kategori c

B' = Jumlah W dari kata unik atau nilai idf yang tidak di kali tf pada seluruh dokumen

Langkah-langkah dalam mengklasifikasi data dengan *Multinomial Naïve Bayes (MNB)* adalah:

1. Menghitung probabilitas prior pada setiap kelas dengan menggunakan rumus 2.3.
2. Menghitung probabilitas kata ke-n pada kelas c dengan menggunakan rumus 2.4.
3. Menghitung probabilitas suatu dokumen dengan menggunakan rumus 2.2.
4. Menentukan kelas dokumen dengan membandingkan nilai probabilitasnya antar kelas. Nilai probabilitas yang tertinggi akan dipilih dalam menentukan penentuan kelasnya.

2.6 *Multivariate Bernoulli*

Multivariate Bernoulli adalah salah satu model algoritma klasifikasi yang dikembangkan dari algoritma *Naïve Bayes Classifier* yang cocok dalam hal pengklasifikasian teks atau dokumen (McCallum & Nigam, 1998). Model dari *Multivariate Bernoulli* memperhitungkan jumlah data yang mengandung kata *term*, bukan frekuensi kemunculan kata. Pada model *Multivariate Bernoulli* menggunakan persamaan sebagai berikut (Karunia, Saptono, & Anggrainingsih, 2017):

$$P(c\,|d) = P\,(c)x\prod_{i=1}^{N} P(fk|c)\;x\;\prod_{i=1}^{M}(1 - P(fk|c)) \quad \textbf{(2.5)}$$

Keterangan:

P(c | d) = Probabilitas suatu dokumen dalam kelas c

P(c) = Probabilitas *prior* dari kelas c

P(fk|c) = Probabilitas setiap kata

Untuk menentukan probabilitas *prior* dari kelas c dihitung dengan menggunakan rumus:

$$P(c) = \frac{Nc}{N} \quad \textbf{(2.6)}$$

Keterangan;

Nc = Banyak kelas c pada semua dokumen

N = Banyak seluruh dokumen

Probabilitas kata ke-n pada kelas c yang digunakan dengan bobot kata TF-IDF dihitung dengan menggunakan rumus:

$$P(fk|c) = \frac{T_{ct}+1}{T_c+\sum c} \quad \textbf{(2.7)}$$

Keterangan:

T_{ct} = Banyak dokumen yang mengandung *term t* pada kelas c

T_c = Jumlah data latih pada setiap kelas c

$\sum c$ = Jumlah kelas atau banyak kategori

2.7 *Rocchio*

Rocchio merupakan salah satu dari algoritma klasifikasi dengan bentuk linear yang menerapkan prinsp dasar *contiguity hypothesis* yang berarti tidak akan terjadi *overlap* antara kelas yang sama dengan kelas yang berbeda. Nilai *centroid* dihitung dengan persamaan (Manning, Raghavan & Schutze, 2009):

$$\overrightarrow{u}(c) = \frac{1}{Dc}\sum d \in Dc\overrightarrow{v}(d) \quad \textbf{(2.8)}$$

Keterangan:

$\overrightarrow{u}(c)$ = Nilai *centroid* dari masing-masing kelas

Dc = Gugus dokumen

$\sum d \in Dc\overrightarrow{v}(d)$ = Jumlah vektor kata dalam kelas c

Untuk melakukan klasifikasi dalam algoritma *rocchio* dilakukan perhitungan *cosine similarity* antara titik d1 dan d2 dengan persamaan (Afriza & Adisantoso, 2017):

$$sim(d1, d2) = \frac{\overrightarrow{v}(d1).\overrightarrow{v}(d2)}{||\overrightarrow{v}(d1)||.||\overrightarrow{v}(d2)||} \quad \textbf{(2.9)}$$

Keterangan:

