JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer | ISSN: 2615-577X (Online)

https://ejournal.upi.edu/index.php/JATIKOM

Implementasi Multinomial Naive Bayes Untuk Klasifikasi Ujaran Kebencian Pada Dataset Kicauan (Twitter) Bahasa Indonesia

Implementation of Naive Bayes Multinomials for Classification of Hate Speech in the Twitter Dataset (Twitter) in Indonesian

Umar Syahid Aulia Rahman [#1], Yudi Wibisono [#2], Eddy Prasetyo Nugroho [#3]

[#]*Departemen Pendidikan Ilmu Komputer Fakultas Pendidikan Matematika dan Ilmu Pengetahuan AlamUniversitas Pendidikan Indonesia*

Bandung, Indonesia

[1]*yousyaheed@student.upi.edu,* [2]*yudi@upi.edu,*

[3]*eddypn@upi.edu*

Abstrak— **Pada paper ini kami membahas klasifikasi ujaran kebencian pada data kicauan (Twitter) dalam bahasa Indonesia dimana pada penelitian ini kami membangun dataset ujaran kebencian pada kicauan bahasa Indonesia dan melakukan pengklasifikasian dengan mengimplementasikan algoritma Multinomial Naive Bayes dengan menggunakan ekstraksi fitur term frequency – inverse document frequency (TF-IDF). Pada penelitian kami melakukan beberapa konfigurasi dalam modifikasi data training untuk mengatasi imbalanced dataset yaitu dengan menggunakan metode random oversampling dan random undersampling. Dari eksperimen tersebut kami melakukan evaluasi menggunakan confusion matrix dan didapatkan hasil implementasi metode Multinomial Naive Bayes dengan modifikasi data training menggunakan random oversampling dengan rasio data testing 10% memiliki hasil yang paling bagus dengan measure sebesar 0.5307.**

Kata Kunci: *Dataset Construction, Hate Speech Classification, Imbalanced Dataset, Multinomial Naive Bayes Classifier, Term Frequency Inverse Document Frequency.*

Abstract— **In this paper we discuss the classification of hate speech on tweet data (Twitter) in Indonesian where in this study we built a dataset of hate speech on Indonesian tweets and classified it by implementing the Naive Bayes Multinomial algorithm using the term frequency - inverse document frequency (TF-IDF) feature extraction. In our study, we performed several configurations in modifying the training data to overcome the imbalanced dataset by using random oversampling and random undersampling methods. From this experiment, we conducted an evaluation using the confusion matrix and the results of the implementation of the Multinomial Naive Bayes method with modified training data using random oversampling with a testing data ratio of 10% has the best results with a measure of 0.5307.**

Keywords: *Dataset Construction, Hate Speech Classification, Imbalanced Dataset, Multinomial Naive Bayes Classifier, Term Frequency Inverse Document Frequency.*

I. PENDAHULUAN

Pengguna Internet Indonesia meningkat pesat tercatat pada 2016 pengguna internet Indonesia sebanyak 79 juta orang[1], [2]. Pada tahun 2019 menurut data Kementerian Komunikasi dan Informatika (KEMKOMINFO) RI pengguna Internet di Indonesia mencapai 150 Juta pengguna dan dari angka tersebut 56% merupakan pengguna media sosial[3].

Salah satu media sosial yang banyak digunakan di Indonesia adalah Twitter berdasarkan data KEMKOMINFO Indonesia berada di urutan kelima pada pengguna media sosial Twitter[4]

Dengan kebebasan yang ada pada media sosial Twitter dan bertambahnya pengguna Twitter mengakibatkan terdapat banyak jenis kicauan di antara jenis kicauan tersebut terdapat kicauan yang mengandung ujaran kebencian[5]. Ujaran kebencian merupakan ungkapan atau bahasa yang digunakan untuk mengekspresikan kebencian terhadap suatu individu atau golongan dengan tujuan untuk menghinakan atau mempermalukan pihak yang ditujukan ujaran kebencian tersebut [6].

