Klasifikasi *Tweet* Berdasarkan Keterkaitan *Tweet* Terhadap Topik Tertentu Pada *Twitter* Menggunakan *Naïve Bayes*

Muhamad Baydhowi [1,*] , Widya Apriliah [2], Ilham Kurniawan [3]

[1] Sistem Informasi; Universitas Bina Insani; Jl. Siliwangi No. 6 Rawa Panjang, Bekasi Barat, Telp 02188958130; e-mail: muhamadbaydhowi@gmail.com

[2,3] Sistem Informasi; Universitas Bina Sarana Informatika; Jl. Banten No. 1 Karangpawitan; Karawang; e-mail : widyaapriliah64@gmail.com, ilhamkurniawan2100@gmail.com

* Korespondensi: e-mail: muhamadbaydhowi@gmail.com

Diterima: 26 Oktober 2019; Review: 12 November 2019; Disetujui: 26 November 2019

Cara sitasi: Baydhowi M. 2019. Klasifikasi *Tweet* Berdasarkan Keterkaitan *Tweet* Terhadap Topik Tertentu Pada *Twitter* Menggunakan *Naïve Bayes*. Information System For Educators and Professionals. 4 (1): 95 – 103.

Abstrak: *Twitter* merupakan salah satu jejaring sosial atau *mikroblog* yang memungkinkan penggunanya untuk mengirim dan membaca pesan yang di sebut kicauan (*tweets*) yang berisi 140 karakter. Indonesia menduduki urutan ke lima jumlah pengguna (*user*) *Twitter* terbanyak di seluruh dunia dan Jakarta adalah kota yang paling aktif diseluruh dunia dalam membuat posting di *Twitter*. Berdasarkan informasi tersebut, kita dapat memanfaatkan data *tweet* untuk kepentingan tertentu seperti mengklasifikasikan *tweet* berdasarkan ketertarikan terhadap topik tertentu dengan kriteria yang telah ditentukan. Hasil penelitian ini diharapkan dapat digunakan untuk mendapatkan *user* potensial yang terkait dengan topik yang telah ditentukan sebagai sasaran marketing dari produk yang berkaitan dengan topik yang telah ditentukan sebelumnya. Namun penentuan klasifikasi *tweet* yang terkait dan tidak terkait dengan suatu topik manjadi kendala bagi para internet marketer. Tujuan penelitian ini adalah untuk menemukan metode klasifikasi *tweet* berdasarkan label terkait dan tidak terkait untuk kategori tertentu berdasarkan isi teks dari *tweet* tersebut. Data *tweet* diolah sehingga membentuk *Bag of Words* yang nantinya akan digunakan sebagai data training untuk melakukan klasifikasi dengan algoritma *Naïve Bayes* terhadap *tweet* yang diinput sebagai data *testing*.

Kata kunci: *Bag of words, Internet marketer,* Klasifikasi, *Naïve Bayes, Tweet*.

Abstract: *Twitter is one of the social network or microblog that allows users to send and read messages called tweets that contain 140 characters. Indonesia ranks the fifth largest number of Twitter users worldwide and Jakarta is the most active city in the world in posting on Twitter. Based on that information, we can utilize tweet data for specific purposes such as classifying tweets based on interest in a particular topic with predetermined criteria. The results of this study are expected to be used to obtain potential users associated with the topic that has been determined as a marketing target of the product relating to a predetermined topic. But determining the classification of related tweets and unrelated tweets to a topic becomes an obstacle for the internet marketers. The purpose of this research is to find the method of tweet classification based on related and unrelated labels for certain categories based on the text content of the tweet. Tweet data is processed and becomes Bag of Words which will be used as training data to classify tweet that inputted as testing data with Naïve bayes algorithm.*

