

Klasifikasi Konten Moderasi Beragama pada Media Sosial dengan Menggunakan Algoritma Naïve Bayes Multinomial Teks

Hery Mustofa^{1,*}, Adzhal Arwani Mahfudh², Nuurun Najmi Qonita³

^{1,2,3} Fakultas Sains dan Teknologi, Teknologi Informasi, Universitas Islam Negeri Walisongo, Semarang, Indonesia

Email: ^{1,*}herymustofa@walisongo.ac.id, ²adzhal@walisongo.ac.id, ³2208096032i@Student.walisongo.ac.id

*) Email Penulis Utama

Abstrak– Fenomena peningkatan penggunaan media sosial telah menjadikan platform ini sebagai media utama penyebaran narasi moderasi beragama. Namun, eskalasi konten tersebut tidak selalu diiringi dengan akurasi, sering kali memuat narasi yang berpotensi menyesatkan. Untuk mengatasi tantangan ini, diperlukan pendekatan *text mining* guna mengklasifikasikan konten narasi moderasi beragama secara otomatis dan akurat. Penelitian ini berhasil melakukan klasifikasi teks moderasi beragama di media sosial menggunakan algoritma *Multinomial Naive Bayes*. Dengan memanfaatkan dataset berjumlah 1996 *tweet* yang telah dilabeli oleh ahli, dari hasil evaluasi diketahui model ini menunjukkan kinerja yang baik, ditandai dengan *accuracy* 0.91, *precision* 0.93, *recall* 0.89, dan *F1-Score* 0.90. Temuan ini mengindikasikan bahwa *Multinomial Naive Bayes* merupakan metode yang efektif untuk identifikasi dan klasifikasi konten moderasi beragama di media sosial.

Kata Kunci: Moderasi beragama, machine learning, teks mining, *Naive Bayes*

Abstract– The significant increase in social media usage has positioned these platforms as a primary medium for disseminating religious moderation narratives. However, the escalating volume of such content is not always accompanied by accuracy, often containing potentially misleading narratives. To address this challenge, a text mining approach is required to automatically and accurately classify religious moderation content. This research successfully performed text classification of religious moderation on social media using the Multinomial Naive Bayes algorithm. Utilizing a dataset of 1996 tweets expertly labeled by specialists, the evaluation results demonstrate that this model exhibits good performance, characterized by an accuracy of 0.91, precision of 0.93, recall of 0.89, and an F1-Score of 0.90. These findings indicate that Multinomial Naive Bayes is an effective method for identifying and classifying religious moderation content on social media.

Keywords: Religious moderation, machine learning, *text mining*, *Naive Bayes*

1. PENDAHULUAN

Pertumbuhan pesat pengguna media sosial memberikan dampak signifikan terhadap dinamika pertukaran informasi dan interaksi sosial masyarakat. Aspek signifikan yang tak boleh diabaikan adalah konten moderasi beragama pada media sosial [1]. Media sosial dapat dijadikan sarana untuk menarasikan moderasi beragama [2]. Konten moderasi beragama di media sosial memiliki peran yang sangat penting dalam menyiarkan moderasi beragama [3]. Konten-konten ini dapat memiliki dampak besar terhadap opini publik, toleransi antar kelompok, dan stabilitas sosial.

Media sosial menjadikan sarana kelompok maupun individu untuk berbagi ide, pandangan, dan nilai-nilai mereka. Media sosial dijadikan saluran atau sarana individu untuk saling berinteraksi, komunikasi, diskusi [4]. Media sosial, dapat dijadikan sarana dalam menggerakkan Moderasi beragama. Konten di media sosial dapat menimbulkan dampak positif maupun negatif, konten yang beredar tersebut kadang juga mengarah pada tindakan ekstremisme maupun tindakan negatif lainnya [5]. Konten moderasi beragama yang beredar di media sosial saat ini sangat banyak sekali, sehingga risiko konten moderasi beragama yang tidak benar akan memicu konflik atau ketegangan antar kelompok. Oleh sebab itu, diperlukan suatu pendekatan model yang efektif untuk mengidentifikasi dan mengklasifikasikan konten-konten ini secara otomatis.

Moderasi beragama mempunyai makna cara beragama dengan menggunakan pendekatan jalan tengah. Dengan begitu seseorang tidak berlebihan-lebihan ketika menjalankan ajaran agamanya [2]. Moderasi beragama sangat dibutuhkan dalam setiap tindakan seseorang untuk selalu menjaga toleransi beragama. Menurut KBBI, moderasi diartikan sebagai pengurangan kekerasan dan penghindaran sikap berlebihan atau ekstrem. [6]. Jika dalam Agama Islam ada konsep *wasythiyah*, sedangkan di agama kristen ada konsep *golden mean*, di Agama Budha juga ada konsep *madyhamika*, dalam *konghucu* juga ada konsep *Zhong Yong*. Ke semuanya istilah tersebut mengacu pada satu makna yang sama yaitu memilih jalan tengah di antara dua kutub tidak *extrem* dan tidak berlebih-lebihan merupakan sikap beragama yang paling ideal [2].

Moderasi beragama dalam konteks media sosial merupakan tema yang semakin relevan di era digital saat ini. Media sosial telah menjadi platform utama untuk menyebarkan ide-ide, termasuk nilai-nilai moderasi beragama. Dalam konteks Indonesia, negara yang memiliki keragaman suku, ras, dan agama, moderasi beragama

sangat penting untuk menciptakan harmoni dan toleransi antar kelompok [7]. media sosial memiliki kekuatan besar dalam membentuk perilaku manusia, dan jika tidak digunakan dengan bijak, dapat menjadi bumerang bagi masyarakat. Oleh karena itu, literasi media sosial menjadi kunci dalam memasyarakatkan sikap moderasi beragama.