Laman Kompas.com merilis pada tahun 2015 terdapat 671 kasus ujaran kebencian yang sudah dilaporkan meliputi pencemaran nama baik, pelecehan, fitnah, provokasi, dan ancaman[7]. Permasalahan ini disikapi pemerintah Indonesia dengan menerbitkan surat edaran mengenai undang-undang no. 11 tahun 2008 mengenai ITE yang mengatur informasi dan transaksi elektronik dan surat edaran Kapolri SE/6/X/2015 mengenai penanganan ujaran kebencian[5], [8].

Dengan banyaknya data kicauan yang ada pada media sosial Twitter akan sulit untuk membedakan kicauan yang mengandung ujaran kebencian dan yang tidak mengandung ujaran kebencian oleh karna itu pada penelitian kali ini kami mengajukan solusi yaitu dengan pengklasifikasian kicauan yang mengandung ujaran kebencian dengan implementasi metode Multinomial Naive Bayes (MNB) yang sudah terbukti cukup optimal

dalam pengklasifikasian teks, terbukti dari penelitian sebelumnya mengenai sentiment analysis pada data Twitter dengan menggunakan berbagai algoritma di antaranya adalah MNB, Support Vector Machine (SVM) dan Logistic Regression (LR) yang menghasilkan akurasi sebesar 73.7% pada algoritma MNB dengan ekstraksi fitur menggunakan TF-IDF dan 73.9% menggunakan ekstraksi fitur count vectorizer [9].

Penelitian mengenai performa MNB selanjutnya adalah penelitian yang dilakukan oleh Daga Gupta [10] mengenai prediksi retweet dan like pada Twitter dengan menggunakan SVM, LR, Random Forest (RF), Neural Network (NN) dan MNB dengan fitur ekstraksi menggunakan TF-IDF dan Doc2vec menghasilkan akurasi untuk prediksi retweet MNB dengan TF-IDF pada data training sebesar 75.9% merupakan yang tertinggi dibanding algoritma lain yang dipakai pada penelitian tersebut, sedangkan akurasi untuk prediksi like dengan TF-IDF menghasilkan akurasi 77.7% dan merupakan yang hasil tertinggi dibanding algoritma lain pada penelitian tersebut.

Penelitian selanjutnya mengenai performa MNB pada klasifikasi teks terbukti pada penelitian [11] mengenai klasifikasi berita pada situs berita CNN Indonesia menggunakan MNB, Multivariate Bernouli dan SVM menghasilkan hasil terbaik ada pada klasifikasi menggunakan metode MNB dengan seleksi fitur menggunakan TF-IDF dengan precision 0.984 dan recall 0.984.

Pada awal dimulai penelitian ini kami tidak menemukan dataset yang relevan oleh karna itu pada penelitian ini kami membangun dataset ujaran kebencian pada kicauan bahasa Indonesia yang kami publikasi pada situs Kaggle .com. Namun pada akhir penelitian kami menemukan dataset yang relevan yang diunggah pada Mei 2020 oleh karna itu kami menjadikan dataset tersebut sebagai rujukan kami[12], [13].

Pada Penelitian ini kontribusi kami adalah sebagai berikut:

1. Membuat *dataset* ujaran kebencian pada kicauan bahasa Indonesia yang kami publikasikan pada situs Kaggle.com.
2. Membuat alur proses pembuatan *dataset* ujaran kebencian pada kicauan bahasa Indonesia.
3. Implementasi metode *Multinomial Naive Bayes* dengan ekstraksi fitur *term frequency inverse document frequency* untuk klasifikasi ujaran kebencian pada dataset yang telah kami bangun.

II. Penelitian Terkait

Penelitian mengenai ujaran kebencian pada data Twitter sudah banyak dilakukan oleh peneliti terdahulu [13]–[17] pada penelitian [16] peneliti menyediakan kriteria kicauan yang mengandung ujaran kebencian yang mana penulis menjadikannya sebagai rujukan untuk pembangunan dataset ujaran kebencian pada kicauan bahasa Indonesia.