Keywords*: Bag of words, Classification, Internet marketer, Naïve bayes, Tweet.*

E-ISSN: 2548-3587 ; 95 – 103

1. Pendahuluan

Perkembangan teknologi dan internet saat ini, diikuti juga dengan perkembangan media sosial yang pesat dan sangat beragam muncul, sehingga memudahan para pengguna teknologi dan internet untuk saling berkomunikasi. Salah satu media sosial yang saat ini banyak dan popular digunakan oleh para pengguna adalah *Twitter*. Kehadiran *Twitter* sebagai salah satu media komunikasi yang memungkinkan para penggunanya untuk mengirim dan membaca pesan yang di sebut kicauan (*tweet*s), dengan pesan yang dapat berisi 140 karakter yang digunakan sebagai media untuk menyebarkan informasi dan para penggunanya pun dapat saling berkomentar mengenai suatu topik yang dibahas.Pada pertengahan tahun 2010 *Twitter* mempunyai pengguna lebih dari 106 juta pengguna yang tersebar diseluruh dunia dan semakin meningkat setiap harinya hingga mencapai sebanyak 300.000 pengguna dan Setiap hari nya *Twitter* mendapatkan lebih dari 3 juta *request*. Berdasarkan angka tersebut menjadikan Indonesia sebagai negara yang menduduki peringkat 8 dalam mengakses situs *Twitter*. Adapun *Twitter* dapat menerima tweet sebanyak 55 juta pesan tweet setiap harinya yang berasal dari pesan tweet penggunanya [Rodiansyah dan Winarko, 2012]. Indonesia menduduki urutan ke lima jumlah pengguna (*user*) *Twitter* terbanyak di seluruh dunia dan lebih lanjut Jakarta menempati posisi sebagai kota yang paling aktif diseluruh dunia dalam membuat posting di *Twitter* [Eddyono, 2013]. Berdasarkan informasi tersebut, kita dapat memanfaatkan data *tweet* tersebut untuk kepentingan tertentu misalkan untuk mengklasifikasikan *tweet* berdasarkan ketertarikan terhadap topik tertentu dengan kriteria yang telah ditentukan.

Data mining merupakan sebuah proses dari *knowledge discovery* (penemuan pengetahuan) dari data yang sangat besar [Han dan Kamber, 2006]. Sementara itu, *text mining* merupakan bidang data *mining* yang bertujuan untuk mengumpulkan informasi yang berguna dari data teks dalam bahasa alami atau proses analisis data teks kemudian mengekstrak informasi yang berguna untuk tujuan tertentu [Witten dan Frank, 2005]. Salah satu teknik dalam data *mining* yaitu klasifikasi, dimana dalam data mining teknik klasifikasi digunakan untuk melakukan penggolongan suatu data [Masripah, 2016] .

Pada penelitian ini menggunakan algoritma *Naïve Bayes classification* untuk mengklasifikasikan *tweet* berdasarkan keterkaitan *tweet* terhadap topik tertentu dengan memanfaatkan *account* yang dinilai punya keterkaitan terhadap topik yang diteliti dengan harapan hasil penelitian ini bisa digunakan untuk mendapatkan *user* yang terkait dengan topik yang telah ditentukan sebagai sasaran marketing dari produk yang berkaitan dengan topik. Pendekatan ini merupakan pendekatan yang mengacu pada teorema *Bayes* yang merupakan prinsip peluang statistika untuk mengkombinasikan pengetahuan sebelumnya dengan pengetahuan baru. Prinsip ini kemudian digunakan untuk memecahkan masalah klasifikasi. Penggunaan algoritma ini dinilai sesuai karena *Naïve Bayes*ian *classification* merupakan salah satu algoritma klasifikasi yang sederhana namun memiliki kemampuan dan akurasi tinggi [Rish, 2001]. Implementasi algoritma *Naïve Bayes* terbukti efektif dalam banyak aplikasi praktis, ternasuk dalam proses klasifikasi teks [Nurelasari, 2018].