Penggunaan *text mining* dalam media sosial terkait dengan narasi beragama memiliki dampak strategis dalam memahami dinamika narasi keagamaan di era digital. Dengan menggabungkan teknik analisis sentimen, ekstraksi informasi, dan klasifikasi teks, maka model akan secara otomatis dapat mengetahui konten moderasi beragama yang tersebar di media sosial.

Text mining adalah pendekatan metode analisis yang menggunakan teknik komputasi untuk melakukan ekstraksi dan identifikasi informasi teks. *Teks mining* dapat melakukan analisis dan evaluasi sentimen yang terkandung dalam informasi teks secara otomatis. Pemodelan *machine learning* yang terintegrasi di dalam informasi *text mining* dapat menjadikan klasifikasi teks secara otomatis ke dalam kategori sentimen [8].

Klasifikasi merupakan metode yang lazim ditemukan dalam data mining. Klasifikasi dapat diterapkan ke dalam beberapa objek data, diantaranya dapat digunakan untuk melakukan klasifikasi dokumen teks. Klasifikasi sebenarnya dapat dilakukan secara manual oleh manusia menggunakan aturan klasifikasi, tetapi jika data yang akan diklasifikasi jumlah besar maka akan kerepotan. Saat ini, penerapan *machine learning* dapat membantu proses klasifikasi dokumen secara otomatis secara lebih efektif dan terstruktur. Klasifikasi terdiri dari dua proses utama yaitu membangun model dari data latih dan menggunakan model tadi untuk melakukan klasifikasi data tes kemudian dilakukan pengukuran model tersebut [9].

Proses *text mining* melibatkan ekstraksi informasi dari kumpulan teks yang besar untuk menemukan pola dan wawasan yang berguna. Dengan semakin banyaknya pengguna internet dan media sosial, seperti Twitter, kebutuhan untuk menganalisis data tekstual menjadi semakin mendesak [10]. *Text mining* dapat digunakan untuk berbagai aplikasi, termasuk analisis sentimen, pengelompokan dokumen, dan klasifikasi teks, yang semuanya bertujuan untuk mengolah data yang tidak terstruktur menjadi informasi yang dapat lebih berguna [11]. Dalam proses *text mining*, terdapat beberapa tahapan penting yang harus dilalui, termasuk *preprocessing*, pengolahan data, dan analisis hasil. *Preprocessing* adalah proses yang terdiri dari langkah-langkah seperti penghilangan kata-kata atau *term* yang tidak penting dan tidak mempunyai makna (*stopword removal*), tokenisasi, dan *stemming*, yang semuanya mempunyai tujuan untuk mempersiapkan dan mengolah data, sehingga data tersebut siap untuk dianalisis [12] [13]. Setelah tahap ini, algoritma seperti *Naive Bayes*, *Support Vector Machine (SVM)*, dan *K-Means* sering digunakan untuk mengklasifikasikan dan mengelompokkan data berdasarkan pola yang ditemukan dalam teks [14] [15].

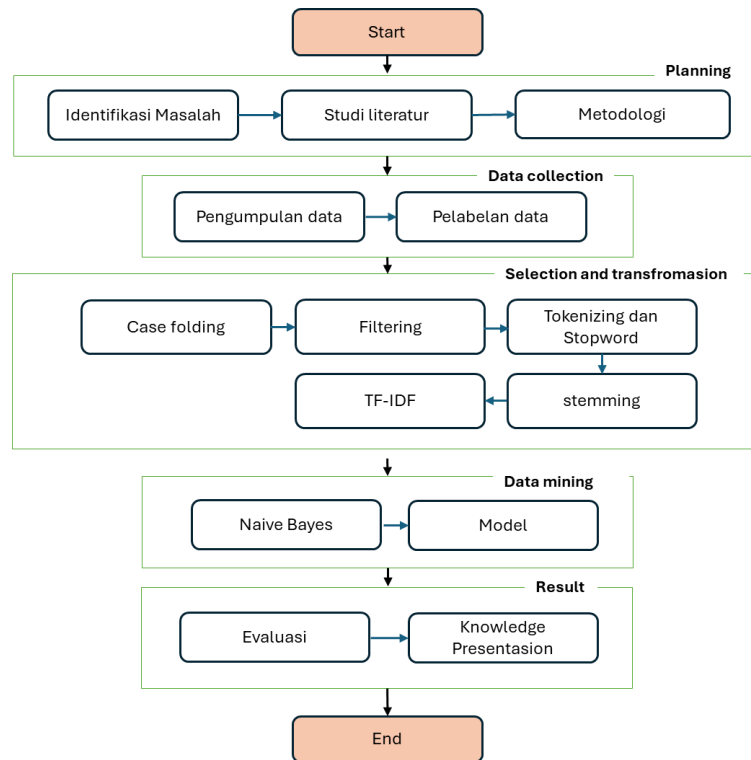
Salah satu metode efektif dalam klasifikasi teks adalah Algoritma *Naive Bayes Multinomial* [16] [17] [18]. Algoritma ini memanfaatkan model probabilitas untuk mengklasifikasikan teks berdasarkan distribusi kata-kata dalam dokumen. Dalam konteks *text mining*, *Naive Bayes* sering digunakan untuk analisis sentimen, pengelompokan dokumen, dan pengenalan pola dalam teks [19] [20] [21]. Keunggulan utama dari algoritma *Naive Bayes* adalah kemampuannya untuk menangani data yang tidak terstruktur dengan baik, seperti teks. Algoritma ini memiliki langkah atau metode untuk menghitung probabilitas setiap kata dalam konteks kelas tertentu, dan kemudian menggunakan probabilitas tersebut untuk menentukan kelas dari dokumen baru. Misalnya, dalam analisis sentimen, *Naive Bayes* dapat digunakan untuk mengklasifikasikan *tweet* atau cuitan sebagai positif dan negatif berdasarkan kata-kata yang muncul dalam *tweet* tersebut [20] [21] [22].