Lalu pada penelitian [14] mengenai klasifikasi ujaran kebencian pada data kicauan bahasa Indonesia dalam kurun waktu Pemilihan Kepala Daerah (PILKADA) DKI Jakarta dengan proses pelabelan atau anotasi dilakukan oleh 30 orang yang dibagi tiap set di beri label oleh 3 orang dengan background yang berbeda dan hanya mengambil data yang disepakati oleh 3 dari 3 pemberi label sebagai ujaran kebencian, dan proses pembangunan model klasifikasi menggunakan metode Naive Bayes, Support Vector Machine, Bayesian Logistic Regression dan Random Forest Decision Tree.

Selanjutnya pada penelitian [15] penelitian mengenai klasifikasi ujaran kebencian pada kicauan bahasa Indonesia dengan label not abusive language, abusive languge but not offensive, offensive language menggunakan metode Naive Bayes, Support Vector Machine dan Random Forest Decision Tree.

Penelitian selanjutnya mengenai klasifikasi ujaran kebencian beserta target ujaran kebenciannya berupa ujaran kebencian terhadap individu, kelompok, agama, ras/etnis, jenis kelamin/ orientasi seksual, fisik/disabilitas dan lainnya dengan 3 level ujaran kebencian yaitu ujaran kebencian lemah, sedang dan kuat. Pengklasifikasian dalam penelitian tersebut menggunakan metode Support Vector Machine (SVM), Naive Bayes (NB), Random Forest Decision Tree (RFDT) classifier dan Binary Relevance (BR), Label Power-set (LP), dan Classifier Chains (CC) sebagai metode data transformasi [13].

Dan penelitian terakhir mengenai klasifikasi ujaran kebencian pada kicauan bahasa Indonesia menggunakan Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words [6].

III. Pembangunan Dataset

Pada Proses Pembangunan dataset kami merujuk pada penelitian[13], [14] dimana proses awal yaitu mengumpulkan data kicauan bahasa Indonesia. Kami mengumpulkan data kicauan pada tanggal 20 Maret 2020 sampai 3 April 2020. Sebelum mengumpulkan data kicauan bahasa Indonesia Penulis melakukan pemetaan ujaran kebencian menjadi 3 kategori berdasarkan penelitian [16], [17] yaitu ujaran kebencian pada ras/etnis, jenis kelamin/orientasi seksual dan individu/antar golongan. Dari 3 kategori itu penulis membuat daftar kata kunci untuk crawling data berikut ini kata kunci yang penulis gunakan dalam proses crawling data kicauan bahasa Indonesia (Tabel 1):

TABEL I

KATA KUNCI UNTUK PROSES CRAWLING

Ras/etnis	**Jenis Kelamin/orientasi seksual**
Cina/ China	cewek/ perempuan/ wanita
Arab	cowok/ laki-laki/ pria
Onta	gay/ homo/ maho
Sipit	lesbi
Negro/ hitam	transgender/ banci/ waria
Kata makian	**Individu/antar golongan**
Babi	Islam
Bego	Teroris
Tolol	Cingkrang
Idiot	Radikal
Tai	NU
Asu	HTI
Bangsat	Khilafah
Kampret	Kristen protestan
Bajingan	Nasrani
Banci	Yesus
Bencong	Salib
Pecun	Katolik
Geblek	Budha
Gila	Hindu
Goblok	Konghucu
Sarap	FPI
Udik/ kampungan	Cebong/ bong
Gembel	Komunis
Setan	Kafir
	Cina
	cacat