2. Metode Penelitian

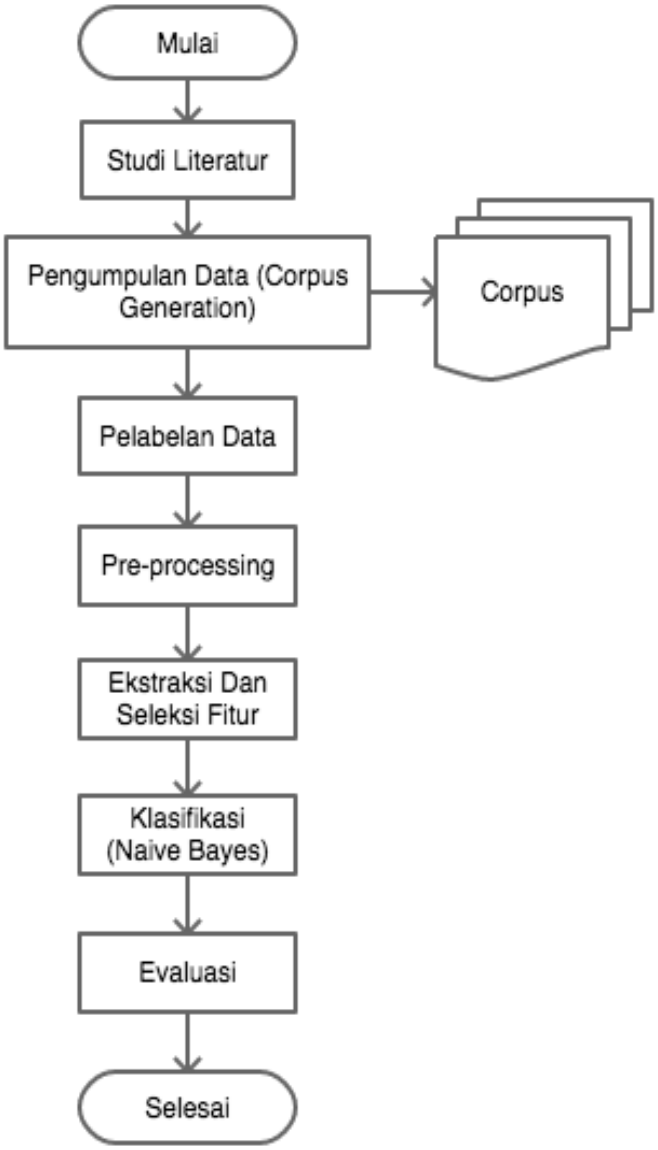
Penelitian ini merupakan *design oriented research* untuk membuat metode baru pengklasifikasian data *Twitter* berdasarkan keterkaitan *tweet* terhadap topik tertentu. Kegiatan penelitian yang dilakukan yaitu 1) **Studi literatur**, bertujuan untuk mencari dan menelusuri informasi-informasi terkait dari topik penelitian. 2) **Pengumpulan Data** *(Corpus Generation),* bertujuan untuk menghasilkan corpus dari pesan-pesan *tweet* yang sesuai untuk diproses. 3) **Pelabelan**, bertujuan untuk memberikan informasi bahwa data *tweet* yang ada terkait atau tidak terkait dengan topik yang ada. 4) ***Pre-processing*** [Allah, 2008]*, Pre-processing* terbagi menjadi beberapa tahap, diantaranya (a) *Tokenization*, mengurai teks menjadi satuan kata penyusun berdasarkan pemisah kata. (b) *Stop word removal*, kata-kata yang sangat umum yang tidak berguna untuk deindex. (c) *Normalization*, melakukan standarisasi penulisan terhadap huruf-huruf yang bisa ditulis menggunakan lebih dari satu cara. (d) *Stemming*, Ekstraksi Dan Seleksi Fitur [Manning et al., 2008], tujuan dari pemilihan fitur adalah mengurangi fitur agar klasifikasi teks lebih efisien dan dapat meningkatkan hasil dari klasifikasi dengan menghilangkan fitur noise. 5) **Klasifikasi**, proses pengelompokkan data *tweet* setelah dimasukkan ke dalam algoritma. Dalam hal ini algoritma *Naïve Bayes*. 6) **Evaluasi**, hasil klasifikasi selanjutnya divalidasi kehandalannya [Allah, 2008] , untuk melaksanakan validasi ini digunakan 10-fold validation. Selain pengukuran validasi dari algoritma yang dipakai dalam penelitian juga untuk

mengukur efektifitas dari *classifier* berdasarkan kategori yang dijadikan label, ada 3 jenis pengukuran (*metrics*) umum yang digunakan yaitu *precision*, *recall*, dan *F1 measure*. Untuk *precision* dan *recall* telah dipaparkan didalam *confusion matrix* pada gambar 4. Untuk *F1 Measure* dapat dihitung berdasarkan keduanya [Chantar, 2013]. Perhitungan *F1 Measure* menggunakan rumus perhitungan sebagaimana pada gambar 1.

$$F\text{-}measure = 2 \cdot \frac{precision.recall}{precision + recall} \quad \text{.....(1)}$$

Gambar 1. Rumus Perhitungan *F1 Measure*

Penelitian ini dibagi menjadi beberapa langkah sebagaimana pada gambar 2.



Sumber: Hasil Penelitian (2019)

Gambar 2: Kerangka Penelitian

E-ISSN: 2548-3587 ; 95 – 103

3. Hasil dan Pembahasan

Pengklasifikasian data *Twitter* pada penelitian ini dilakukan secara bertahap, sebagaimana pada pembahasan ini.

