EVALUATION OF CAMPAIGN CATEGORIES ON KITABISA.COM BY NAIVE BAYES CLASSIFIER METHOD

Dwi Putri Pertiwi[1], Wiranto[2], Rini Anggrainingsih[3]
Department of Informatics, Faculty of Mathematics and Natural Science,
Universitas Sebelas Maret
Email: [1]dwiputripertiwi@student.uns.ac.id, [2]wiranto@staff.uns.ac.id, [3]rini.anggrainingsih@staff.uns.ac.id

ABSTRACT

Kitabisa.com is a crowdfunding platform in Indonesia. To help donors choose a campaign that suit their preferences, Kitabisa.com categorizes campaigns manually when campaigners create campaign page. However, there are many options of categories offered so that is possible for campaigners choose wrong campaign category. The Naive Bayes Classifier method can be used to classify campaigns, so it generates recommendations for Kitabisa.com simplifies the campaign categories that can minimize campaigners mistake in choosing categories. Naive Bayes Classifier classifies each campaign based on title, short description, and full description. Document Frequency Improved (DFM) as feature selection implemented for improving the classification performance. This study used 7992 campaign data as training data and 888 campaign data as testing data. The testing applied 5 types of threshold value and using k-fold 10 cross-validations. The best results are shown in the model classification using 5 categories with 3.0 threshold level. The result is an average value of accuracy 90,89%, precision 89.24%, and recall 81.31%.

Keywords: *campaign, classification, document frequency improved, naive bayes classifier*

1. PENDAHULUAN

Kitabisa.com adalah platform untuk menggalang dana dan berdonasi secara online (*crowdfunding*) di Indonesia. Penggalangan dana yang diajukan bisa berbentuk apa saja, mulai dari membantu keluarga terdekat atau teman yang membutuhkan biaya medis, membangun infrastruktur hingga memberikan bantuan pendidikan kepada yang membutuhkan.

Untuk membantu donatur memilih penggalangan dana yang sesuai dengan preferensinya, Kitabisa.com mengelompokkan semua penggalangan dana menjadi 21 kategori, yaitu Balita & Anak Sakit, Bantuan Medis & Kesehatan, Beasiswa & Pendidikan, Bencana Alam, *Birthday Fundraising*, Difabel, *Family for Family*, Hadiah & Apresiasi, Karya Kreatif (Film, Buku, dll), Kegiatan Sosial, Kemanusiaan, Lingkungan, Menolong Hewan, Modal Usaha, Panti Asuhan, Produk & Inovasi, Rumah Ibadah, *Run For Charity*, Sarana & Infrastruktur, Zakat dan Kategori Lainnya [1].

Selama ini di Kitabisa.com pengelompokkan penggalangan dana dilakukan secara manual oleh penggalang dana pada saat membuat halaman penggalangan dana. Dengan banyaknya pilihan kategori yang ditawarkan tersebut, sehingga masih memungkinkan terjadinya *human error* pada saat memilih kategori penggalangan dana, yaitu 885 dari 9.844 data penggalangan dana yang digunakan pada penelitian ini.

Selama ini di Kitabisa.com pengelompokkan penggalangan dana dilakukan secara manual oleh penggalang dana pada saat membuat halaman penggalangan dana. Dengan banyaknya pilihan kategori yang ditawarkan tersebut, sehingga masih memungkinkan terjadinya *human error* pada saat memilih kategori penggalangan dana, yaitu 885 dari 9.844 data penggalangan dana yang digunakan pada penelitian ini. Hal ini menyebabkan donatur menemukan penggalangan dana yang tidak sesuai dengan kategori yang diinginkan.Oleh karena itu, diperlukan suatu metode yang dapat mengelompokkan penggalangan dana pada kategori yang sesuai, sehingga dapat menghasilkan kategori penggalangan dana yang optimal untuk meminimalisir kesalahan penggalang dana dalam memilih kategori. Metode pengolah teks yang diperlukan untuk pengelompokkan penggalangan dana adalah *text mining*. *Text mining* dapat mengekstrak informasi yang berguna dari data teks yang tidak terstruktur melalui identifikasi dan eksplorasi pola yang menarik [2].

Salah satu teknik dalam *text mining* adalah klasifikasi. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai prototipe untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya. Dalam membangun model, diperlukan suatu algoritma untuk membangunnya, yaitu disebut algoritma pelatihan (*learning algorithm*). Ada banyak algoritma pelatihan yang sudah dikembangkan oleh para peneliti, seperti *K-Nearest Neighbor*, *Naive Bayes Classifier*, C4.5, *Artificial Neural Network, Support Vector Machine,* dan sebagainya [3].

Algoritma yang banyak digunakan dalam klasifikasi teks salah satunya adalah *Naive Bayes Classifier* (NBC) yang memiliki beberapa kelebihan antara lain, sederhana, cepat dan berakurasi tinggi. Metode *Naive Bayes Classifier* (NBC) untuk klasifikasi teks menggunakan atribut kata yang muncul dalam satu dokumen sebagai dasar klasifikasinya. Algoritma klasifikasi *Naive Bayes* memanfaatkan teori probabilitas yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumya [4]. Penelitian mengenai klasifikasi teks telah dilakukan dengan algoritma *Naive Bayes* [5] [6] [7]. Selain itu penelitian mengenai pendekatan klasifikasi teks menggunakan *feature selection* juga telah dilakukan [8] [9].

Berdasarkan uraian tersebut, pengelompokkan penggalangan dana pada penelitian ini menerapkan metode klasifikasi *Naive Bayes Classifier* dengan 18 kategori yang digunakan. Tujuan penelitian ini adalah memberikan rekomendasi untuk Kitabisa.com dalam menyederhanakan kategori penggalangan dana, sehingga meminimalisir kesalahan penggalang dana dalam memilih kategori.

2. TEXT MINING

Text mining adalah proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah besar sumber data tak terstruktur. *Text mining* memiliki tujuan dan menggunakan proses yang sama dengan data mining, namun memiliki masukan yang berbeda. Masukan untuk text mining adalah data yang tidak (atau kurang) terstruktur, seperti dokumen Word, PDF, kutipan teks, dll,

sedangkan masukan untuk data mining adalah data yang terstruktur [2].

Struktur data yang baik dapat memudahkan proses komputerisasi secara otomatis. Pada *text mining*, informasi yang akan digali berisi informasi-informasi yang strukturnya sembarang. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam *data mining*, yang biasanya akan menjadi nilai-nilai numerik. Proses ini sering disebut *text preprocessing* . Setelah data menjadi data terstruktur dan berupa nilai numerik, maka data dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut [2].

Salah satu teknik yang sering digunakan dalam text mining adalah klasifikasi. Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu pembangunan model sebagai prototipe untuk disimpan sebagai memori dan penggunaan model tersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui di kelas mana objek data tersebut dalam model yang sudah disimpannya. Contoh aplikasi yang sering ditemui adalah pengklasifikasian jenis hewan, yang mempunyai sejumlah atribut. Dengan atribut tersebut, jika ada hewan baru, kelas hewannya bisa langsung diketahui [3].