Penelitian ini bertujuan untuk membuat model klasifikasi otomatis pada konten moderasi beragama pada media sosial dengan mengimplementasikan Algoritma *Naive Bayes Multinomial*. Dengan adanya model ini, sistem dapat mendeteksi secara efektif dan meminimalkan dampak negatif konten moderasi beragama dan memberikan kontribusi positif terhadap menciptakan lingkungan daring yang lebih aman dan harmonis.

Melalui pendekatan ini, kita dapat mengetahui dinamika narasi konten beragama di media sosial, dan mengidentifikasi pola-pola perilaku yang tidak moderat di media sosial serta mengetahui strategi moderasi yang lebih cerdas untuk meningkatkan kualitas interaksi di dunia maya. Selain itu, penelitian ini jika dapat memberikan hasil kontribusi pada literatur tentang analisis teks, klasifikasi menggunakan model *text mining* pada konten narasi moderasi beragama yang berada di media sosial.

2. METODE PENELITIAN

Langkah dalam penelitian ini berdasarkan pada metode *Knowledge Discovery Data from Data (KDD)* [23]. Urutan proses penelitian ini dapat di ilustrasikan seperti pada gambar 1 di bawah ini.



Gambar 1. Alur Penelitian

2.1 Metode Pengumpulan Data (Data Collection)

Dalam penelitian ini, *datasets* dikumpulkan dari platform media sosial *Twitter* (sekarang dikenal sebagai *X*). Untuk tujuan analisis klasifikasi *teks*, metode pengumpulan data yang paling umum dan efektif adalah teknik *web scraping* dan *crawling*. Kedua teknik ini memungkinkan peneliti untuk memperoleh sejumlah besar data dari platform *Twitter (X)* secara otomatis.

Melalui *web scraping* dan *crawling*, peneliti dapat mengakses dan mengumpulkan data *tweet* dengan efisiensi tinggi. Data yang dikumpulkan mencakup berbagai informasi penting seperti *teks tweet*, tanggal publikasi, dan meta data relevan lainnya. Penting untuk dicatat bahwa proses pengumpulan data ini harus memperhatikan rentang waktu data *tweet* guna memastikan relevansi dan konsistensi *dataset* dengan tujuan penelitian.

Metode ini dapat memungkinkan peneliti untuk mengumpulkan dan mengolah data dalam jumlah besar, yang sangat penting untuk analisis yang akurat dan representatif. Dalam mengambil data dari *twitter* akan dipertimbangkan beberapa parameter diantaranya yang pertama terkait dengan rentang waktu pengambilan. Yang kedua adalah terkait dengan *keyword*. Pengambilan kata kunci yang tepat akan menghasilkan *datasets* yang akurat, sehingga akan mempengaruhi hasil dari sebuah klasifikasi penelitian.

Setelah pengumpulan data, kemudian dilanjutkan dengan proses pelabelan data. Proses pelabelan dilakukan secara manual oleh ahli. Ahli akan membaca *tweet* yang didapat kemudian melakukan pelabelan ke dalam dua jenis label, yaitu label radikal dan moderat.

2.2 Selection and Transformation

Di dalam *Knowledge Discovery from Data (KDD)*, tahap pemilihan dan transformasi data merupakan dua langkah krusial yang berkontribusi pada keberhasilan proses penemuan pengetahuan. Pemilihan data (*selection*) adalah langkah awal yang melibatkan pengidentifikasian dan pengambilan *subset* data yang relevan dari kumpulan data yang lebih besar. Hal ini dilakukan untuk menjamin bahwa hanya data yang berkualitas dan relevan yang digunakan dalam analisis lebih lanjut. Pemilihan data yang tepat dapat meningkatkan efisiensi dan efektivitas proses KDD secara keseluruhan. Selain itu, pemilihan data yang baik juga membantu mengurangi kompleksitas dan waktu pemrosesan yang diperlukan dalam tahap-tahap berikutnya.

Dalam proses ini akan dilakukan beberapa tahapan *selection* dan *transformation* diantaranya dengan melakukan *Case folding*, *Filtering*, *Tokenizing dan Stopword*, *Stemming* selanjutnya dengan *TF-IDF* atau *Term Frequency-Inverse Document Frequency*.

2.2.1 Case Folding

Tahap *case folding* akan dilakukan penyeragaman huruf, di mana huruf besar atau capital akan diubah ke huruf kecil, kemudian juga akan dilakukan penghilangan tanda baca dan angka.

2.2.2 Filtering

Pada tahap *filtering* akan dilakukan pembersihan data dengan melakukan pengilangan tanda baca, *mention*, *username*, *link* atau *URL*, *hashtag* dan simbol atau karakter.

2.2.3 Tokenizing dan stop word removal

Proses tokenisasi merupakan langkah awal yang sangat penting. Tokenisasi merupakan proses memecah teks menjadi bagian atau unit yang lebih kecil, yang kemudian disebut sebagai *token*. *Token* ini bisa berupa kata, frasa, atau simbol, tergantung pada tujuan analisis. Proses ini memungkinkan analisis lebih lanjut terhadap data teks yang tidak terstruktur, seperti yang sering ditemukan dalam media sosial, artikel, cuitan *twitter* atau dokumen lainnya. *Stop word removal* merupakan proses penghilangan kata yang tidak mempunyai makna atau tidak mempunyai arti yang signifikan. Kata tersebut bisa berupa kata sambung atau kata penunjuk tempat seperti dan, di, dan lain sebagainya.

Secara keseluruhan, tokenisasi adalah langkah fundamental dalam proses penambangan teks yang memungkinkan pengolahan dan analisis data teks yang tidak terstruktur. Proses ini tidak hanya memecah *text* menjadi lebih kecil, tetapi juga mempersiapkan data untuk analisis lebih lanjut yang dapat memberikan wawasan berharga dari data yang ada.