Pengumpulan data (Corpus Generation)

Proses pengambilan *tweet* dilakukan dengan menarik data yang berasal dari beberapa *account* yang memiliki keterkaitan dengan beberapa topik yang telah ditentukan dengan harapan data tersebut bisa digunakan untuk mengelompokkan *tweet* lainnya dengan topik terkait.

Proses pembentukan *corpus* terbagi menjadi beberapa langkah, yaitu: 1) Koneksi ke API *Twitter* dengan menggunakan *Twitter*OAuth dari Abraham Williams dengan mempergunakan Bahasa pemrograman PHP. 2) Tarik data tweet berdasarkan topik tertentu dari account tertentu dengan menambahkan filter berdasarkan *account* atau *screen*_name *Twitter* pada library *Twitter* OAuth, *account* terbagi menjadi 3 kategori berdasarkan topik terkait yaitu Retail, Proyek dan Pendidikan. Hasil akhir dari tahapan ini adalah *tweet corpus* yang selanjutkan akan diberi label.

Pelabelan data

Setelah tahapan dalam pengumpulan data dilakukan, tahap selanjutnya adalah dengan melakukan proses pelabelan terhadap tweet *corpus*. Adapun dalam tahap ini dilakukan dengan cara membaca setiap *tweet* yang ada berdasarkan pemahaman penulis dan kemudian diberikan informasi klasifikasi yang sesuai dan dalam hal ini nilai nol untuk tidak terkait dan nilai satu untuk klasifikasi terkait. Hasil akhir dari tahap ini adalah daftar tweet yang sudah berisi klasifikasi berdasarkan label yang diberikan secara manual, untuk selanjutnya dilakukan proses *pre-processing.*

Pre-processing

Pada proses ini dilakukan beberapa tahap, diantaranya: 1) *Tokenization*, tahapan ini dibagi menjadi empat proses yaitu (a) Dibersihkan dari objek yang bersifat url, contohnya:Yuk borong semua rasa Muffin #BreadLife dimakan sambil minum teh. Pasti asyik! ;) https://t.co/25Z9YjlCSB. Menjadi: Yuk borong semua rasa Muffin #BreadLife dimakan sambil minum teh. Pasti asyik! ;) (b) Dibersihkan dari karakter @ yang berarti memention account tertentu. (c) Dibersihkan dari karakter hashtag (#). (d) Untuk *tweet* yang bersifat re*tweet* dikeluarkan dari data yang digunakan. 2) Stop word removal, berdasarkan tokenization, token-token tersebut difilter kembali untuk dihilangkan setiap kata yang termasuk stop word berdasarkan stop word list dari https://github.com/masdevid/ID-Stopwords/blob/master/id.stopwords.02.01.2016.txt. 3) Normalization, seperti Yummy... manjadi yummy, *Twitter*, menjadi *Twitter*, 2017! menjadi 2017, Kalender... menjadi Kalender.

Bag of Words (BoW)

Berdasarkan hasil dari tahap *pre-processing* yang telah dilakukan sebelumnya, maka dapat didapatkan data yang bisa dipergunakan untuk membuat *Bag of Words*. Sebagaimana data dapat dilihat pada tabel 1 dibawah ini.

Tabel 1. Perbandingan Hasil Pre-processing (dalam persentase)

No.	Topik	*Tweet*	Token	Term Unik
1	Retail	44,67	40,72	40,73
2	Projek	11,11	14,42	14,63
3	Pendidikan	44,42	44,86	44,64

Sumber: Hasil Penelitian (2019)

Ekstraksi dan seleksi fitur

Berdasarkan *Bag of Words* dapat ditentukan bahwa data yang akan dijadikan sebagai data *training* adalah *term* yang memiliki frekuensi kemunculan lebih atau sama dengan lima. Adapun data dapat dilihat pada tabel 2.