$sim(d1, d2)$ = Jarak kecocokan antara dokumen uji terhadap kelas

$\overrightarrow{v}(d1)$ = Nilai vektor centroid setiap kelas

$\overrightarrow{v}(d2)$ = Nilai vektor data uji

$||\overrightarrow{v}(d1)||$ = Nilai panjang vektor centroid

$||\overrightarrow{v}(d1)||$ = Nilai panjang vektor data uji

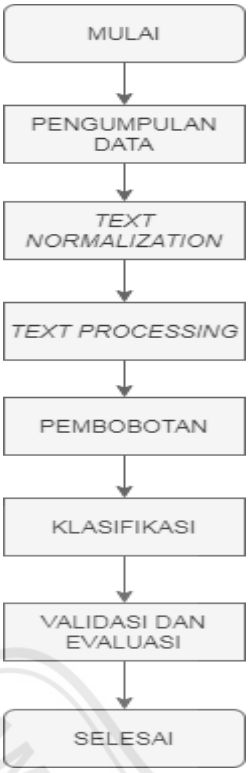
Langkah-langkah dalam mengklasifikasi data dengan *Rocchio algorithm* adalah:

1. Menghitung nilai *centroid* dengan menggunakan rumus 2.8.
2. Menghitung nilai *cosine similarity* dengan rumus 2.9.
3. Membandingkan nilai *cosine similarity* pada setiap kelas. Nilai *cosine similarity* yang tertinggi maka semakin mirip data uji terhadap kelas tersebut.

3. METODOLOGI

3.1 Rancangan Penelitian

Rancangan penelitian yang akan digunakan dalam penelitian ini antara lain mengumpulkan data, *text processing*, pembobotan kata, klasifikasi, validasi dan evaluasi.



Gambar 1. Alur Penelitian

3.2 Pengumpulan Data

Pada penelitian ini, data yang akan digunakan adalah data berita *hoax* dan benar. Data tersebut diambil melalui situs web *turnbackhoax.id*, yang menyediakan data berita *hoax* dan benar yang sudah diklasifikasikan berdasarkan hasil diskusi dan penelusuran fakta yang dilakukan oleh tim MAFINDO yang terdiri dari Muhammad Khairil Haesy M.Hum, Dedy Helsyanto S.IP, Bentang Febrylian S.IKOM dan Syarief Ramaputra S.IKOM. Berita yang diklasifikasi berasal dari berbagai media sosial atau dari situs web. Pada situs *turnbackhoax.id* terdapat beberapa kategori berita yaitu berita yang disebarkan dengan gambar, video, ataupun teks narasi. Data yang digunakan pada penelitian ini dipilih berdasarkan berita yang menggunakan teks narasi.

3.3 Text Normalization

Pada proses ini berfungsi untuk memperbaiki kata-kata yang ada pada teks narasi berita. Kata-kata seperti singkatan, bahasa gaul, atau *typo* dalam penulisannya agar kata tersebut sesuai dengan maksud dari penulisan narasi tersebut. Pada proses ini akan dibuat sebuah kamus dalam *file* teks yang mengandung kata singkatan, bahasa gaul, atau *typo* dan juga kata ganti yang akan digunakan pada perbaikan kata tersebut.

3.4 Text Processing

Text processing dilakukan untuk mempersiapkan kata pada data narasi sehingga bersih dari *noise* sebelum dilakukan pembobotan. Proses dari *text processing* antara lain *case folding, tokenizing, filtering,* dan *stemming.*

3.5 Pembobotan

Teknik pembobotan digunakan untuk menghitung nilai bobot suatu kata dalam dokumen dengan menggunakan algoritma TF-IDF. Langkah pertama dalam mencari nilai bobot suatu kata adalah dengan menghitung *term frequency* pada setiap data narasi. Langkah kedua adalah menghitung nilai *inverse document frequency* (IDF). Langkah terakhir adalah menghitung nilai bobot TF-IDF dengan mengkalikan nilai TF dengan IDF.

3.6 Klasifikasi

Proses klasifikasi yang digunakan pada penelitian ini menggunakan metode *Multinomial Naïve Bayes (MNB), Multivariate Bernoulli* dan *Rocchio Algorithm*. Pada tahap klasifikasi akan menentukkan kelas pada data uji.