2.2.4 Stemming

Stemming merupakan salah satu teknik penting dalam proses penambangan teks (*text mining*) yang bertujuan untuk mengurangi kata-kata atau *term* ke bentuk kata dasar (*root form*). *Stemming* berfungsi untuk menyederhanakan analisis teks dengan mengeliminasi variasi morfologis dari kata-kata, sehingga memudahkan pengelompokan dan klasifikasi data teks. Proses ini sangat relevan dalam berbagai aplikasi, termasuk analisis sentimen, klasifikasi dokumen, dan ekstraksi informasi. Dalam melakukan *stemming* akan menggunakan *library sastrawi*.

2.2.5 TF/IDF

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang umum digunakan dalam pengolahan informasi dan pengambilan data untuk menilai relevansi suatu kata dalam sebuah dokumen dalam koleksi dokumen. Metode ini mengukur pentingnya suatu kata berdasarkan frekuensi kemunculannya dalam dokumen tertentu *Term Frequency (TF)* dan seberapa jarang kata tersebut muncul di seluruh koleksi dokumen *Inverse Document Frequency (IDF)*. Semakin suatu kata muncul dalam dokumen tertentu dan semakin jarang kata tersebut muncul di dokumen lain, semakin tinggi bobot yang diberikan kepada kata tersebut. Hal ini membuat TF-IDF sangat efektif dalam meningkatkan akurasi klasifikasi teks dan pengambilan informasi.

2.3 Data Mining

Di dalam tahap ini akan dilakukan proses pengklasifikasian data menggunakan algoritma data mining. *Dataset* akan dibagi menjadi 2 bagian, yaitu data latih (*training data*) dan data testing (*testing data*). *Data latih* akan membentuk model. Model tersebut kemudian diuji dengan menggunakan data testing.

Kemudian akan dilakukan klasifikasi dengan menggunakan algoritma *Naive Bayes Multinomial*. Dalam konteks klasifikasi dokumen, *Naive Bayes Multinomial* sangat cocok untuk menangani data teks yang memiliki distribusi kata yang berbeda, seperti dalam analisis sentimen dan deteksi ujaran kebencian di media sosial. *Tool* yang digunakan menggunakan *Python* dengan memanfaatkan salah satu *library scikit-learn*.

Metode *Naive Bayes* adalah sekumpulan supervised learning algorithm yang berdasarkan teorema bayes. Rumus dasar *Naive Bayes* dapat dinotasikan seperti berikut :

$$P(C|X) = \frac{P(X|C).P(C)}{P(X)} \quad 1)$$

Dimana

$P(C|X)$: probabilitas kelas C berdasarkan data X (*posterior*)

$P(X|C)$: *likelihood*, probabilitas data X jika kelas C diketahui.

$P(C)$: *Prior*, probabilitas awal kelas C .

$P(X)$: probabilitas data X secara keseluruhan.

2.4 Hasil

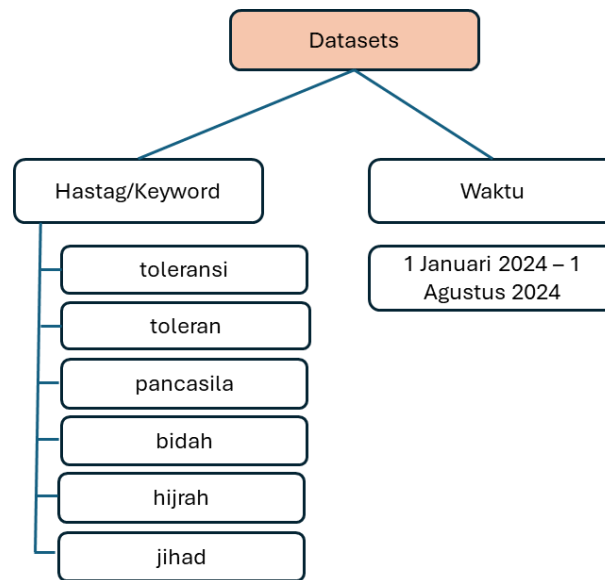
Hasil dari penelitian ini akan terlihat kemampuan dalam membangun model yang akurat dalam melakukan klasifikasi dokumen. Hal tersebut diukur menggunakan *confusion matrix* yang menghasilkan nilai *Accuracy model*. Selanjutnya itu juga yang dihasilkan dalam penelitian ini adalah bagaimana mempresentasikan hasil dari model klasifikasi dalam bentuk grafik sehingga mudah di pahami oleh pembaca. Penyajian grafik disajikan dalam bentuk *wordcloud*.

3. HASIL DAN PEMBAHASAN

Dalam bagian ini dibahas secara mendalam setiap langkah-langkah yang dilakukan di dalam penelitian berdasarkan dari metode penelitian yang telah di jabarkan sebelumnya

3.1 Data Collection

Pengumpulan data dilakukan dari pengambilan data dari *twitter* atau *x*. *Keyword* yang digunakan dalam penelitian ini adalah menggunakan kata *toleransi*, *toleran*, *pancasila*, *bidah*, *hijrah*, *moderat*. Dengan pembatasan limit setiap *crawling* 300 data. Hal tersebut dilakukan karena *twitter* melakukan pembatasan setiap kali melakukan *crawling data*. *Crawling data* dengan *hashtag* dapat terlihat pada gambar 2.