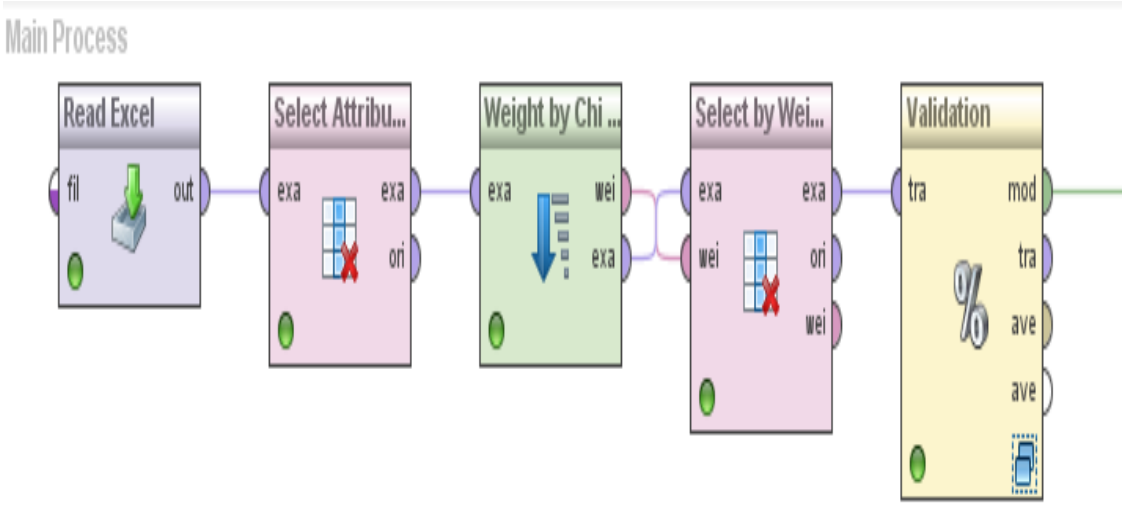
Tabel 2. Perbandingan jumlah term unik setelah *filtering* frekuensi

No.	Topik	Term_unik	%
1	Retail	6.099	42.40
2	Projek	1.852	12,88
3	Pendidikan	6.432	44,72
Jumlah		14.383	100

Sumber: Hasil Penelitian (2019)

Klasifikasi *Naïve Bayes*

Berdasarkan *dataset* yang didapatkan dari proses *pre-processing* dibuatlah sebuah model yang akan dipergunakan untuk percobaan klasifikasi dan menerapkan algoritma *Naïve Bayes* dalam model.



Sumber: Hasil Penelitian (2019)

Gambar 3. Design Model Klasifikasi *Naïve Bayes*

Pada gambar 3 dapat dijelaskan urutan prosesnya sebagai berikut: *1) Read Excel*, Operator yang digunakan untuk memuat file dataset training atau file vector document dalam bentuk excel. 2) *Select Attributes*, menentukan atribut mana saja yang akan dipergunakan dalam proses klasifikasi. 3) *Weight by Chi Square*, memberikan bobot nilai kepada setiap atribut. 4) *Select by Weight*, memilih sejumlah atribut dengan batasan bobot yang diinginkan. 5) *Validation*, Proses ini berisi proses klasifikasi dengan menggunakan *Naïve Bayes* dan penambahan *performance validation* dengan k-fold *validation*.