3.7 Validasi dan Evaluasi

Proses validasi dilakukan pada kumpulan *dataset* berita hoax dan benar. Untuk mendapatkan hasil akurasi yang optimal maka akan dilakukan proses validasi dengan menggunakan *K-Fold Cross Validation*. Pada proses validasi tersebut akan membagi *dataset* menjadi beberapa bagian yang terdiri dari data latih dan data uji. Pada Penelitian ini ditentukan nilai *fold K* yang digunakan bernilai 2, 4, 5, 8 dan 10. Karena tidak ada aturan formal dalam pemilihan nilai pada *K-Fold Cross Validation* (Kuhn & Johnson, 2013), maka pemilihan nilai *fold K* tersebut diambil nilai yang habis dibagi atau tidak menyisahkan nilai, sehingga pada setiap partisi akan memiliki nilai yang seimbang.

Proses terakhir pada penelitian ini adalah melakukan evaluasi terhadap hasil klasifikasi yang telah dilakukan. Proses evaluasi ini dilakukan untuk mendapatkan nilai akurasi dari penggunaan beberapa algoritma yang telah digunakan sebelumnya. Perhitungan nilai akurasi pada evaluasi merupakan hasil dari perhitungan *K-Fold Cross Validation* yang memunculkan beberapa nilai akurasi dari beberapa banyak pengujian pada *Fold-K.* Dari hasil nilai yang didapatkan akan dipilih nilai akurasi yang paling optimum dari beberapa hasil pengujian.

4. HASIL DAN PEMBAHASAN

Data yang akan digunakan berjumlah 200 data berita terdiri dari 110 data berita *hoax* dan 90 data berita benar. Dari 200 data tersebut selanjutnya akan dilakukan *cleaning* agar bersih dari *noise* atau kata yang tidak diperlukan seperti kata penghubung, bahasa gaul, dll. Setelah data terkumpul selanjutnya dilakukan proses *text normalization* yang

berfungsi untuk mengubah kata singkatan atau bahasa gaul. Setelah dilakukan *text normalization* masuk ke dalam tahap *text processing.* Pada proses ini terdiri dari beberapa bagian yaitu *case folding, tokenizing, filtering* dan *stemming*. Selanjutnya dilakukan pembagian data atau partisi data dengan menggunakan *K Fold Cross Validation* untuk mendapatkan model terbaik yang kemudian akan diujikan ke dalam data uji baru. Penggunaan K Fold pada penelitian ini adalah 2, 4, 5, 8 dan 10. Setelah dilakukan pembagian data maka masuk kedalam tahap klasifikasi menggunakan algoritma *Multinomial Naïve Bayes (MNB), Multivariate Bernoulli* dan *Rocchio*. Berikut hasil terbaik yang didapatkan dari hasil klasifikasi algoritma *Multinomial Naïve Bayes (MNB)* ditampilkan Gambar 2.

K-FOLD	AKURASI	PRESISI	*RECALL*
2	78 %	76,47 %	89,65 %
4	80%	76,47 %	92,86 %
5	82,50 %	76 %	95 %
8	84 %	77,78 %	100 %
10	85 %	81,81 %	90 %

Gambar 2. Hasil Klasifikasi *MNB*

Berikut hasil terbaik yang didapatkan dari hasil klasifikasi algoritma *Multivariate Bernoulli* ditampilkan Gambar 3.

K-FOLD	AKURASI	PRESISI	*RECALL*
2	75 %	70,37 %	98,28 %
4	80%	74,36 %	100 %
5	85 %	81,25 %	100 %
8	84 %	78,94 %	100 %
10	85 %	81,25 %	100 %

Gambar 3. Hasil Klasifikasi *Bernoulli*

Berikut hasil terbaik yang didapatkan dari hasil klasifikasi algoritma *Rocchio* ditampilkan Gambar 4.