Gambar 2. Hashtag atau keyword dataset

Dalam pengumpulan data ini akan menggunakan *library Tweet Harvest*. Rentang waktu pengambilan data yaitu antara tanggal 1 Januari 2024 sampai dengan 01 Agustus 2024. Dengan *keyword* sebagai berikut *toleransi*, *toleran*, *pancasila*, *bidah*, *hijrah*, *moderat*. *Script python* untuk mengambil *keyword* ditunjukkan pada gambar 3.

```

1  twitter_auth_token =
   '3b13 [REDACTED]'
2  filename = 'toleran.csv'
3  search_keyword = 'toleran since:2023-04-01 until:2024-08-01
   lang:id'
4  limit = 300
5
6  !npx -y tweet-harvest@2.6.1 -o "{filename}" -s "
   {search_keyword}" --tab "LATEST" -l {limit} --token
   {twitter_auth_token}
7
    
```

Gambar 3. Script Scraping Tweet Harvers

Hasil dari *crawling data* di dapat hasil berikut, yang terdiri dari atribut *conversation_id_str*, *create_at*, *favorite_count*, *full_text*, *id_str*, *img_url*, *in_reply_to_screename*, *lang*, *location*, *quote_count*, *reply_count*, *retwet_count*, *tweet_url*, *user_id_str*, dan *username*.

conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_count	tweet_url	user_id_str	username	
0	1818624086484172805	Wed Jul 31 12:45:53 +0000 2024	0	@hyunjin maaf yah gaada toleransi kk	1818629154042266058	NaN	hyunjin	in	Brook Town	0	1	0	https://x.com/brocalypse/status/18186291540422...	123267690987757058	brocalypse
1	18186335919169116552	Wed Jul 31 12:40:09 +0000 2024	0	@distracthythm @Apokrif23 Aku2 istamopobhla	1818627709888872555	NaN	distracthythm	in	Everywhere	0	1	0	https://x.com/hadesianah/status/18186277098888...	1309439378	hadesianah
2	1818566525156261985	Wed Jul 31 12:19:08 +0000 2024	0	@lwasturgo Baunya bukan wangi (seperti bunga)...	1818622423559991468	NaN	lwasturgo	in	Indonesia	0	0	0	https://x.com/mami_su37049/status/18186224235...	1742286629031845888	mami_su37049
3	1818599048854880427	Wed Jul 31 12:16:46 +0000 2024	0	@twify ngga ada toleransi	1818621826597363758	NaN	twify	in	tl	0	0	0	https://x.com/pevtrpn/status/1818621826597363758...	896322446285979649	pevtrpn
4	1818621528243925264	Wed Jul 31 12:15:35 +0000 2024	0	Fans kpop kalo udah bawa-bawa kematian seseorang...	1818621528243925264	NaN	NaN	in	NaN	0	0	0	https://x.com/teumenon/status/181862152824392...	1394002346326183940	teumenon
110	1818324812948996552	Tue Jul 30 16:36:32 +0000 2024	0	Gatau ya tingkat toleransi ku tinggi kekuk mau...	1818324812948996552	NaN	NaN	in	NaN	0	0	0	https://x.com/sebatianco/status/18183248129489...	1280456129730891776	sebatianco
111	1818300398157119883	Tue Jul 30 16:22:04 +0000 2024	0	@yumperees Masalahnya toleransi mereka bener...	1818321169193615524	NaN	lhaechiel	in	NaN	0	0	0	https://x.com/lhaechiel/status/1818321169193...	1664788063732772864	lhaechiel
112	1818310685153206299	Tue Jul 30 16:12:15 +0000 2024	0	@budibongg Toleransi tingkat tinggi	1818318702041014377	NaN	budibongg	in	Indramayu, Indonesia	0	0	0	https://x.com/malagrela/status/181831870204101...	1357069646	malagrela
113	1818318630494535850	Tue Jul 30 16:11:58 +0000 2024	0	Gada semua aku toleransi tinggi kan dulu aku...	1818318630494535850	NaN	NaN	in	ULTRAZ	0	0	0	https://x.com/agentsEQUIN/status/1818318630494...	1221524857487101956	agentsEQUIN
114	1818316405055574133	Tue Jul 30 16:03:08 +0000 2024	0	Banyakah sebenarnya aku illah sama orang d...	1818316405055574133	NaN	NaN	in	Athena Tpeason	0	0	0	https://x.com/athenajpoason/status/18183164050...	1496534852807847936	athenajpoason

Gambar 4. Hasil Scraping Tweet

Selanjutnya *datasets* akan dibagi menjadi dua bagian, yaitu data testing (*data testing*) dan data latih (*data training*). Semua *datasets* akan dilabeli oleh ahli. *Data testing* yaitu *datasets* yang digunakan untuk menguji model. Sedangkan data training yaitu *datasets* yang akan digunakan sebagai *datasets* yang membangun sebuah model. *Datasets tweet* akan dikumpulkan berdasarkan hastag atau keyword tertentu. Dari hasil pengumpulan *datasets* didapatkan hasil yang tampak pada tabel sesuai pada tabel 1.

Tabel 1. Hasil crawling berdasar keyword

No	Keyword	Tweet
1	Toleransi	356
2	Toleran	309
3	Pancasila	311
4	Bidah	330
5	Hijrah	342
6	jihad	348
Total		1.996

3.2 Pelabelan Data

Proses pelabelan data dilabeli secara manual, oleh ahli yang mampu membedakan cuitan moderat dan tidak moderat. Dikarenakan *datasets* yang didapatkan dari *twitter* atau *x* sangat beragam, maka dalam penelitian ini penulis membagi kategori cuitan *twitter* ke dalam beberapa klasifikasi. Masing-masing ahli akan diberikan *tweet* kemudian ahli tersebut akan melakukan klasifikasi menurut persepsi ahli tersebut. Selanjutnya setiap *tweet* akan diklasifikasi ke dalam klasifikasi moderat dan radikal. Moderat jika pakar menilai bahwa *tweet* tidak mengandung ajakan yang baik, sesuai dengan *framework* moderasi beragama kemenag. Sedangkan tidak moderat jika terdapat hal-hal yang bertentangan dengan *framework* moderasi beragama kemenag.

Tabel 2. Klasifikasi Dataset

No	Klasifikasi	Tweet
1	Moderat	814
2	Radikal	1182
TOTAL		1996

3.2 Selection dan Transformation

Langkah pertama dalam tahap ini yaitu melakukan load data dari *datasets* yang sudah di persiapkan pada tahap sebelumnya.