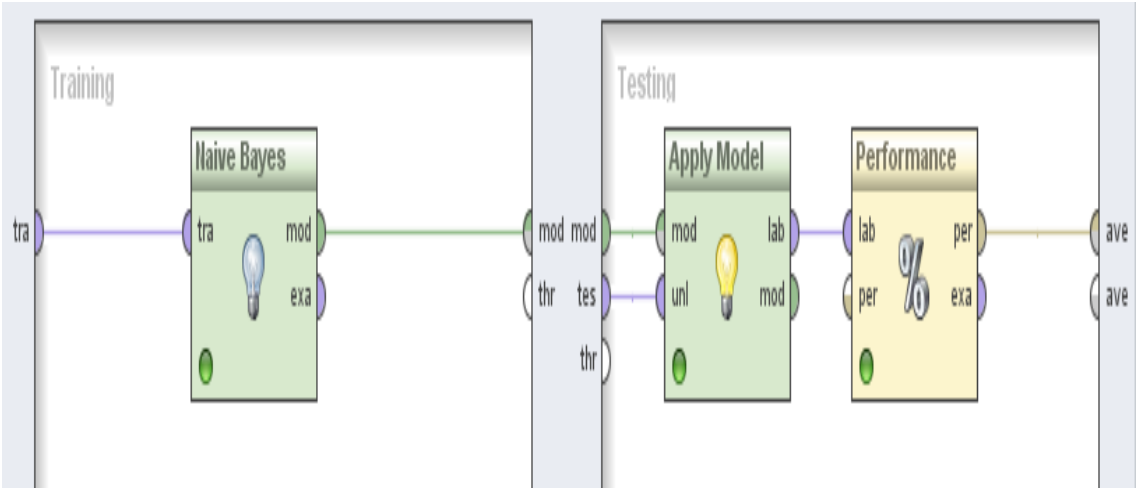
Evaluasi

Setelah model klasifikasi dibuat, selanjutnya adalah melakukan tahap evaluasi terhadap model klasifikasi yang di rancang. Adapun untuk evaluasi metode yang digunakan adalah *10-fold cross validation*. Pada gambar 4 merupakan desain proses yang digunakan untuk evaluasi.

10-fold validation bekerja dengan membagi *dataset* masukkan menjadi 10 bagian yang sama rata. 9 bagian kemudian ditraining sedangkan yang 1 bagian lainnya digunakan untuk testing. Proses ini diulang sebanyak 10 kali untuk setiap bagian sehingga setiap bagian dari kesepuluh bagian pernah menjadi bagian dari data untuk testing.

Operator *validation* melakukan proses *10-fold validation* ini kedalam algoritma *Naïve Bayes*. Untuk setiap percobaan akan dihitung akurasinya. Akurasi akhir adalah nilai rata-rata dari akurasi kesepuluh percobaan tersebut. Hasilnya dapat disajikan dalam bentuk *confusion matrix*. Berikut adalah *confussion matrix* hasil percobaan terhadap dataset yang ada.

E-ISSN: 2548-3587 ; 95 – 103



Sumber: Hasil Penelitian (2019)

Gambar 4. Desain proses *10-fold validation* untuk *Naïve Bayes*.

Semua Topik

Berdasarkan hasil percobaan yang dilakukan didapatkan bahwa nilai akurasi yang didapatkan dengan menggunakan 14.383 atribut untuk semua topik tweet adalah sebesar 50.66% seperti pada gambar 5 dibawah ini.

	true Tidak Terkait	true Terkait	class precision
pred. Tidak Terkait	1045	1076	49.27%
pred. Terkait	1053	1141	52.01%
class recall	49.81%	51.47%	

Sumber: Hasil Penelitian (2019)

Gambar 5. *Confusion Matrix* Semua Topik

Topik Retail

Berdasarkan hasil percobaan yang dilakukan didapatkan bahwa nilai akurasi yang didapatkan dengan menggunakan 6.099 atribut untuk topik retail adalah sebesar 49.89% seperti pada gambar 6.

	true Terkait	true Tidak Terkait	class precision
pred. Terkait	268	294	47.69%
pred. Tidak Terkait	623	645	50.87%
class recall	30.08%	68.69%	

Sumber: Hasil Penelitian (2019)

Gambar 6. *Confussion Matrix* Topik Retail

Topik Projek

Berdasarkan hasil percobaan yang dilakukan didapatkan bahwa nilai akurasi yang didapatkan dengan menggunakan 1.852 atribut dengan topik projek adalah sebesar 51.26% seperti pada gambar 7.

	true Terkait	true Tidak Terkait	class precision
pred. Terkait	57	62	47.90%
pred. Tidak Terkait	209	228	52.17%
class recall	21.43%	78.62%	

Sumber: Hasil Penelitian (2019)

Gambar 7. *Confussion Matrix* Topik Projek

Topik Pendidikan

Berdasarkan hasil percobaan yang dilakukan didapatkan bahwa nilai akurasi yang didapatkan dengan menggunakan 6.432 atribut topik pendidikan adalah sebesar 52.33% seperti pada gambar 8.

	true Tidak Terkait	true Terkait	class precision
pred. Tidak Terkait	220	263	45.55%
pred. Terkait	657	790	54.60%
class recall	25.09%	75.02%	

Sumber: Hasil Penelitian (2019)

Gambar 8. *Confussion Matrix* Topik Pendidikan

Untuk mengukur efektifitas dari *classifier* berdasarkan kategori yang dijadikan label, ada 3 jenis pengukuran (*metrics*) umum yang digunakan yaitu *precision*, *recall*, dan *F1 measure*. Berikut pengukuran perhitungan nilai F1 Measure yang didapatkan dalam penelitian ini, sebagai berikut adalah hasil yang diperoleh.

Semua Kategori

Berikut data dalam tabel nilai *F1 Measure* yang didapatkan dalam penelitian untuk semua kategori dapat dilihat pada tabel 3.