Pada penelitian ini kami menggunakan API Twitter untuk pengumpulan data kicauan dengan menggunakan *library* Tweepy menggunakan bahasa pemrograman python di bawah ini merupakan kode program yang kami gunakan:

```
import tweepy
import csv

consumer_key = 'consumer key anda'
consumer_secret = 'consumer secret anda'
access_token = 'token number anda'
access_token_secret = 'access token anda'
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

search_q= 'kata kunci'
filename = 'twitter_data_analysis '+search_q+'.csv'

with open (filename, 'a+', newline='') as csvFile:
    for tweet in tweepy.Cursor(api.search, q=search_q +"-filter:retweets",tweet_mode='extended', lang = 'id', count=100).items():
        csv.writer(csvFile).writerow([tweet.created_at,
        tweet.full_text,tweet.retweet_count,tweet.id])
csvFile.close()
```

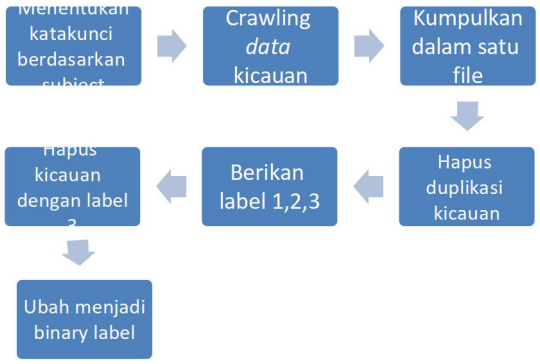
Tahap selanjutnya kami menggabungkan kicauan yang terpisah berdasarkan kata kunci dalam bentuk file csv menjadi satu file lalu disimpan ulang dalam bentuk xlsx, lalu kami menghapus kicauan yang berulang menggunakan bahasa pemrograman python. Di bawah ini kode program untuk menghapus kicauan yang berulang:

```
import pandas as pd
import xlrd
import xlsxwriter as xlsw

data = pd.read_excel (r'tempat disimpan file xlxs kicauan', dtype = str)
data.sort_values("TWEET", inplace = True)
df= data.drop_duplicates( subset=None, keep='first', inplace=False)
writer = pd.ExcelWriter('nama file baru untuk data yang sudah dihapus duplikasinya.xlsx', engine='xlsxwriter')
df.to_excel(writer, sheet_name="data", index=False)
writer.save()
print('OK')
```

Tahap selanjutnya kami memberikan label pada semua kicauan dengan label 1,2,3. Dimana label 1 untuk kicauan yang tidak mengandung ujaran kebencian, 2 untuk kicauan yang mengandung ujaran kebencian dan 3 untuk kicauan yang tidak relevan atau spam merujuk pada [16], [17] untuk proses pelabelan. Lalu kami menghapus kicauan dengan label 3 untuk membersihkan dataset dari kicauan spam dan terakhir kicauan dengan label 1 dan 2 diubah menjadi 0 dan 1 menjadi dataset dengan *binary class*.

Dari proses pembangunan *dataset* didapatkan data kicauan dengan label 0 sebanyak 5757 dan data kicauan dengan label 1 sebanyak 1126. Berikut ini alur proses pembangunan *dataset* berdasarkan alur proses di atas (Gambar 1):



Gambar 1 Alur proses pembangunan dataset

IV. PRAPROSES DAN EKSTRAKSI FITUR

Tahap selanjutnya setelah pembangunan dataset adalah melakukan praproses dataset. Di bawah ini tahap praproses yang dilakukan dalam penelitian ini merujuk pada penelitian sebelumnya mengenai praproses pada data Twitter [1]:

1. Ubah jadi huruf kecil
2. Hapus URL
3. Hapus mention
4. Hapus tanda baca
5. Hapus nomor
6. Hapus kata dengan hanya 1 huruf
7. Hapus stopword

Untuk stopword kami menggunakan library python sastrawi ditambah dengan kamus stopword yang kami bangun untuk melengkapi kata yang belum masuk dalam kamus stopword dalam library sastrawi [18].

Proses selanjutnya adalah tokenisasi dan stemming. tokenisasi dalam penelitian ini kami menggunakan library word_tokenize dalam modul NLTK dan stemming menggunakan fungsi stemmerfactory dalam library sastrawi.

Selanjutnya kami membagi dataset menjadi data training dan testing dengan rasio yang disesuaikan dengan konfigurais eksperimen.