3.2.1 Load Data

Penulis menggunakan *library pandas* dan menggunakan bahasa pemrograman python. File *datasets* sebelumnya disimpan dalam *format xlsx* dengan nama file *datasets-mb1.xls*. File tersebut akan di buka ke dalam sistem dengan menggunakan *library pandas*.

```
import pandas as pd
import re

data = pd.read_excel('dataset-mb1.xlsx')
data.info()
```

Gambar 5. Script Load Data

Dari hasil pengolahan data didapat 1996 baris data yang memiliki 3 atribut yaitu *idc*, *text*, *label* dimana *idc* memiliki tipe data *int64*, *text* memiliki tipe data *object* yang berisi teks cuitan dan label memiliki tipe data *object* yang berisi klasifikasi data yang mempunyai 2 kategori yaitu radikal dan moderat. berikut contoh data ditampilkan maka akan terlihat seperti berikut. Terdiri dari 3 atribut yaitu *idc*, *text* dan *label*.

idc	text	label
0	1 Bekerja mencari nafkah jika diniatkan dengan benar makan sama dengan jihad.. Wajar aja sih sebagian orang kalo berangkat kerja berasa berangkat perang	moderat
1	2 Kami pun paham jika semua kasus dan eksistensi terorisme di Indonesia dan sekitar juga masih operasi intelijen Barat/Pihak ² Terkait. Bukan Murni Operasi Nasionalisme Islamis (Jihad Murni). Pihak Propaganda dan Pihak Alat Nya sama sama teroris	moderat
2	3 4 Jihad Sederhana Membantu Palestina: 1. Jihad Medsos 2. Jihad Ekonomi 3. Jihad Sosial 4. Jihad Hati https://t.co/qCX0iy3rTC	moderat
3	4 BARUSAN BGT LIAT ANAK XTC ROLLING DI CIMAH I EHHH SKRG ADA PEMBACOKAN DI BONKOP	radikal
4	5 @ernestprakasa Jihad yang paling utama adalah seseorang berjihad (berjuang) melawan dirinya dan hawa nafsunya (Hadits shahih diriwayatkan oleh Ibnu Najjar dari Abu Dzarr). Sumber: ² https://t.co/cBd6GgYavD Copyright ² © 2024 https://t.co/YIE2UilfGm	moderat
5	6 Yuk saksikan *Buka Bareng Kita* edisi Ahad 31 Maret 2024 pukul 16.45 WIB bersama *Ust. M. Karebet Widjajakusuma*. https://t.co/uVcj4ChSxR Jangan lupa untuk share informasi ini . #kisah #konstantinopel #jihad #RamadhanBarengKita #BukaBarengKita #BerkahDenganSyariahKaffah https://t.co/l0viDA8YaB	moderat

Gambar 6. Tampilan data dengan 3 atribut

3.2.2 Case Folding

Tahap *case folding* dilakukan penyeragaman atau normalisasi huruf, seperti perubahan huruf besar akan diubah ke dalam huruf kecil, penghapusan *mention* (@username), penghapusan URL, penghapusan *hashtag*, penghapusan tanda baca, penghapusan karakter non-alfanumerik, penghapusan angka 0-9, normalisasi kata seperti “yg” menjadi “yang”, “tp” menjadi “tapi”, bgt menjadi begitu, liat menjadi lihat, “gak” menjadi “tidak”, “org” menjadi “orang”, “rt” menjadi “kutip”.

	text	text_clean
0	Bekerja mencari nafkah jika diniatkan dengan benar makan sama dengan jihad.. Wajar aja sih sebagian orang kalo berangkat kerja berasa berangkat perang	bekerja mencari nafkah jika diniatkan dentidakn benar makan sama dentidakn jihad wajar aja sih sebagian orang kalo berangkat kerja berasa berangkat perang
1	Kami pun paham jika semua kasus dan eksistensi terorisme di Indonesia dan sekitar juga masih operasi intelijen Barat/Pihak ² Terkait. Bukan Murni Operasi Nasionalisme Islamis (Jihad Murni). Pihak Propaganda dan Pihak Alat Nya sama sama teroris	kami pun paham jika semua kasus dan eksistensi terorisme di indonesia dan sekitar jutidak masih operasi intelijen baratapiihak terkait bukan murni operasi nasionalisme islamis jihad murni pihak propatidaknda dan pihak alat nya sama sama teroris
2	4 Jihad Sederhana Membantu Palestina: 1. Jihad Medsos 2. Jihad Ekonomi 3. Jihad Sosial 4. Jihad Hati https://t.co/qCX0iy3rTC	jihad sederhana membantu palestina jihad medsos jihad ekonomi jihad sosial jihad hati

Gambar 7. Hasil Case Folding

3.2.3 Tokenizing

Untuk melakukan tokenizing, dalam penelitian ini menggunakan library NLTK. Oleh karena ini perlu melakukan import NLTK terlebih dahulu. Library NLTK banyak dipakai di dalam penelitian sentimen analisis. Selanjutnya ditampilkan hasil `text_clean` yang telah dilakukan proses tokenizing, terlihat pada atribut token.

	text_clean	token
0	bekerja mencari nafkah jika diniatkan dentidakn benar makan sama dentidakn jihad wajar aja sih sebagian orang kalo berangkat kerja berasa berangkat perang	[bekerja, mencari, nafkah, jika, diniatkan, dentidakn, benar, makan, sama, dentidakn, jihad, wajar, aja, sih, sebagian, orang, kalo, berangkat, kerja, berasa, berangkat, perang]
1	kami pun paham jika semua kasus dan eksistensi terorisme di indonesia dan sekitar jutidak masih operasi intelijen baratapiihak terkait bukan murni operasi nasionalisme islamisl jihad murni pihak propatidaknda dan pihak alat nya sama sama teroris	[kami, pun, paham, jika, semua, kasus, dan, eksistensi, terorisme, di, indonesia, dan, sekitar, jutidak, masih, operasi, intelijen, baratapiihak, terkait, bukan, murni, operasi, nasionalisme, islamisl, jihad, murni, pihak, propatidaknda, dan, pihak, alat, nya, sama, sama, teroris]