K-FOLD	AKURASI	PRESISI	*RECALL*
2	76 %	78,33 %	81,03 %
4	82%	85,19 %	82,14 %
5	85 %	81,81 %	90 %
8	88 %	86,67 %	92,86 %
10	95 %	90 %	100 %

Gambar 4. Hasil Klasifikasi *Rocchio*

Dari hasil klasifikasi yang didapatkan, akan diambil model terbaiknya. Untuk *MNB, Bernoulli* dan *Rocchio* akan diambil model pada *fold k=10* karena memiliki hasil yang terbaik dari semua *fold k* yang digunakan. Selanjutnya model akan digunakan sebagai data training untuk pengujian terhadap data uji baru. Data yang digunakan sebagai pengujian berjumlah 50 data yang terdiri dari 25 data berita *hoax* dan 25 data berita benar. Hasil dari pengujian data baru terhadap model terbaik ditampilkan pada Gambar 5.

ALGORITMA	AKURASI	PRESISI	*RECALL*
MNB	74 %	83,33 %	60 %
Bernoulli	70 %	62,50 %	100 %
Rocchio	76 %	88,24 %	60 %

Gambar 5. Hasil Prediksi Data Baru

Berdasarkan Hasil klasifikasi dengan menggunakan data baru didapatkan hasil akurasi, presisi dan *recall* pada algoritma *Multinomial Naïve Bayes (MNB)* sebesar 74% untuk akurasi, 83,33% untuk presisi dan 60% untuk *recall*. Sedangkan pada algoritma *Multivariate Bernoulli* mendapatkan hasil sebesar 70% untuk akurasi, 62,50% untuk presisi dan 100% untuk *recall*. Dan pada algoritma *Rocchio* mendapatkan hasil sebesar 76% untuk akurasi, 88,24% untuk presisi dan 60% untuk *recall*. Hasil yang didapatkan dari

validasi data mengalami penurunan, hal ini disebabkan karena adanya karakteristik baru yang ada pada data validasi yang tidak dimiliki pada data model yang menyebabkan penurunan hasil dari segi akurasi, presisi dan *recall*.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari hasil penelitian yang telah dilakukan adalah:

1. Didapatkan hasil presisi yang paling tinggi pada klasifikasi konten berita *hoax* berbahasa Indonesia diantara algoritma *Multinomial Naïve Bayes (MNB)*, *Multivariate Bernoulli* dan *Rocchio* adalah pada algotima Rocchio yaitu sebesar 88,24%.
2. Didapatkan hasil akurasi yang paling tinggi pada klasifikasi konten berita *hoax* berbahasa Indonesia diantara algoritma *Multinomial Naïve Bayes (MNB)*, *Multivariate Bernoulli* dan *Rocchio* adalah pada algotima Rocchio yaitu sebesar 76%.
3. Didapatkan hasil *recall* yang paling tinggi pada klasifikasi konten berita *hoax* berbahasa Indonesia diantara algoritma *Multinomial Naïve Bayes (MNB)*, *Multivariate Bernoulli* dan *Rocchio* adalah pada algotima Bernoulli yaitu sebesar 100%.
4. Dari ketiga algoritma yang digunakan pada data konten berita *hoax* berbahasa Indonesia, kinerja masing-masing algoritma memiliki kelebihan dan kekurangan. Dari kinerja akurasi *Rocchio* (76%) lebih baik dari algoritma *Multinomial Naïve Bayes (MNB)* (74%) dan *Multivariate Bernoulli* (70%), sedangkan *Multinomial Naïve Bayes (MNB)* (74%) lebih baik dari *Multivariate Bernoulli* (70%). Dari kinerja presisi algoritma *Rocchio* (88,24%) lebih baik dari algoritma *Multinomial Naïve Bayes (MNB)* (83,33%) dan *Multivariate Bernoulli* (62,50%), sedangkan *Multinomial Naïve Bayes (MNB)* (83,33%) lebih baik dari *Multivariate Bernoulli* (62,50%). Dari kinerja *recall* algoritma *Multivariate Bernoulli* (100%) lebih baik dari algoritma *Multinomial Naïve Bayes (MNB)* (60%) dan *Rocchio* (60%), sedangkan *Multinomial Naïve Bayes (MNB)* (60%) dan *Rocchio* (60%) memiliki kinerja yang sama dari segi *recall*.