3. FEATURE SELECTION

Pada text classification terdapat suatu permasalahan, yaitu adanya fitur-fitur yang berdimensi tinggi atau kata-kata unik yang sering muncul di semua kelas. *Feature selection* adalah metode yang digunakan untuk mengurangi jumlah fitur dengan memilih fitur yang relevan dengan kelasnya. Sehingga, *feature selection* dapat meningkatkan performa klasifikasi dan menghindari terjadinya masalah *overfitting data* [9].

Salah satu teknik seleksi fitur yang paling sederhana namun memiliki kinerja yang cukup baik adalah *Document Frequency Thresholding* yang bersifat class *independent*. Document Frequency merupakan banyaknya jumlah dokumen yang mengandung term tertentu. Term yang jarang muncul memiliki kemungkinan besar tidak memberikan informasi spesifik. Begitupun jika term tersebut terlalu sering muncul pada banyak dokumen, maka dianggap bahwa term tersebut merupakan term yang umum dan tidak akan mempengaruhi kinerja pediksi secara keseluruhan [10].

Dengan mereduksi fitur yang digunakan dalam proses klasifikasi, akan meningkatkan kinerja klasifikasi. Terdapat 3 syarat suatu data input dinyatakan sangat membantu dalam proses klasifikasi [7], yaitu:

3.1 *Concentration Degree*

Dalam suatu data set dengan berbagai kelas atau kategori, jika fitur kata muncul di satu atau sedikit kelas tapi tidak muncul di kelas lain, fitur kata tersebut memberikan informasi spesifik yang kuat dan sangat membantu dalam proses klasifikasi. Formula yang digunakan untuk menunjukkan rasio seberapa tinggi konsentrasi suatu fitur dalam suatu kelas ditunjukkan pada Persamaan 3.1

$$Concentration\ degree = \frac{DF(t, c_i)}{\left(1 + \sum_{j=1|j\neq i}^{n} DF(t, c_j)\right)} \quad (3.1)$$

Keterangan:
$DF(t, c_i)$ = jumlah dokumen kelas c_i yang mengandung *term t*
$DF(t, c_j)$ = jumlah dokumen kelas lain selain c_i yang mengandung term t

3.2 *Disperse Degree*

Jika suatu fitur kata muncul di satu kelas, fitur ini memiliki korelasi yang kuat dengan kelas tersebut. Sehingga, fitur sangat membantu proses klasifikasi apabila tersebar dalam satu kelas. *Disperse degree* menunjukkan tingkat persebaran suatu fitur dalam satu kelas. Sebagai contoh terdapat *m* kelas yang berbeda, C = {c_1, c_2, ..., c_m}, Persamaan 3.2 digunakan untuk mengetahui derajat persebarannya.

$$Disperse\ Degree = \frac{DF(t, c_i)}{N(c_i)} \quad (3.2)$$

Keterangan:
$N(c_i)$ = jumlah dokumen pada kelas c_i

3.3 *Contribution Degree*

Jika sebuah fitur kata hanya muncul di satu atau sedikit kelas dan kata tersebut tersebar dalam banyak dokumen pada kelas yang sama serta TF atau frekuensi kemunculannya pada satu dokumen tinggi, maka kata tersebut memiliki kontribusi yang tinggi terhadap kelas yang bersangkutan. *Contribution degree* merupakan penyederhanaan dari *Expectation Crossing Entropy (ECE). Expectation Crossing Entropy (ECE)* adalah salah satu jenis seleksi fitur yang mempertimbangkan TF (*Term Frequemcy*) dan relasi antara *term* / fitur kata dengan kelas. Nilai ECE yang besar menunjukkan bahwa fitur kata tersebut semakin informatif dan membantu proses klasifikasi. Formula ECE yang disederhanakan digunakan untuk menentukan nilai kontribusi suatu fitur terhadap suatu kelas. Formula yang telah disederhanakan ditunjukkan pada Persamaan 3.3.

$$Contribution\ degree = P(c_i, t)\ log \frac{p(c_i|t)}{p(c_i)} \quad (3.3)$$

Keterangan:
$P(c_i, t)$ = probabilitas gabungan kelas c_i dan *term* t
$P(c_i|t)$ = probabilitas kelas c_i yang mengandung *term t*
$P(c_i)$ = probabililtas kelas c_i

Berdasarkan 3 prinsip suatu fitur dikatakan sangat membantu dalam proses klasifikasi di atas dan metode *Document Frequency (DF)*, didapatkan metode *feature selection* baru yang disebut *Document Frequency Improved (DFM)* [7]. Persamaan *Document Frequency Improved (DFM)* ditunjukkan pada Persamaan 3.4.

$$DFM(t, c_i) = \frac{DF(t, c_i)}{\left(1 + \sum_{j=1|j\neq i}^{n} DF(t, c_j)\right)} + \frac{DF(t, c_i)}{N(c_i)} + p(c_i, t) log \frac{p(c_i|t)}{p(c_i)} \quad (3.4)$$

4. NAIVE BAYES CLASSIFIER

Teorema Bayes merupakan teorema yang mengacu konsep probabilitas bersyarat. Secara umum terorema Bayes dapat dinotasikan pada Persamaan 4.1.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (4.1)$$

Keterangan :
A : Data dengan *class* yang belum diketahui
B : Hipotesis data A merupakan suatu *class spesifik*
$P(B|A)$: Probabilitas hipotesis B berdasar kondisi A (*conditional/posterior probability)*
$P(B)$: Probabilitias hipotesis B (*prior probability*)
$P(A|B)$: Probabilitas A berdasar kondisi pada hipotesis B
$P(A)$: Probabilitas dari A

Metode Naive Bayes Classifier (NBC) merupakan salah satu metode yang dapat mengklasifikasikan teks. Kelebihan NBC adalah algoritmanya sederhana tetapi memiliki akurasi yang tinggi. Dalam

algoritma NBC setiap data direpresentasikan dengan pasangan atribut "$a_1, a_2, a_3, \ldots, a_n$" dimana a_1 adalah kata pertama, a_2 adalah kata kedua dan seterusnya. Sedangkan V adalah himpunan kategori dokumen. Pada saat klasifikasi, algoritma akan mencari probabilitas tertinggi dari semua kategori dokumen yang diujikan (V_{MAP}). Adapun persamaan V_{MAP} ditunjukkan pada Persamaan 4.2.

$$V_{MAP} = \underset{v_j \in V}{argmax} P(v_j) \prod P(a_i|v_j) \quad (4.2)$$

Dimana $argmax$ adalah fungsi yang mengembalikan indeks dari nilai maksimum dari sekumpulan himpunan data. Nilai $P(v_j)$ dihitung pada saat data *training*, didapat dengan rumus pada Persamaan 4.3

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (4.3)$$

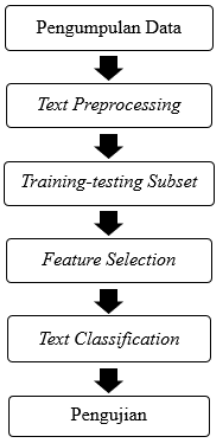
Dimana $|doc\ j|$ merupakan jumlah dokumen yang memiliki kategori j dalam $training$. Sedangkan $|training|$ merupakan jumlah dokumen dalam contoh yang digunakan untuk $training$. Untuk probabilitas kata a_i untuk setiap kategori $P(a_i|v_j)$, dihitung pada saat *training.*

$$P(a_i|v_j) = \frac{n_i + 1}{|n + kosakata|} \quad (4.3)$$

Dimana n_i adalah jumlah kemunculan kata a_i dalam dokumen yang berkategori v_j , sedangkan n adalah banyaknya seluruh kata dalam data dengan kategori v_j dan $|kosakata|$ adalah banyaknya kata dalam contoh *training* [11].