Gambar 8. Hasil Tokenizing

3.2.4 Stopword

Stop word removal merupakan proses penghilangan kata yang tidak bermakna. Dalam hal ini menggunakan *stopword* yang dikembangkan oleh library sastrawi. *Stopword* tersebut diantaranya adalah sebagai berikut :

[*'yang', 'untuk', 'pada', 'ke', 'para', 'namun', 'menurut', 'antara', 'dia', 'dua', 'ia', 'seperti', 'jika', 'jika', 'sehingga', 'kembali', 'dan', 'tidak', 'ini', 'karena', 'kepada', 'oleh', 'saat', 'harus', 'sementara', 'setelah', 'belum', 'kami', 'sekitar', 'bagi', 'serta', 'di', 'dari', 'telah', 'sebagai', 'masih', 'hal', 'ketika', 'adalah', 'itu', 'dalam', 'bisa', 'bahwa', 'atau', 'hanya', 'kita', 'dengan', 'akan', 'juga', 'ada', 'mereka', 'sudah', 'saya', 'terhadap', 'secara', 'agar', 'lain', 'anda', 'begitu', 'mengapa', 'kenapa', 'yaitu', 'yakni', 'daripada', 'itulah', 'lagi', 'maka', 'tentang', 'demi', 'dimana', 'kemana', 'pula', 'sambil', 'sebelum', 'sesudah', 'supaya', 'guna', 'kah', 'pun', 'sampai', 'sedangkan', 'selagi', 'sementara', 'tetapi', 'apakah', 'kecuali', 'sebab', 'selain', 'seolah', 'seraya', 'seterusnya', 'tanpa', 'agak', 'boleh', 'dapat', 'dsb', 'dst', 'dll', 'dahulu', 'dulunya', 'amu', 'demikian', 'tapi', 'ingin', 'juga', 'nggak', 'mari', 'nanti', 'melainkan', 'oh', 'ok', 'seharusnya', 'sebetulnya', 'setiap', 'setidaknya', 'sesuatu', 'pasti', 'saja', 'toh', 'ya', 'walau', 'tolong', 'tentu', 'amat', 'apalagi', 'bagaimanapun'*]

Hasil dari skrip tersebut seperti atribut berikut, *token* merupakan hasil sebelum di lakukan *stop word removal*, sedangkan atribut *stop* merupakan hasil setelah di lakukan *stopword removal*.

	token	stop
0	[bekerja, mencari, nafkah, jika, diniatkan, dentidakn, benar, makan, sama, dentidakn, jihad, wajar, aja, sih, sebagian, orang, kalo, berangkat, kerja, berasa, berangkat, perang]	[bekerja, mencari, nafkah, diniatkan, dentidakn, benar, makan, sama, dentidakn, jihad, wajar, aja, sih, sebagian, orang, kalo, berangkat, kerja, berasa, berangkat, perang]
1	[kami, pun, paham, jika, semua, kasus, dan, eksistensi, terorisme, di, indonesia, dan, sekitar, jutidak, masih, operasi, intelijen, baratapiihak, terkait, bukan, murni, operasi, nasionalisme, islamisl, jihad, murni, pihak, propatidaknda, dan, pihak, alat, nya, sama, sama, teroris]	[paham, semua, kasus, eksistensi, terorisme, indonesia, jutidak, operasi, intelijen, baratapiihak, terkait, bukan, murni, operasi, nasionalisme, islamisl, jihad, murni, pihak, propatidaknda, pihak, alat, nya, sama, sama, teroris]

Gambar 9. Hasil stopword removal

3.2.5 Stemming

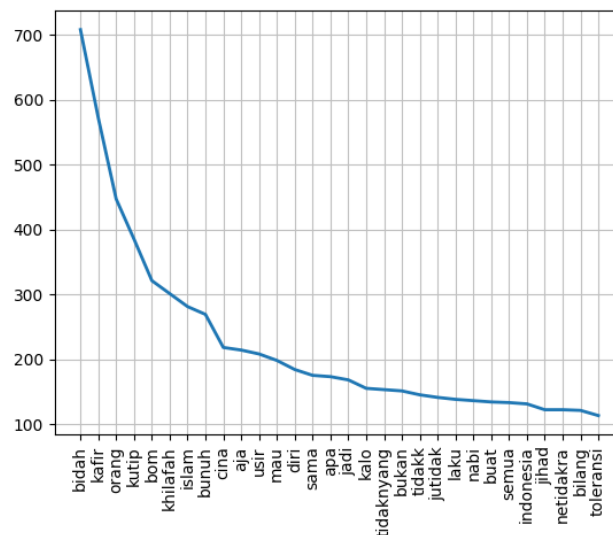
Stemming text adalah proses mencari kata dasar. *Stemming* dalam bahasa Indonesia berarti mencari kata dasar dengan menghilangkan imbuhan atau awalan suatu kata. Dalam penerapnya dalam *script python stemming* menggunakan *library sastrawi*. Atribut *stop* merupakan teks yang belum di lakukan *stemming*, sedangkan atribut *stemming* merupakan teks yang sudah dilakukan *stemming*. Proses *stemming* memerlukan waktu yang cukup lama. Dari *datasets* akan terlihat beberapa kata berubah menjadi kata dasar. Kata sebelum *stemming* berada pada kolom *stop*, sedangkan kata yang sudah di *stemming* berada pada kolom *stemed*. Seperti kata “bekerja” menjadi “kerja”, kata “diajak” menjadi “ajak”, kata “saksikan” menjadi “saksi”, dan lain sebagainya.