5.2 Saran

Berdasarkan penelitian yang telah dijabarkan di atas, adapun beberapa saran yang diberikan untuk penelitian selanjutnya adalah sebagai berikut:

1. Diharapkan untuk peneliti selanjutnya agar menggunakan data berita dalam kurun waktu terbaru untuk dapat meningkatkan hasil yang didapatkan.
2. Diharapkan untuk peneliti selanjutnya bisa menerapkan atau membangun pembuatan sistem klasifikasi konten berita *hoax* dengan penelitian ini sebagai acuan.
3. Diharapkan untuk peneliti selanjutnya menggunakan data berita yang lebih banyak, karena semakin banyak data yang digunakan semakin banyak kata pula yang dapat diolah dan dapat

memungkinkan untuk mendapatkan hasil yang lebih optimal.

6. DAFTAR PUSTAKA

Feldman, R & Sanger, J., (2007). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructered Data*. New York: Cambridge University Press.

Han, J., Kamber, M, & Pei, J. 2011. *Data Mining: Concept and Techniques, Third Edition*. New York: Elsevier.

Harlian, M. 2006. *Machine Learning Text Categorization*. University of Texas, Austin.

Hermawati, F.A. 2013. *Data Mining*. Surabaya: Andi.

Havrlant, L., & Kreinovich, V. 2014. "A Simple Probabilistic of Term Frequency-Invers Document Frequency (TF-IDF)". *International Journal of General System*. University of Texas.

Karunia, S.A., Saptono, R., & Anggrainingsih, R. 2017. "Online News Classification Using Naïve Bayes Classifier with Mutual Information for Feature Selection". *ITSMART Vol. 6 No.1*. Universitas Sebelas Maret.

Kuhn, M., & Johnson, K. 2013. *Applied Predictive Modeling*. New York: Springer.

Larose, D.T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Willey & Sons, Inc.

Manning, C., Raghavan, P., & Schutze, H. 2015. *Introduction to Information Retrieval*. New York: Cambridge University Press.

McCallum, A., & Nigam, K. 1998. *A Comparison of Event Models for Naïve Bayes Text Classification*. Pittsburgh: Carnegie Mellon University.

Pantouw, J.C.W. 2017. Perbandingan Klasifikasi Rocchio dan Multinomial Naïve Bayes pada Analisis Sentimen Data Twitter Bahasa Indonesia. Institut Pertanian Bogor, Bogor.

Priansya, S. 2017. Social Media Text Normalization Using Word2vec, Levenshtein Distance, & Jaro-Winkler Distance. Institut Teknologi Sepuluh Nopember Surabaya, Final Project-KS 141501.

Rahman, A., Wiranto, & Doewes, A. 2017. "Online News Classification Using Multinomial Naive Bayes". *ITSMART Vol. 6 No.1*. Universitas Sebelas Maret.

Rasywir, E., & Purwarianti, A. 2015. "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin". *Jurnal Cybermatika Vol. 3 No.2*. Institut Teknologi Bandung.

Taprial, V., & Kanwar, P. 2012. *Understanding Social Media*. New York: Bookboon.

Triawati., & Chandra. 2009. Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia. Institut Teknologi Telkom, Bandung.

Turban, E., Aronson, J.E., & Liang, T.P. 2005. *Decision Support System and Intelligent System.* Yogyakarta: Andi Offset.

Wisaksono, A., & Mujiyatna, I.G. 2017. Klasifikasi Berita Berkategori Olahraga dengan Algoritma Multivariate Bernoulli Naïve Bayes dan Multinomial

Naïve Bayes. Gadjah Mada University Press, Yogyakarta.

Zhang, Y., Gong, L., & Wang, Y. 2005. "An improved TF-IDF approach for text classification". *Journal of Zhejiang University SCIENCE Vol. 6 No.1*. Tersedia di https://doi.org/10.1631/jzus.2005.A0049.

Manning, C., Raghavan, P., & Schutze, H. 2015. *Introduction to Information Retrieval*. New York: Cambridge University Press.