5. METODOLOGI

Metode penelitian pada klasifikasi kategori penggalangan dana ini terdiri dari 6 tahapan, yaitu tahap pengumpulan data, tahap *text preprocessing*, tahap *training-testing subset*, tahap *feature selection*, tahap *text classification* dan tahap evaluasi. Tahapan-tahapan tersebut digambarkan seperti Gambar 5.1.



Gambar 5.1 Tahapan Penelitian

5.1 Pengumpulan Data

Pada tahap ini dilakukan pengambilan data yang akan diolah pada penelitian ini yaitu penggalangan dana di website Kitabisa.com (https://kitabisa.com). Data diperoleh dari PT Kita Bisa Indonesia atau lebih dikenal dengan Kitabisa.com. Data yang akan digunakan sebagai masukan pada proses klasifikasi adalah bagian atribut judul, ajakan singkat, cerita lengkap dan kategori sebagai labelnya.

Sebelum dilakukan klasifikasi, data yang digunakan dipilah-pilah dahulu dan hanya mengambil data penggalangan dana dengan kategori dan atribut yang dibutuhkan saja. Data yang menggunakan bahasa selain bahasa indonesia, dan bagian atribut lainnya akan dihapus

5.2 *Text Preprocessing*

Pada tahap *preprocessing*, pertama kali data dilakukan proses *case folding*, proses ini akan menghilangkan *tag html*, mengubah semua huruf dalam dokumen menjadi huruf kecil dan menghapus karakter selain huruf 'a' sampai dengan 'z'. Data hasil *case folding* yang masih berupa kalimat akan dipotong berdasarkan tiap kata penyusunnya. Proses ini disebut dengan *tokenizing*. Hasil dari proses *tokenizing* ini menghasilkan fitur yang digunakan sebagai data pembelajaran mesin oleh *NBC*. Selanjutnya melakukan proses *stemming*, yaitu mengubah kata ke dalam bentuk dasarnya (*root word*) dengan menghilangkan imbuhan awalan dan akhirannya. Proses ini menggunakan *library* sastrawi yang menggunakan algoritma Nazief & Adrian [12]. Kata-kata yang dianggap tidak penting akan di hapus dari hasil *stemming*. Proses ini disebut *stopword removal. Stopword list* berbahasa Indonesia yang digunakan didapatkan dari penelitian yang dilakukan oleh Fadillah Z Tala (2003) [13].

Kitabisa.com tidak memiliki aturan tata bahasa dalam pembuatan penggalangan dana, sehingga terdapat penggalangan dana dengan bahasa formal, semi formal dan non formal, memiliki tipo, dan menggunakan berbagai jenis bahasa, misalnya campuran bahasa indonesia dan bahasa inggris. Oleh karena itu, dilakukan percobaan *text preprocessing* dengan *filtering* untuk menghindari kata-kata yang tidak beraturan tersebut menjadi informasi yang spesifik pada saat proses *feature selection*. Penelitian ini menerapkan 2 jenis *filtering,* yaitu filter KBBI dan filter manual. Filter KBBI, yaitu menghapus kata-kata yang tidak terdapat di Kamus Besar Bahasa Indonesia (KBBI) dengan melakukan *scraping* dari *website* KBBI (https://kbbi.web.id). Filter manual yaitu menghapus kata-kata yang tidak termasuk ke dalam penggabungan semua list fitur dari filter manual pada setiap kategori penggalangan dana berdasarkan hasil filter KBBI.

5.3 *Training-testing Subset*

Tahap *training-testing subset* adalah membagi data menjadi dua bagian yaitu data pelatihan dan data pengujian. Dalam penelitian ini digunakan metode *k-fold cross validation* dengan nilai *k-fold* = 10, yaitu membagi *dataset* menjadi 10 bagian dengan persentase sama untuk masing-masing kelas (*stratified partition*), 1/10 data sebagai data pengujian (*testing*) dan 9/10 data sebagai data pelatihan (*training*) [14]. Hasil dari *10-fold cross validation* adalah 10 kombinasi *training-testing subset* yang akan digunakan dalam proses *feature selection* dan *text classification.*

5.4 *Feature Selection*

Penerapan *feature selection* diawali dengan menghitung nilai *Document Frequency Improved* (*DFM*) dengan cara menghitung banyaknya dokumen yang mengandung *term t*, kemudian menghitung nilai *concentration degree, disperse degree,* dan *contribution degree*. Nilai DFM ini akan digunakan sebagai pertimbangan apakah kata akan digunakan sebagai fitur pada saat proses klasifikasi atau dihilangkan. Semakin besar nilai DFM maka fitur semakin dianggap penting.

Tahapan selanjutnya, yaitu menentukan besar *threshold* untuk fitur yang diseleksi. Penentuan *threshold* dilakukan beberapa kali sehingga ditemukan hasil paling optimal. Setelah ditentukan, fiturfitur dengan nilai di luar batas *threshold* akan di hapus. Fitur-fitur yang telah lolos seleksi akan digunakan sebagai fitur dalam proses klasifikasi dengan Naive Bayes Classifier.

Penerapan *feature selection* diawali dengan menghitung *nilai Document Frequency Improved* (DFM) dengan cara menghitung banyaknya penggalangan dana yang mengandung *term t* di setiap kombinasi *training subset*, kemudian menghitung nilai *concentration degree, disperse degree,* dan *contribution degree*.

Nilai DFM ini akan digunakan sebagai pertimbangan apakah kata akan digunakan sebagai fitur pada saat proses klasifikasi atau dihilangkan. Semakin besar nilai DFM maka fitur semakin dianggap penting.

Tahapan selanjutnya, yaitu menentukan besar *threshold* untuk fitur yang diseleksi. Penentuan threshold dilakukan beberapa kali sehingga ditemukan hasil paling optimal. Pada penelitian ini dilakukan pada Model 1-4 sebanyak 5 kali, yaitu 1.0, 1.5, 2.0, 2.5, dan 3.0, dan pada Model 5-6 sebanyak 9 kali, yaitu 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, dan 5.0. Penjelasan terkait masing-masing Model dijelaskan pada Bagian 5.5. Fitur – fitur dengan nilai di bawah batas *threshold* akan dihapus, sedangkan fitur-fitur yang telah lolos seleksi akan digunakan sebagai fitur dalam proses klasifikasi dengan *Naive Bayes Classifier.*

5.5 *Text Classification*

Proses *text classification* dibagi menjadi dua tahap, yaitu *training* dan *testing*. Pada proses *training* akan membangun model sebagai dasar dalam menentukan klasifikasi dengan menghitung probabilitas pada setiap kategori atau $P(v_j)$ dan probabilitas setiap kata pada setiap kategori atau $\prod P(a_i|v_j)$. Pada proses *testing* akan menghitung probabilitas penggalangan dana dari data *testing* dengan model yang telah dibangun pada saat proses *training*. Klasifikasi ditentukan dari nilai terbesar perhitungan V_{MAP} atau $\underset{v_j \in V}{argmax} P(v_j) \prod P(a_i|v_j)$ untuk setiap kategori. Model klasifikasi yang digunakan pada penelitian ini adalah sebagai berikut:

1. Model 1
   Model 1 adalah percobaan klasifikasi pada semua kategori dengan menggunakan hasil *text preprocessing* dengan filter KBBI.