Tahap selanjutnya adalah menangani dataset yang tidak seimbang menggunakan oversampling dan undersampling [19]. Proses ini dilakukan karna dataset yang tidak seimbang akan mempengaruhi hasil dari klasifikasi [19], [20].

Tahap selanjutnya pada penelitian ini kami menggunakan metode ekstraksi fitur TF-IDF dimana TF-IDF terbukti bagus untuk meningkatkan performa algoritma dalam masalah klasifikasi teks [9]–[11].

Pembobotan menggunakan TF-IDF meliputi berbagai tahap antara lain:

1. Penghitungan term frequency (tf_{dt})
2. Penghitungan document frequency (df_t)
3. Penghitungan inverse document frequency (idf_t)
4. Penghitungan term frequency – inverse document frequency

Rumus penghitungan term frequency dengan normalisasi adalah sebagai berikut:

$$tf_{dt} = \frac{f_{dt}}{\max_{j \in d} f_{dj}} \qquad \text{* MERGEFORMAT (0.1)}$$

Dimana:

- f_{dt} = frekuensi kemunculan *term* t pada dokumen j
- $\max f_{dj}$ = total *term* pada dokumen j .

Penghitungan df_t didapat dengan menghitung jumlah dokumen yang mengandung *term* t . Selanjutnya dari df_t dapat ditemukan nilai idf_{td} dengan rumus di bawah ini:

$$idf_{td} = \log\left(\frac{N}{df_t}\right) \qquad \text{* MERGEFORMAT (0.2)}$$

Tahap terakhir dari proses ekstraksi fitur yaitu penghitungan TF-IDF dengan cara mengalikan tf_{dt} dengan idf_{td} .

V. METODE PEMBANGUNAN MODEL

Tahap selanjutnya adalah pembangunan model menggunakan *Multinomial Naive Bayes* dengan rumus persamaan di bawah ini [21]–[23]:

$$C_{MAP} = \underset{c_k \in C}{\arg\max}\, P(c_k)\prod_{i=1}^{n} P(x_j \mid c_k) \qquad (2.1)$$

$$= P(c_k) \times P(x_{j=1} \mid c_k) \times \ldots \times P(x_{j=n} \mid c_k) \qquad (2.2)$$

Dimana n merupakan jumlah term pada dokumen testing dan parameter $P(x_j \mid c_k)$ didapatkan dengan menggunakan laplacean prior dengan menghitung jumlah kejadian x_j pada data training kelas c_k [22].

$$P(x_j \mid c_k) = \frac{1 + ft_i}{Nc_k + n} \qquad (2.3)$$

Dimana:

- ft_i = frekuensi *term* ke i pada *data training* pada kelas c_k

- Nc_k = Jumlah seluruh kata pada kelas c_k pada *data training*
- n = Jumlah seluruh kata unik pada *data training*

Implementasi TF-IDF pada *laplacean prior* [24]:

$$P(x_j \mid c_k) = \frac{W_t + 1}{(\sum W_t \in c_k) + n} \qquad (2.4)$$

Dimana:

- W_t = Nilai pembobotan TF-IDF dari term t di kelas c_k
- $\sum W_t \in c_k$ = Jumlah total pembobotan TF-IDF dari keseluruhan term yang berada di kelas c_k .
- n = Jumlah seluruh kata unik pada *data training*

VI. Eksperimen dan Hasil

Eksperimen dilakukan dengan melakukan pengklasifikasian *data testing* dengan konfigurasi rasio *data testing* 10 dan 25 serta menggunakan konfigurasi *data training modification* menggunakan *oversampling, undersampling* dan tidak memakai *data training modification*.