Tabel 3. *F1 Measure* untuk Semua Kategori

	Class precision	**Class recall**	**F1 Measure**
Terkait	52,01%	51,87%	51,94 %
Tidak Terkait	49,27%	49,81%	49,54%

Sumber: Hasil Penelitian (2019)

Kategori Retail

Berikut data dalam tabel nilai F1 Measure yang didapatkan dalam penelitian untuk kategori retail dapat dilihat pada tabel 4.

Tabel 4. *F1 Measure* untuk Kategori Retail

	Class precision	**Class recall**	**F1 Measure**
Terkait	47,69%	30.08%	36,90 %
Tidak Terkait	50,87%	68,69%	58,45%

Sumber: Hasil Penelitian (2019)

Kategori Proyek

Berikut data dalam tabel nilai F1 Measure yang didapatkan dalam penelitian untuk kategori Proyek dapat dilihat pada tabel 5.

E-ISSN: 2548-3587 ; 95 – 103

Tabel 5. *F1 Measure* untuk Kategori Proyek

	Class precision	Class recall	F1 Measure
Terkait	47,90%	21,43%	29,61%
Tidak Terkait	52,17%	78,62%	62,72%

Sumber: Hasil Penelitian (2019)

Kategori Pendidikan

Berikut data dalam tabel nilai F1 Measure yang didapatkan dalam penelitian untuk kategori pendidikan dapat dilihat pada tabel 6.

Tabel 6. *F1 Measure* untuk Kategori Pendidikan

	Class precision	Class recall	F1 Measure
Terkait	45,55%	25,09%	32,36%
Tidak Terkait	54,60%	75,02%	63,20%

Sumber: Hasil Penelitian (2019)

Berdasarkan hasil penelitian yang telah dilakukan didapatkan bahwa: 1) Algoritma klasifikasi *Naïve Bayes* dapat digunakan untuk mengklasifikasikan data tweet berbahasa Indonesia dengan mengacu pada keterkaitan terhadap topik-topik tertentu. 2) Hasil akurasi dari proses penelitian ini mendapatkan nilai <70% sehingga dapat dikatakan bahwa hipotesis kedua ini belum terbukti dikarenakan nilai yang ada dalam hipotesis adalah >70% untuk proses klasifikasi data tweet berbahasa Indonesia berdasarkan keterkaitan terhapap topik-topik tertentu.

4. Kesimpulan

Berdasarkan hasil percobaan yang dilakukan telah didapatkan bahwa nilai akurasi yang didapatkan dari semua topik termasuk ke dalam kategori kurang baik dengan tingkat kesalahan diatas 40% dan bahkan ada yang mempunyai akurasi dibawah 50%. Proses peningkatan efisiensi klasifikasi tweet ini bisa ditambahkan dengan menggunakan metode lain atau memperbaiki metode yang pernah dipergunakan sebelumnya. Adapun dari proses *cleansing tweet* bisa ditambahkan agar *term* yang dihasilkan bisa lebih bersih.

Referensi

Allah FA. 2008. Information Retrieval: Applications to English and Arabic Documents.

Chantar HKH. 2013. New Techniques for Arabic Document Classication. 165 p.

Eddyono AS. 2013. Twitter: Kawan, Sekaligus Lawan Bagi Redaksi Berita. J. Commun. Spectr. 3: 47–65.

Han J, Kamber M. 2006. Data mining: Data mining concepts and techniques, 2e. Stephan A, editor. San Fransisco: Morgan Kaufman. 1-745 p.

Manning CD, Raghavan P, Schutze H. 2008. Introduction to Information Retrieval. 307-309 p.

Masripah S. 2016. Komparasi Algoritma Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit. Bina Insa. ICT J. 3: 187–193.

Nurelasari E. 2018. Komparasi Algoritma Naive Bayes Dengan Support Vector Machine Berbasis Particle Swarm Optimization untuk Prediksi Kesuburan. Bina Insa. ICT J. 5: 61–70.

Rish I. 2001. An empirical study of the naive bayes classifier. IJCAI 2001 Work. Empir. methods

Artif. Intell. 3: 41–46.

Rodiansyah SF, Winarko E. 2012. Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification. Indones. J. Comput. Cybern. Syst. 6: 91–100.

Witten IH, Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques (Google eBook), 2e. Diane Cerra. 664 p.