2. Model 2
   Model 2 adalah percobaan klasifikasi pada semua kategori dengan menggunakan hasil *text preprocessing* dengan filter manual.

3. Model 3
   Model 3 adalah pengembangan dari Model 1. Kategori yang digunakan adalah hasil pengelompokkan kategori berdasarkan kategori yang mendapatkan perlakuan khusus oleh Tim Kitabisa.com, yaitu BS, MK, dan BA dan kategori yang saling memiliki keterkaitan. Kategori yang saling memiliki keterkaitan dipilih berdasarkan penyebaran hasil klasifikasi pada kategori lain dari hasil rata-rata *confusion matrix* dari 10 kombinasi *training-testing subset* pada Model 1-2.

4. Model 4
   Model 4 adalah pengembangan dari Model 2. Kategori yang digunakan sama seperti pada Model 3.

5. Model 5
   Model 5 adalah pengembangan dari Model 3. Kategori yang digunakan adalah hasil evaluasi dari Model 3-4 berdasarkan penyebaran hasil klasifikasi pada kategori lain dari hasil rata-rata *confusion matrix* dari 10 kombinasi *training-testing subset* pada Model 3-4.

6. Model 6
   Model 6 adalah pengembangan dari Model 4. Kategori yang digunakan sama seperti pada Model 5.

5.6 Pengujian

Tahapan terakhir dari penelitian yaitu pengujian untuk menganalisis performa hasil klasifikasi kategori penggalangan dana. Analisis dilakukan dengan menghitung presisi, recall dan akurasi dari hasil klasifikasi menggunakan *confusion matrix*. *Confusion matrix* terdiri dari elemen yang diklasifikasikan secara benar dan tidak benar dari setiap kelas. Salah satu manfaat dari confusion matrix adalah memberikan kemudahan untuk melihat terjadinya kesalahan sistem dalam meletakkan kelas klasifikasi [15]. Tabel 5.1 menunjukkan penggunaan confusion matrix.

Tabel 5.1 Confusion matrix

Kelas Sebenarnya	Kelas Hasil				
	C_1	C_2	...	C_n	
C_1	TP_C_1	Error	...	Error	Total_C_1
C_2	Error	TP_C_2	...	Error	Total_C_2
⋮	:	:	⋮	:	⋮
C_n	Error	Error	...	TP_C_n	Total_C_n
	Terprediksi_C_1	Terprediksi_C_2	...	Terprediksi_C_n	

Precision adalah ketepatan data yang terklasifikasi dengan benar. Rumus *precision* ditunjukkan pada Persamaan 5.1. *Recall* adalah sensitifitas data yang terklasifikasi terhadap kelas yang sebenarnya . Rumus *recall* ditujukkan pada Persamaan 5.2. *Accuracy* adalah hasil data yang terklasifikasi dengan benar secara keseluruhan. Rumus akurasi ditujukkan pada Persamaan 5.3.

$$accuracy = \frac{TP(c_1)+TP(c_2)+\cdots+TP(c_n)}{Total(c_1)+Total(c_2)+\cdots+Total(c_n)} \quad (5.1)$$

$$p(c_i) = \frac{TP(c_i)}{Terprediksi(c_i)} \quad (5.2)$$

$$r(c_i) = \frac{TP(c_i)}{Total(c_i)} \quad (5.3)$$

Keterangan:
$TP(c_i)$: Jumlah data yang terklasifikasi dengan benar pada kelas i
$Terprediksi(c_i)$: Jumlah seluruh data yang terprediksi di kelas i
$Total(c_i)$: Jumlah seluruh *dataset* pada kelas i

Parameter hasil akurasi, presisi, *recall* dapat dikatakan menghasilkan hasil klasifikasi yang bagus atau tidak dengan menggunakan pedoman parameter hasil klasifikasi yang ditampilkan di tabel 5.2 [17].

Tabel 5.2 Pedoman Parameter Hasil Klasifikasi

Rentang	Hasil Klasifikasi
90-100%	*Excellent Classification*
80-90%	*Good Classification*
70-80%	*Fair Classification*
60-70%	*Poor Classification*
50-60%	*Failure*

6. HASIL DAN PEMBAHASAN

6.1 Pengumpulan Data

Pengambilan data diperoleh dari dari PT Kita Bisa Indonesia atau lebih dikenal dengan Kitabisa.com. Data yang diambil adalah data penggalangan dana dari Januari 2016 sampai Juni 2018 dengan jumlah data sebanyak 11.243 penggalangan dana. Kategori Zakat, kategori Sarana & Infrastruktur dan Kategori Lainnya dihapus, sehingga data yang digunakan sebanyak 9.844 penggalangan dana. Kategori Zakat dihapus karena hanya digunakan oleh organisasi dan melalui verifikasi manual oleh Tim Kitabisa.com, kategori Sarana & Infrastruktur dihapus karena merupakan kategori yang dapat menyebar ke banyak kategori, sedangkan Kategori Lainnya dihapus karena memiliki konten yang terlalu luas. Dari hasil *text preprocessing* 9.844 penggalangan dana, terdapat 885 penggalangan dana yang tidak menghasilkan fitur setelah dilakukan

filter manual. Setelah dilakukan pengecekan, ternyata sebanyak 885 penggalangan dana tersebut adalah penggalangan dana dengan kategori yang tidak sesuai dengan isi kontennya, artinya penggalang dana salah memilih kategori pada saat membuat halaman penggalangan dana.