Pengukuran hasil konfigurasi menggunakan algoritma *confusion matrix* yang merupakan metode yang merepresentasikan informasi mengenai seberapa sering suatu kelas diprediksi sebagai kelas tersebut dan seberapa sering kelas tersebut salah diprediksi sebagai kelas lain [25]. Akurasi dari klasifikasi yang diukur menggunakan metode *confusion matrix* biasanya ditampilkan dalam ringkasan dalam bentuk *precision, recall/sensitivity*, *f-measure* dan akurasi[19]. Di bawah ini merupakan tabel konfigurasi eksperimen dan hasil (Tabel 2):

TABEL II
KONFIGURASI DAN HASIL EKSPERIMEN

Konfigurasi	**1**	**2**	**3**
Data Uji	*Test 25%*	Test 10%	Test 10%
Praproses	Ya	Ya	Ya
Stemming	Ya	Ya	Ya
Data training modification	Tidak	*Oversampling*	*Undersampling*
TP	29	91	110
FN	232	26	7
TN	1434	436	216
FP	19	135	355
Recall	0.1082	0.7778	0.9402
Precision	0.6042	0.4027	0.2366
F-Measure	0.1835	0.5307	0.3781

VII. Analisis Hasil

Dari konfigurasi eksperimen pada Tabel 2 didapatkan hasil f-measure yang beragam dan dapat diketahui bahwa hasil eksperimen yang paling bagus adalah pada konfigurasi 2 dengan melakukan data training modification dengan oversampling dengan rasio data testing 10% dengan nilai f-measure 0.5307. Dari hasil tersebut dapat disimpulkan bahwa dengan bertambahnya frekuensi suatu term dalam kelas minoritas dapat meningkatkan ketepatan klasifikasi kelas tersebut namun hal tersebut dapat mempengaruhi ketepatan klasifikasi kelas mayoritas terbukti dari hasil false positive sejumlah 135 sehingga menyebabkan hasil precision yang tidak bagus dan menyebabkan hasil f-measure yang tidak bagus.

Pada klasifikasi menggunakan data training modification undersampling dengan rasio data testing 10%pada tabel 2 konfigurasi 3 didapatkan f-measure yang kurang baik yaitu 0.3781, hal ini disebabkan hasil precision yang tidak bagus karna berkurangnya data training kelas mayoritas sehingga nilai false positive menjadi sangat banyak. dan juga dari tabel di atas diketahui bahwa ketepatan klasifikasi kelas minoritas cukup baik yaitu sejumlah 110 dari 117 data kelas minoritas, tapi terjadi penurunan yang signifikan dalam ketepatan klasifikasi kelas mayoritas dimana pada data training modification oversampling didapatkan ketepatan kelas mayoritas sejumlah 436 dari 571, sedangkan pada data training modification undersampling ketepatan klasifikasi kelas mayoritas sejumlah 216 dari 571 data.

Dengan melihat hasil eksperimen ini dapat disimpulkan data training yang dimodifikasi dengan undersampling menjadikan banyak kata pada kelas mayoritas yang berkurang frekuensinya sehingga mengurangi bobot kata tersebut dan juga menjadikan banyak kata pada kelas mayoritas yang seharusnya ada pada kamus TF-IDF menjadi tidak ada sehingga menyebabkan banyak data kelas mayoritas yang diprediksi sebagai kelas minoritas dan menyebabkan turunnya precision dari model klasifikasi tersebut.

Dari Tabel 2 juga dapat disimpulkan bahwa dataset yang tidak seimbang tidak dianjurkan diproses langsung karna dapat mempengaruhi hasil klasifikasi untuk lebih condong pada kelas mayoritas terbukti dari hasil konfigurasi 1 pada tabel di atas. Dan juga dapat disimpulkan bahwa modifikasi data training dapat meningkatkan akurasi dan f-measure dari model klasifikasi.

