Dari hasil evaluasi menunjukkan nilai presisi, recall, dan akurasi yang sangat tinggi sehingga dapat dikatakan bahwa sistem telah berhasil mengklasifikasikan data dengan baik dan benar.

SIMPULAN

Klasifikasi dengan metode Naive Bayes termasuk dalam Klasifikasi Linear. Metode ini efektif untuk mengelompokkan komentar yang termasuk dalam spam dan bukan spam pada situs YouTube, dengan akurasi sebesar 96% apabila menggunakan nilai laplace 0 dan akurasi sebesar 98% jika menggunakan nilai laplace 1.

DAFTAR PUSTAKA

Alberto TC, Lochter JV, dan Almeida TA. 2015. YouTubeSpam: Comment Spam Filtering on YouTube. Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA'15), 1-6, Miami, FL, USA [diakses pada 29 Des 2017]. Tersedia pada http://bit.ly/2CflBQh.

Cran. Stemming Words [diakses pada 29 Des 2017]. Tersedia pada https://cran.r-project.org/web/packages/corpus/vignettes/stemmer.html.

Dragut Eduard *et all*. 2009. Stop Word and Related Problems in Web Interface Integration. Lyon(France) : VLDB Endowment. [diakses pada 29 Desember 2017]. Tersedia pada http://www.vldb.org/pvldb/2/vldb09-384.pdf.

Garonfolo HJ. 2015. Text Classification using a K Nearest Neighbour Model [diakses pada 29 Des 2017]. Tersedia pada http://garonfolo.dk/herbert/2015/05/r-text-classification-using-a-k-nearest-neighbour-model/.

Han et al. 2012. Data Mining: Concept and Techniques. Ed ke-3. Massachusetts (US): Morgan Kauffman. [diakses pada 29 Des 2017]. Tersedia pada http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

Kurniawan B, Effendi S, Sitompul OS. 2012. Klasifikas Konten Berita dengan Metode *Text Mining*. Jurnal Dunia Teknologi Informasi, Vol. 1, 14-19. [diakses pada 29 Des 2017]. Tersedia pada https://jurnal.usu.ac.id/index.php/duniait/article/viewFile/409/212.

Lang DT. 2004. Word Stemming in R. UC Davis, Department of Statistics. [diakses pada 29 Des 2017]. Tersedia pada http://www.omegahat.net/Rstem/stemming.pdf.

Manning CD, Raghavan P, Schütze H. 2008. An Introduction to Information Retrieval. Cambridge (UK): Cambridge University Press. [diakses pada 29 Des 2017]. Tersedia pada https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf.

Meira W dan Zaki MJ. 2014. Data mining and analysis: fundamental concepts and algorithms. Cambridge (UK): Cambridge University Press. [diakses pada 29 Des 2017]. Tersedia pada https://repo.palkeo.com/algo/information-retrieval/Data%20mining%20and%20analysis.pdf.

Proofpoint. Protect Against Global Spam. [diakses pada 29 Des 2017]. Tersedia pada https://www.proofpoint.com/us/solutions/social-media-malicious-content-remediation.

Pantouw JCW. 2017. Perbandingan Klasifikasi Rocchio dan Multinomial Naïve Bayes pada Analisis Sentimen Data Twitter Bahasa Indonesia. Bogor: Institut Pertanian Bogor, Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam. [diakses pada 29 Des 2017]. Tersedia pada http://repository.ipb.ac.id/bitstream/handle/

123456789/87230/G17app.pdf?sequence=1&isAllowed=y.

Putra Aswinda Prima. 2017. Analisis Sentimen Data Twitter menggunakan Naïve Bayes dengan Negation Handling pada Data Twitter Bahasa Indonesia. Bogor : Institut Pertanian Bogor, Departemen Ilmu Komputer, Fakultas Matematika dan Imu Pengetahuan Alam. [diakses pada 29 Des 2017]. Tersedia pada http://repository.ipb.ac.id/bitstream/handle/123456789/87311/G17jcw.pdf?sequence=1&isAllowed=y.

R Documentation. Wordstem [diakses pada 29 Des 2017]. Tersedia pada https://www.rdocumentation.org/packages/RTextTools/versions/1.4.2/topics/wordStem

R Documentation. Stopwords [diakses pada 29 Des 2017]. Tersedia pada https://www.rdocumentation.org/packages/qdap/versions/0.2.5/topics/stopwords

Sinaga ST dan Khodra ML. 2014. Restricted Content Classification Based On Videa Metadata And Comments (Case Study : YouTube.com). *Jurnal Ilmiah Kursor Menuju Solusi Teknologi Informasi*. vol 7 no 4. pp 165-172 [diakses pada 29 Des 2017]. Tersedia pada http://kursorjournal.org/index.php/kursor/article/view/31/24

Tala Fadillah Z. 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Universiteit van Amsterdam the Netherlands, Institute for Logic, Language and Computation. [diakses pada 29 Des 2017]. Tersedia pada http://www.illc.uva.nl/Research/Publications/Reports/MoL-2003-02.text.pdf.

UCI. YouTube Spam Collection Data Set. [diakses pada 29 Des 2017]. Tersedia pada https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection#

Vijayarani S, Ilamathi J, Nithya. 2015. Preprocessing Techniques for Text Mining - An Overview. International Journal of computer Science & communication Networks, Vol 5(1),7-16. [diakses pada 29 Des 2017]. Tersedia pada http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf

YouTube. Statistics. [diakses pada 29 Des 2017]. Terdapat pada https://www.youtube.com/yt/press/statistics.