Dari hasil *text preprocessing* 9.844 penggalangan dana, terdapat 885 penggalangan dana yang tidak menghasilkan fitur setelah dilakukan filter manual. Setelah dilakukan pengecekan, ternyata sebanyak 885 penggalangan dana tersebut adalah penggalangan dana dengan kategori yang tidak sesuai dengan isi kontennya, artinya penggalang dana salah memilih kategori pada saat membuat halaman penggalangan dana. Penggalangan dana sebanyak 885 tersebut tidak digunakan dalam penelitian ini, sehingga data yang digunakan menjadi 8.959 penggalangan dana. Tabel 6.1 menampilkan jumlah penggalangan dana tiap kategori

Tabel 6.1 Jumlah Data Penggalangan Dana

Kategori	Jumlah Penggalangan Dana	Jumlah Kata Filter KBBI	Jumlah Kata Filter Manual
Balita & Anak Sakit	869	3,633	717
Bantuan Medis & Kesehatan	1,148	4,271	864
Beasiswa & Pendidikan	1,281	4,660	717
Bencana Alam	156	1,762	385
Birthday Fundraising	84	1,276	371
Difabel	123	1,943	469
Family For Family	1,115	2,956	577
Hadiah & Apresiasi	184	1,737	384
Karya Kreatif	189	2,556	515
Kegiatan Sosial	912	4,630	777
Kemanusiaan	456	3,214	633
Lingkungan	115	2,088	487
Menolong Hewan	190	1,813	455
Modal Usaha	756	3,253	593
Panti Asuhan	283	2,453	449
Produk & Inovasi	107	1,732	416
Rumah Ibadah	980	3,857	577
Run For Charity	11	482	155
Total	**8,959**	**48,316**	**9,541**

6.2 *Text Preprocessing*

Tahap *text preprocessing* pada penelitian ini terdiri dari lima proses, yaitu:

1. *Case folding*: menghilangkan *tag html*, mengubah semua huruf dalam dokumen menjadi huruf kecil dan menghapus karakter selain huruf 'a' sampai dengan 'z'
2. *Tokenizing:* memotong kalimat berdasarkan tiap kata penyusunnya
3. *Stemming:* mengubah kata-kata ke dalam bentuk dasarnya
4. *Stopword removal:* menghapus kata-kata yang dianggap tidak penting
5. *Filtering:* filter KBBI dan filter manual

Tabel 6.2 menampilkan hasil text preprocessing pada salah satu penggalangan dana

Tabel 6.2 Hasil *Text Preprocessing*

Kategori	Balita & Anak Sakit
Judul	Adam Fabumi Foundation
Ajakan Singkat	Terinspirasi dari semangat Adam Fabumi, AFF berniat untuk membantu bayi dan balita yang terlahir dengan kondisi spesial dan membutuhkan dukungan
Cerita Lengkap	- Adam Fabumi Foundation telah melaksanakan #Adamberbagi dengan membagikan sticker secara cuma-cuma kepada teman seperjuangan yang menggunakan sonde untuk makan dan minum - Berkomitmen untuk membantu perjuangan teman-teman Pejuang Cilik dengan membuka penggalangan dana bagi yang memiliki kondisi serupa dengan Adam Fabumi <img src=\" xxxx\"> Adam Fabumi Foundation (AFF) pertama kali didirikan Bulan Juli, tahun 2017..... (dst)
Data	Adam Fabumi Foundation Terinspirasi dari semangat Adam Fabumi, AFF berniat untuk membantu bayi dan balita yang terlahir dengan kondisi spesial dan membutuhkan dukungan - Adam Fabumi Foundation telah melaksanakan #Adamberbagi dengan membagikan sticker secara cuma-cuma kepada teman seperjuangan yang menggunakan sonde untuk makan dan minum - Berkomitmen untuk membantu perjuangan teman-teman Pejuang Cilik dengan membuka penggalangan dana bagi yang memiliki kondisi serupa dengan Adam Fabumi <img src=\" xxxx\"> Adam Fabumi Foundation (AFF) pertama kali didirikan Bulan Juli, tahun 2017..... (dst)
Text Preprocessing	"adam", "fabumi", "foundation", "inspirasi", "semangat", "adam", "fabumi", "aff", "niat", "bantu", "bayi", "balita", "lahir", "kondisi", "spesial", "butuh", "dukung", "adam", "fabumi", "foundation", "laksana", "adamberbagi", "sticker", "teman", "juang", "sonde", "makan", "minum", "komitmen", "bantu", "juang", "teman", "teman", "juang", "cilik", "buka", "galang", "dana", "milik", "kondisi", "rupa", "adam", "fabumi", "adam", "fabumi", "foundation", "aff", "kali", "juli", ... (dst)
Filter KBBI	"adam", "inspirasi", "semangat", "adam", "niat", "bantu", "bayi", "balita", "lahir", "kondisi", "spesial", "butuh", "dukung", "adam", "laksana", "teman", "juang", "makan", "minum", "komitmen", "bantu", "juang", "teman", "teman", "juang", "cilik", "buka", "galang", "dana", "milik", "kondisi", "rupa", "adam", "adam", "kali", "juli", ... (dst)
Filter Manual	"inspirasi", "bayi", "balita", "lahir", "kondisi", "spesial", "laksana", "juang", "makan", "juang", "juang", "kondisi", ... (dst)

6.3 *Training-testing subset*

Metode *k-fold cross validation* digunakan untuk membagi data pelatihan (*training*) dan data pengujian (*testing*) dengan nilai *k-fold* = 10 seperti yang telah dijelaskan di sub bagian 4.3. Sebelum melakukan proses *cross validation,* data akan dibulatkan ke bawah dengan kelipatan 10 untuk pemerataan masing-masing *subset*. Misalnya jumlah kategori Bantuan Medis & Kesehatan sebanyak 1148 dibulatkan menjadi 1140. Selanjutnya, masing-masing kategori akan diambil secara acak berdasarkan hasil perhitungan jumlah kategori yang telah dibulatkan tadi.

6.4 *Feature Selection*

Tahap *feature selection* ini menggunakan 10 kombinasi data *training* yang telah dilakukan pada tahap *training-testing subset*. Penerapan *feature selection* diawali dengan menghitung nilai *Document Frequency Improved (DFM)* untuk setiap kata yang ada di dalam data *training* menggunakan Persamaan 3.4. Hasil yang diperoleh kemudian disimpan dalam bentuk file masing-masing kategori penggalangan dana. Tahap selanjutnya yaitu menentukan besar *threshold* untuk fitur yang diseleksi. Pada penelitian ini menggunakan *threshold* 5-9 kombinasi, Model 1-4 sebanyak 5 kali, yaitu 1.0, 1.5, 2.0, 2.5, dan 3.0, dan Model 5-6 sebanyak 9 kali, yaitu 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, dan 5.0. Fitur – fitur dengan nilai di bawah batas *threshold* akan dihapus, sedangkan fitur-fitur yang telah lolos seleksi akan digunakan sebagai fitur dalam proses klasifikasi dengan *Naive Bayes Classifier*. Tabel 6.3, menampilkan salah satu hasil *feature selection* dengan metode DFM menggunakan hasil *text preprocessing* dengan filter KBBI.