Untuk memvalidasi kode program implementasi Mulitnomial Naive Bayes yang sudah kami buat kami melakukan percobaan klasifikasi dengan menggunakan library scikit-learn Multinomial NB dan menghasilkan hasil f-measure yang sesuai dengan hasil eksperimen dengan menggunakan kode program yang telah kami implementasikan. Kami juga mencoba melakukan eksperimen klasifikasi dengan dataset yang digunakan pada penelitian sebelumnya yang dilakukan oleh [13]Kami mencoba dataset tersebut dengan kode program kami dan menghasilkan f-measure sebesar 0.7814 lalu

kami juga mencoba mencampurkan dataset tersebut dengan dataset yang sudah kami buat dan menghasilkan f-measure sebesar 0.6748. Hal ini membuktikan bahwa kode program yang kami buat sudah benar dan sudah cukup baik melihat hasil penelitian sebelumnya dimana hasil terbesar pada eksperimen kedua pada penelitian tersebut adalah akurasi sebesar 0.7736 dengan menggunakan Random Forest Decision Tree dan hasil implementasi kami menggunakan dataset tersebut menghasil kan f-measure sebesar 0.7814.

Dari hasil eksperimen pada Tabel 2 dapat dilihat bahwa hasil dari proses testing terbesar adalah sebesar 0.5307 dimana hasil tersebut kurang baik dibandingkan dengan hasil penelitian sebelumnya oleh [14]. Pada penelitian [14] hasil terbaik sebesar 81.7, hipotesis kami hal tersebut didapatkan karna proses pengambilan dataset yang lebih fokus kepada satu tema kata kunci sehingga dataset yang didapatkan menghasilkan pola yang baik sedangkan dataset yang kami buat menggunakan kata kunci yang cukup banyak dan melebar.

Selanjutnya kami membandingkan proses anotasi atau pelabelan pada penelitian [14]dengan penelitian kami dimana pada penelitian tersebut tiap kicauan dilabeli oleh 3 orang dengan total 30 pemberi label, lalu menghapus kicauan yang tidak disepakati oleh 3 orang sehingga label yang dihasilkan pada dataset tersebut sudah melalui kesepakatan 3 orang dan dapat dianggap sudah divalidasi sedangkan dataset yang kami buat hanya dilabeli oleh penulis tanpa adanya validasi oleh ahli atau validasi 3 orang seperti penelitian tersebut sehingga hal tersebut dapat menyebabkan hasil dataset yang buruk yang dapat menyebabkan hasil f-measure yang kecil.

VIII. KESIMPULAN

Proses Pembangunan dataset yang menggunakan fokus tema yang banyak menyebabkan hasil implementasi Multinomial Naive Bayes pada dataset tersebut menjadi tidak baik. Proses pembangunan dataset yang hanya memakai satu pemberi label menjadikan dataset menjadi subyektif dan tidak dapat divalidasi kebenaran ataupun ketepatan dataset. Harapan kami untuk penelitian selanjutnya apabila ingin memakai dataset yang kami buat agar melakukan pelabelan ulang dan melakukan validasi dengan cara kesepatan 100%.

Dari eksperimen yang telah dilakukan dapat disimpulkan bahwa algoritma Multinomial Naive Bayes dalam pengklasifikasian ujaran kebencian pada dataset kicauan bahasa Indonesia menghasilkan hasil yang kurang baik yaitu f-measure 0.5307 pada data training yang menggunakan oversampling dengan rasio data testing 10%. Hal ini disebabkan proses pembangunan dataset yang tidak sesuai standar.

REFERENSI

[1] S. Mujilahwati, "Pre-Processing Text Mining Pada Data Twitter," *Semin. Nas. Teknol. Inf. dan Komun.*, vol. 2016, no. Sentika, pp. 2089–9815, 2016.

[2] M. S. H. Mujahiddin, "Model Penggunaan Media Sosial," *J. Interak. J. Ilmu Komun.*, vol. 1, pp. 142–155, 2017.

[3] L. Rizkinaswara, "Pengguna Internet Indonesia," *KEMENTRIAN KOMUNIKASI DAN INFORMATIKA RI*, 2019. [Online]. Available: https://aptika.kominfo.go.id/2019/08/penggunaan-internet-di-indonesia/#:~:text=Di Indonesia%2C pengguna internet mencapai,dengan persentase penetrasi sebesar 53%25. [Accessed: 09-Jun-2020].