Tabel 6.3 hasil *feature selection* dengan metode DFM

Kategori	Fitur: "sakit"	Kategori	Fitur: "sakit"
	DFM		DFM
Balita & Anak Sakit	2.57571494	Kegiatan Sosial	0.26347406
Bantuan Medis & Kesehatan	2.80863969	Kemanusiaan	0.45187982
Beasiswa & Pendidikan	0.06302728	Lingkungan	0.0513146
Bencana Alam	0.15174189	Menolong Hewan	0.56121367
Birthday Fundraising	1.16827535	Modal Usaha	0.16471849
Difabel	0.59502816	Panti Asuhan	0.06618239
Family For Family	0.91430539	Produk & Inovasi	0.12417771
Hadiah & Apresiasi	0.06976462	Rumah Ibadah	0.01035058
Karya Kreatif	0.05067343	*Run For Charity*	0.55647354

6.5 *Text Classification*

Tahap *text classification* ini menggunakan 10 kombinasi data *training-testing* yang telah dilakukan pada tahap *training-testing subset*. Klasifikasi dilakukan dengan menghitung nilai probabilitas masing-masing fitur yang telah terseleksi dengan DFM. Setiap *term* pada data *testing yang* sesuai dengan fitur-fitur tersebut dihitung probabilitas kemunculannya dalam kelas kemudian dikalikan dengan probabilitas kelasnya. Setelah menghitung semua kemungkinan kelas yang memiliki relasi dengan data pengujian, dipilih kelas dengan nilai probablitas paling tinggi sebagai hasil klasifikasinya. Hasil klasifikasi disajikan menggunakan *confusion matrix.*

6.6 Pengujian

Hasil *confusion matrix* yang dihasilkan dari tahap *text classification* dianalisis hasilnya dengan menghitung nilai presisi, *recall*, dan akurasi menggunakan Persamaan 5.1-5.3. Perhitungan nilai presisi, *recall*, dan akurasi pada penelitian ini dilakukan pada 10 kombinasi *training-testing subset*, 6 kombinasi Model, dan 5-9 kombinasi *threshold,* 5 kombinasi *threshold* pada Model 1-4 dan 9 kombinasi *threshold* pada Model 5-6, sehingga terdapat 380 hasil *precision*, *recall*, dan akurasi.

Hasil rata-rata pengujian pada Model 1 dan Model 2 dari 10 kombinasi *training-testing subset* ditunjukkan pada Tabel 6.4 – 6.5. Kolom *highlight* pada Tabel 6.4 - 6.5 merupakan *threshold* paling optimal karena memiliki hasil presisi, *recall*, dan akurasi yang paling tinggi dari semua *threshold* yang digunakan.

Tabel 6.4 Hasil Pengujian Model 1

Threshold	*Accuracy*	*Precision*	Recall
1	47.85%	45.42%	39.85%
1.5	37.05%	41.48%	31.72%
2	41.72%	39.44%	33.37%
2.5	38.89%	38.67%	28.29%
3	32.41%	39.94%	23.25%

Tabel 6.5 Hasil Pengujian Model 2

Threshold	*Accuracy*	*Precision*	Recall
1	47.07%	46.79%	44.31%
1.5	46.68%	48.67%	35.98%
2	46.79%	45.32%	33.20%
2.5	42.04%	44.71%	30.57%
3	41.14%	39.25%	30.30%

Berdasarkan pedoman parameter hasil klasifikasi pada Tabel 5.2, hasil pengujian Model 1 pada Tabel 6.4 dan Model 2 pada Tabel 6.5 telah gagal terklasifikasi. Tabel 6.6 menampilkan rata-rata hasil presisi dan *recall* tiap kategori pada Model 1-2 dengan *threshold* paling optimal, yaitu Model 1 dengan *threshold* 1.0 dan Model 2 dengan *threshold* 1.0. Kolom *highlight* pada Tabel 6.6 menujukkan nilai presisi dan *recall* dengan nilai kurang dari 60%.

Tabel 6.6 Hasil rata-rata presisi dan *recall* Model 1-2

KATEGORI	Model 1		Model 2	
	Precision	*Recall*	*Precision*	*Recall*
Balita & Anak Sakit	53.07%	69.77%	60.33%	57.91%
Bantuan Medis & Kesehatan	65.65%	33.51%	64.42%	34.91%
Beasiswa & Pendidikan	70.51%	44.14%	62.15%	46.56%
Bencana Alam	67.25%	55.33%	56.66%	66.00%
Birthday Fundraising	10.98%	62.50%	21.00%	75.00%
Difabel	15.34%	10.83%	45.24%	13.33%
Family For Family	36.19%	97.66%	59.16%	91.08%
Hadiah & Apresiasi	22.75%	10.00%	1.83%	5.56%
Karya Kreatif	39.25%	35.56%	57.08%	28.33%
Kegiatan Sosial	47.53%	10.00%	41.43%	11.65%
Kemanusiaan	40.90%	9.33%	24.35%	20.44%
Lingkungan	33.58%	57.27%	39.50%	58.18%
Menolong Hewan	86.30%	35.79%	83.23%	59.47%
Modal Usaha	84.04%	28.67%	77.53%	36.93%
Panti Asuhan	40.40%	60.71%	49.85%	49.29%
Produk & Inovasi	18.63%	12.00%	9.66%	30.00%
Rumah Ibadah	84.65%	74.29%	88.32%	62.86%
Run For Charity	0.59%	10.00%	0.53%	50.00%

Berdasarkan hasil rata-rata *confusion matrix* dari 10 kombinasi *training-testing subset* pada Model 1-2, hasil evaluasi Model 1-2, yaitu sebagai berikut:

1. Terdapat beberapa kategori yang saling memiliki keterkaitan, yaitu kategori Bantuan Medis & Kesehatan dengan kategori Balita & Anak Sakit, kategori *Family for Family* dengan kategori Modal Usaha, dan kategori Kegiatan Sosial dengan kategori Kemanusiaan
2. Terdapat banyak kategori yang memiliki keterkaitan dengan lebih dari 2 kategori, yaitu kategori *Birthday Fundraising*, kategori *Family For Family*, kategori Difabel, kategori Kegiatan Sosial, kategori Karya Kreatif, kategori Lingkungan, kategori Panti Asuhan, kategori Produk & Inovasi, kategori Bencana Alam, kategori Hadiah & Apresiasi, kategori Kemanusiaan, dan kategori *Run For Charity*
3. Mayoritas kategori penggalangan dana dapat terklasifikasi ke kategori Kegiatan Sosial dan kategori Kemanusiaan.

Percobaan pada Model 3 dan Model 4 dilakukan dengan mengurangi jumlah kategori penggalangan dana dengan cara mengelompokkan kategori penggalangan dana berdasarkan kategori yang saling memiliki keterkaitan dari hasil evaluasi Model 1-2 dan kategori yang mendapatkan perlakuan khusus oleh tim operasional Kitabisa.com, misalnya kategori Bencana Alam yang tidak dikenakan biaya administrasi. Tabel 6.7 menampilkan kategori yang digunakan, serta jumlah data pelatihan dan data pengujian pada Model 3-4

Tabel 6.7 Jumlah data pelatihan dan data pengujian pada Model 3-4

Kategori	Jumlah Penggalangan Dana	*Training*	*Testing*
Bencana Alam	150	135	15
Kesehatan • Balita & Anak Sakit • Bantuan Medis & Kesehatan	2,010	1,809	201
Menolong Hewan	190	171	19
Pendidikan • Beasiswa & Pendidikan	1,280	152	128
Rumah Ibadah	980	882	98
Sosial & Kemanusiaan • Sisa kategori lainnya pada penelitian ini	4,330	3,897	433

Hasil rata-rata pengujian pada Model 3-4 dari 10 kombinasi *training-testing subset* ditunjukkan pada Tabel 6.8-6.9. Kolom

highlight pada Tabel 6.8-6.9 merupakan *threshold* paling optimal karena memiliki hasil presisi, *recall*, dan akurasi yang paling tinggi dari semua *threshold* yang digunakan.

Tabel 6.8 Hasil pengujian Model 3

Threshold	*Accuracy*	*Precision*	*Recall*
1.0	67.11%	69.11%	44.43%
1.5	71.26%	74.47%	51.45%
2.0	71.11%	71.71%	51.50%
2.5	73.24%	70.55%	52.83%
3.0	73.02%	66.99%	47.89%

Tabel 6.9 Hasil pengujian Model 4

Threshold	*Accuracy*	*Precision*	*Recall*
1	74.35%	70.09%	66.14%
1.5	77.23%	72.99%	66.84%
2	76.73%	71.01%	68.00%
2.5	75.23%	69.54%	63.18%
3	74.07%	68.54%	54.95%

Berdasarkan pedoman parameter hasil klasifikasi pada Tabel 5.2, hasil pengujian Model 3 pada Tabel 6.8 telah berhasil terklasifikasi dengan cukup baik, namun hasil *recall* berada di rentang 50%-60% yang menunjukkan bahwa penggalangan dana tidak sensitif terhadap kategori, sedangkan hasil pengujian Model 4 pada Tabel 6.9 telah berhasil cukup baik, meskipun hasil *recall* berada di rentang 60%-70%. Tabel 6.10 menampilkan rata-rata hasil presisi dan *recall* tiap kategori pada Model 3-4 dengan *threshold* paling optimal, yaitu Model 3 dengan *threshold* 2.5 dan Model 4 dengan *threshold* 1.5. Kolom *highlight* pada Tabel 6.10 menujukkan nilai presisi dan *recall* dengan nilai kurang dari 60%.

Tabel 6.10 Hasil rata-rata presisi dan *recall* Model 3-4

KATEGORI	Model 5		Model 6	
	Precision	*Recall*	*Precision*	*Recall*
Bencana Alam	11.53%	23.33%	25.05%	60.00%
Kesehatan	88.03%	76.72%	88.98%	82.19%
Menolong Hewan	93.89%	35.79%	86.01%	60.00%
Pendidikan	68.60%	23.91%	70.02%	43.13%
Rumah Ibadah	91.30%	66.02%	92.46%	67.04%
Sosial & Kemanusiaan	69.96%	91.22%	75.40%	88.66%

Berdasarkan hasil rata-rata *confusion matrix* dari 10 kombinasi *training-testing subset* pada Model 3-4, hasil evaluasi Model 3-4, yaitu sebagai berikut:

1. Kategori Sosial & Kemanusiaan memiliki keterkaitan dengan semua kategori yang diperkuat dengan dasar penggalangan dana yang bersifat sosial
2. Kategori yang sangat berpengaruh terhadap kategori Sosial dan Kemanusiaan adalah kategori Bencana Alam, namun kategori Bencana Alam harus berdiri sendiri karena mendapatkan perlakuan khusus dari Tim Operasional Kitabisa.com.

Oleh sebab itu, percobaan pada Model 5 dan Model 6 dilakukan dengan hanya menggunakan 5 kategori, yaitu Bencana Alam, Kesehatan, Menolong Hewan, Pendidikan, dan Rumah Ibadah. Kelima kategori tersebut adalah kategori yang dapat berdiri sendiri atau memiliki sedikit kemungkinan saling berkaitan antar 5 kategori tersebut dan terdapat kategori yang mendapatkan perlakuan khusus oleh tim operasional Kitabisa.com. Kategori Sosial dan Kemanusiaan tidak dimasukkan ke dalam proses klasifikasi, namun jika terdapat penggalangan dana yang tidak memiliki fitur setelah dilakukan filter manual pada 5 kategori tersebut, maka penggalangan dana otomatis terklasifikasi sebagai kategori Sosial dan Kemanusiaan. Tabel 6.11 menampilkan jumlah data pelatihan dan data pengujian pada Model 5-6

Tabel 6.11 Jumlah data pelatihan dan data pengujian pada Model 5-6

Kategori	Jumlah Penggalangan Dana	*Training*	*Testing*
Bencana Alam	184	135	15
Kesehatan	2091	1809	201
Menolong Hewan	204	171	19
Pendidikan	1322	1152	128
Rumah Ibadah	990	882	98

Hasil rata-rata pengujian pada Model 5 dan Model 6 dari 10 kombinasi *training-testing subset* ditunjukkan pada Tabel 6.12-6.13. Kolom *highlight* pada Tabel 6.12-6.13 merupakan *threshold* paling optimal karena memiliki hasil presisi, *recall*, dan akurasi yang paling tinggi dari semua *threshold* yang digunakan. Model 5 dan Model 6 menggunakan 4 *threshold* tambahan karena jika hanya menggunakan *threshold* yang sama seperti Model 1-4, grafik selalu naik dan memungkinkan *threshold* paling optimal tidak berada pada rentang *threshold* 1.0-3.0.

Tabel 6.12 Hasil pengujian Model 5

Threshold	*Accuracy*	*Precision*	*Recall*
1.0	79.22%	77.44%	65.45%
1.5	84.38%	81.13%	71.33%
2.0	86.68%	83.28%	73.71%
2.5	87.79%	84.29%	75.53%
3.0	87.31%	85.65%	73.97%
3.5	88.50%	86.16%	75.27%
4.0	89.89%	87.46%	78.23%
4.5	89.54%	84.66%	74.43%
5.0	89.52%	82.38%	72.13%

Tabel 6.13 Hasil pengujian Model 6

Threshold	*Accuracy*	*Precision*	*Recall*
1.0	89.74%	84.24%	82.42%
1.5	89.67%	85.84%	81.01%
2.0	90.41%	88.19%	81.49%
2.5	90.61%	89.05%	80.60%
3.0	90.89%	89.24%	81.31%
3.5	90.63%	88.92%	80.82%
4.0	90.67%	89.31%	80.86%
4.5	89.61%	86.27%	75.32%
5.0	88.44%	79.44%	70.38%

Berdasarkan pedoman parameter hasil klasifikasi pada Tabel 5.2, hasil pengujian Model 5 pada Tabel 6.12 telah berhasil terklasifikasi dengan baik pada *threshold* 4.0, sedangkan hasil pengujian Model 6 pada Tabel 6.13 telah terklasifikasi dengan sangat baik pada *threshold* 3.0. Tabel 6.14 menampilkan rata-rata hasil presisi dan *recall* tiap kategori pada Model 5 dan Model 6 dengan *threshold* paling optimal, yaitu Model 5 dengan *threshold* 4.0 dan Model 6 dengan *threshold* 3.0. Kolom *highlight* pada Tabel 6.14 menujukkan nilai presisi dan *recall* dengan nilai kurang dari 80%.