[4] KEMKOMINFO, "Kominfo : Pengguna Internet di Indonesia 63 Juta Orang," 2017. [Online]. Available: https://kominfo.go.id/index.php/content/detail/3415/Kominfo%3A Pengguna Internet di Indonesia 63 Juta Orang/0/berita_satker. [Accessed: 09-Jun-2020].

[5] R. R. Setyaningrum and R. A. Dwilestari, "Literacy Class Based on Pop Up Book As Media To Minimize the Effect of Hate Statements (Saracen) on Teenage of Adolescent In Digital Era," no. 2008, pp. 536–541, 2017.

[6] M. M. Munir, M. A. Fauzi, and R. S. Perdana, "Implementasi Metode Backpropagation Neural Network berbasis Lexicon Based Features dan Bag of Words Untuk Identifikasi Ujaran Kebencian Pada Twitter," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 10, pp. 3182–3191, 2018.

[7] A. N. K. Movanita, "2016, Konten Berisi Ujaran Kebencian Paling Banyak Diadukan ke Polisi," 2016. [Online]. Available: https://nasional.kompas.com/read/2017/03/26/08465611/2016.konten.berisi.ujaran.kebencian.paling.banyak.diadukan.ke.polisi. [Accessed: 09-Jun-2020].

[8] Polri, *Surat Edaran Kapolri Mengenai Penanganan Ujaran Kebencian*. 2015.

[9] M. D. Devi and N. Saharia, "Learning Adaptable Approach to Classify Sentiment with Incremental Datasets.," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 2426–2434, 2020.

[10] I. Daga, A. Gupta, R. Vardhan, and P. Mukherjee, "Prediction of Likes and Retweets Using Text Information Retrieval," *Procedia Comput. Sci.*, vol. 168, pp. 123–128, 2020.

[11] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017.

[12] I. F. Putra, "Indonesian Abusive and Hate Speech Twitter Text," 2020. [Online]. Available: https://www.kaggle.com/ilhamfp31/indonesian-abusive-and-hate-speech-twitter-text. [Accessed: 03-Jun-2020].

[13] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," in *Proceedings of the Third Workshop on Abusive Language Online*, 2019, pp. 46–57.

[14] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 233–238.

[15] M. O. Ibrohim and I. Budi, "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media," *Procedia Comput. Sci.*, vol. 135, pp. 222–229, 2018.

[16] Z. Waseem and D. Hovy, "Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter," *Proc. NAACL student Res. Work.*, pp. 88–93, 2016.

[17] M. C. Anam and M. Hafiz, "Surat Edaran Kapolri Tentang Penanganan Ujaran Kebencian (Hate Speech) dalam Kerangka Hak Asasi Manusia," *J. Keamanan Nas.*, vol. 1, no. No. 3, pp. 342–364, 2015.

[18] H. A. Robbani, "Sastrawi 1.0.1." Python Package Index, 2016.

[19] H. He and Gracia, "Learning from imbalanced data," *IEEE Trans. Knowl. data Eng.*, vol. 21, no. 9, pp. 1263-1284., 2009.

[20] G. Lemaˆ, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, 2017.

[21] D. D. Lewis, "Naiv e (B a y e s) at Forty : The Independence

Assumption in Information Retrieval," in *European conference on machine learning*, 1998, no. x, pp. 4–15.

[22] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 3339, pp. 488–499, 2004.

[23] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telemat. Informatics*, vol. 48, p. 101345, May 2020.

[24] A. Rahman, W. Wiranto, and A. Doewes, "Online News Classification Using Multinomial Naive Bayes," *ITSMART J. Teknol. dan Inf.*, vol. 6, no. 1, pp. 32–38, 2017.

[25] S. Ruuska, W. Hämäläinen, S. Kajava, M. Mughal, P. Matilainen, and J. Mononen, "Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle," *Behav. Processes*, vol. 148, pp. 56–62, Mar. 2018.