Tabel 6.14 Hasil rata-rata presisi dan *recall* Model 5-6

KATEGORI	Model 5		Model 6	
	Precision	*Recall*	*Precision*	*Recall*
Bencana Alam	70.45%	62.00%	75.08%	69.33%
Kesehatan	91.53%	97.01%	91.75%	97.26%
Menolong Hewan	95.61%	55.79%	95.39%	63.16%
Pendidikan	87.46%	88.98%	87.27%	92.81%
Rumah Ibadah	92.25%	87.35%	96.69%	83.98%

Berdasarkan hasil rata-rata *confusion matrix* dari 10 kombinasi *training-testing subset* pada Model 5-6, hasil evaluasi Model 5-6 adalah sebagai berikut:

1. Penggalangan dana dengan kategori Bencana Alam memiliki keterkaitan dengan kategori Kesehatan, kategori Pendidikan dan kategori Rumah Ibadah
2. Penggalangan dana dengan kategori Menolong Hewan memiliki keterkaitan dengan kategori Kesehatan.
3. Model 6 memiliki nilai akurasi yang paling tinggi, tetapi model ini harus dilakukan pengecekan fitur-pada setiap

kategori secara berkala karena memungkinkan terdapat fitur baru yang dapat digunakan sebagai kata kunci pada kategori tertentu. Namun, meskipun Model 5 memiliki nilai akurasi yang tidak lebih tinggi dari Model 6, tapi Model 5 dapat otomatis melakukan filter fitur dengan *scraping* KBBI.

4. Model 5 memiliki hasil presisi dan *recall* yang berada di rentang 50%-100%, sedangkan Model 6 memiliki hasil presisi dan *recall* yang berada di rentang 60%-100%.

Berdasarkan hasil evaluasi pada Model 1-6, klasifikasi penggalangan dana pada Kitabisa.com menggunakan metode *Naive Bayes Classifier* menghasilkan kategori yang optimal pada Model 6 dengan *threshold* 3.0. Model 6 adalah klasifikasi kategori penggalangan dana menggunakan 5 kategori, yaitu Bencana Alam, Kesehatan, Menolong Hewan, Pendidikan, dan Rumah Ibadah. Kelima kategori tersebut adalah kategori yang dapat berdiri sendiri atau memiliki sedikit kemungkinan saling berkaitan antar 5 kategori tersebut dan terdapat kategori yang mendapatkan perlakuan khusus oleh Tim Kitabisa.com. Kategori Sosial dan Kemanusiaan tidak dimasukkan ke dalam proses klasifikasi, namun jika terdapat penggalangan dana yang tidak memiliki fitur setelah dilakukan filter manual pada 5 kategori tersebut, maka penggalangan dana otomatis terklasifikasi sebagai kategori Sosial dan Kemanusiaan. Kelima kategori tersebut dapat dijadikan rekomendasi untuk Kitabisa.com dalam menyederhanakan kategori supaya meminimalisir kesalahan penggalang dana dalam memilih kategori.

7. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa klasifikasi kategori penggalangan dana di Kitabisa.com menggunakan metode *Naive Bayes Classifier* menghasilkan kategori penggalangan dana yang optimal dengan menggunakan 5 kategori, yaitu Bencana Alam, Kesehatan, Menolong Hewan, Pendidikan, dan Rumah Ibadah. Kategori Sosial dan Kemanusiaan tidak dimasukkan ke dalam proses klasifikasi, namun jika terdapat penggalangan dana yang tidak memiliki fitur setelah dilakukan filter manual pada 5 kategori tersebut, maka penggalangan dana otomatis terklasifikasi sebagai kategori Sosial dan Kemanusiaan. Kelima kategori tersebut dapat dijadikan rekomendasi untuk Kitabisa.com dalam menyederhanakan kategori supaya meminimalisir kesalahan penggalang dana dalam memilih kategori penggalangan dana.

8. SARAN

Untuk pengembangan selanjutnya, diharapkan melakukan evaluasi kategori penggalangan dana menggunakan teknik *text clustering* dan menggunakan data penggalangan dana yang sesuai dengan kategorinya dengan cara mengevaluasi dan menentukan kategori penggalangan dana tersebut secara manual. Selain itu, diharapkan menggunakan data penelitian minimal 500 penggalangan dana pada setiap kategori. Penelitian berikutnya juga diharapkan menggunakan filter KBBI dengan cara melakukan scraping website https://kbbi.kemdikbud.go.id/ agar pengecekan kata sesuai KBBI menjadi jauh lebih akurat, namun website tersebut memiliki batasan akses, sehingga harus mengajukan akses terlebih dahulu kepada pihak yang berwenang.

9. DAFTAR PUSTAKA

[1] Kitabisa Team, "Kitabisa.com," PT Kita Bisa Indonesia, [Online]. Available: https://kitabisa.com/explore/all. [Accessed 13 December 2018].

[2] Feldman, R. & Sanger, J., The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, New York: Cambridge University Press, 2007.

[3] Prasetyo, E, Data Mining Konsep dan Aplikasi Menggunakan Matlab, Yogyakarta: Andi, 2012.

[4] D. Yanti, Analisis akurasi algoritma, Universitas Sumatera Utara, 2013.

[5] Rizqiyani, et.al, "Klasifikasi Judul Buku dengan Algoritma Naïve Bayes dan Pencairan Buku pada Perpustakaan Jurusan Teknik Elektro," *Jurusan Teknik Elektro, Universitas Negeri Semarang,* vol. 9, p. 2, 2019.

[6] Saptono, et.al., Text Classification using Naive Bayes Updateable Algorithm in SBMPTN Test Questions, Surakarta: Research Gate, 2016.

[7] Zheng & Feng, "Feature Selection Method Based on Improved Document Frequency," *TELOMNIKA,* vol. 2, pp. 905-910, 2014.

[8] Wongso, et.al., "News Article Text Classification in Indonesian Language.," *2nd Internasional Conference on Computer Science and Computational Inteligence 2017, ICCSCI 2017,* 2017.

[9] Chatcharaporn, et.al., Comparison of feature selection and Classification Algorithm Restaurant Dataset Classification, Thailand: Proceedings of the 11th Conference on Latest Advances in Systems Science & Computational Intelligence. 2012., 2012.

[10] Nallaswamy, R, "A Study on Analysis of SMS Classification Using Document Frequency Threshold," *I.J. Information Engineering and Electronic Business. MECS.,* pp. 44-50, 2012.

[11] Ariadi & Fithriasari, Klasifikasi Berita Indonesia Menggunakan Metode Naive Bayesian Classification dan Support Vector Machine dengan Conflix Stripping Stemmer, Surabaya: Jurnal Sains dan Seni ITS Vol. 4, No. 2, (2015) 2337-3520 (2301-928X Print), 2015.

[12] Librian. et.al, "High quality stemmer library for Indonesian Language.," Sastrawi, 2017. [Online]. Available: https://github.com/sastrawi/sastrawi. [Accessed 30 September 2018].

[13] Tala, F.Z, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," Universiteti van Amsterdam The Netherlands, 2003.

[14] Santosa, B., Data Mining: Teknik Pemanfaatam Data untuk Keperluan Bisnis, Yogyakarta: GRAHA ILMU, 2007.

[15] Rokach & Maimon, Data Mining With Decision Trees, Israel: World Scientific Publishing Co. Pte. Ltf, 2015.

[16] D. M. W. Powers, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, 2007.

[17] Gorunescu, Data Mining: Concepts, Models and Techniques, Springer, 2011.