	stop	stemmed
0	[bekerja, mencari, nafkah, diniatkan, dentidakn, benar, makan, sama, dentidakn, jihad, wajar, aja, sih, sebagian, orang, kalo, berangkat, kerja, berasa, berangkat, perang]	[kerja, cari, nafkah, niat, dentidakn, benar, makan, sama, dentidakn, jihad, wajar, aja, sih, bagi, orang, kalo, berangkat, kerja, asa, berangkat, perang]
1	[paham, semua, kasus, eksistensi, terorisme, indonesia, jutidak, operasi, intelijen, baratapiihak, terkait, bukan, murni, operasi, nasionalisme, islamis, jihad, murni, pihak, propatidaknda, pihak, alat, nya, sama, sama, teroris]	[paham, semua, kasus, eksistensi, terorisme, indonesia, jutidak, operasi, intelijen, baratapiihak, kait, bukan, murni, operasi, nasionalisme, islamis, jihad, murni, pihak, propatidaknda, pihak, alat, nya, sama, sama, teroris]
2	[jihad, sederhana, membantu, palestina, jihad, medsos, jihad, ekonomi, jihad, sosial, jihad, hati]	[jihad, sederhana, bantu, palestina, jihad, medsos, jihad, ekonomi, jihad, sosial, jihad, hati]

Gambar 10. Hasil Stemming

3.2.6 TF-IDF

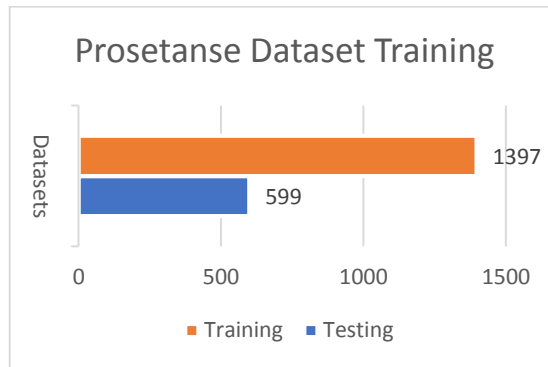
Sebelum dilakukan TF/IDF (*Term Frequency-Inverse Document Frequency*) akan dipetakan dalam grafik kemunculan kata terbanyak. Hasilnya terlihat pada grafik berikut.



Gambar 11. Sebaran Kata

3.3 Data Mining

Sebelum melakukan pengolahan menggunakan *algoritma machine learning*, maka perlu dilakukan pembagian *datasets* terlebih dahulu. *Datasets* dibagi atau di split menjadi 2 bagian yaitu yang pertama disebut sebagai *data training* dan yang kedua disebut sebagai *data testing*. fungsi *train_test_split* yang diambil dari pustaka *scikit-learn* digunakan untuk membagi *datasets* menjadi dua bagian, yaitu set pelatihan dan set pengujian. Fitur yang digunakan (*df['st']*) merupakan data teks, sedangkan target atau label (*df['label']*) adalah kelas yang akan diprediksi oleh model. Parameter *test_size=0.3* menunjukkan bahwa 30% dari data akan dialokasikan untuk pengujian, sementara 70% sisanya untuk pelatihan. Dengan demikian, jika ada 1.996 sampel dalam *datasets*, 1.397 sampel akan digunakan untuk pelatihan dan 599 sampel untuk pengujian. Prosentase sebaran *datasets* training dapat dilihat seperti pada gambar 12 dibawah ini



Gambar 12. Prosentase Datasets Training

Selain itu, penggunaan *stratify=df['label']* memastikan bahwa distribusi kelas dalam set pelatihan dan pengujian tetap proporsional, menjaga keseimbangan data di kedua set. Hal ini penting untuk menghindari bias dalam pembagian data, terutama dalam kasus di mana distribusi kelas tidak merata. Parameter *random_state=42* memberikan hasil yang konsisten dengan mengatur *seed* acak, sehingga pembagian data dapat direproduksi dengan hasil yang sama setiap kali skrip dijalankan.

Selanjutnya klasifikasi dengan *Naive Bayes Multinomial*. Dalam penelitian ini, untuk membantu penelitian, peneliti memanfaatkan *library sklearn* yang telah menyediakan fitur model *Naive Bayes Multinomial*. Yang pertama dilakukan adalah melakukan import library tersebut. Dimulai dari *Naive Bayes Multinomial*. Kemudian masing-masing model akan di inputkan dengan *datasets* training yang sudah disiapkan sebelumnya. Dalam hal ini *datasets* yang dipakai dengan prosentase 30 % data testing, dan 70% data training. *X_train* merepresentasikan atribut atau feature *datasets*, sedangkan *y_train* merupakan target atau label *datasets*.

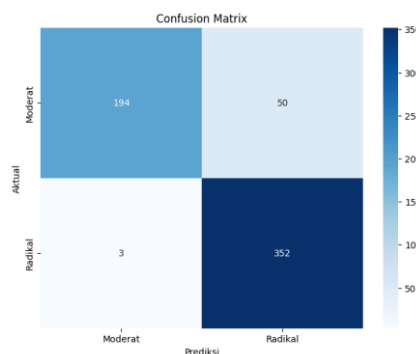
```
from sklearn.naive_bayes import MultinomialNB

modelNb = MultinomialNB()
modelNb.fit(X_train, y_train)
```

Gambar 13. Script Algoritma Naive Bayes Multinomial

3.4 Hasil / Evaluasi Model

Model akan dievaluasi dengan menggunakan metode *confusion matrix*. Dalam *confusion matrix* akan dicari nilai *accuracy*, *precision*, *recall*, *F1-scores*. Dalam script python Performa algoritma *Naive Bayes Multinomial* secara keseluruhan menunjukkan ketangguhan dalam mengklasifikasikan kelas "Radikal", yang dibuktikan dengan jumlah *False Negative* yang sangat rendah. Namun, masih ada tantangan dalam mengurangi jumlah *False Positive*, yang terlihat dari banyaknya data "Moderat" yang salah diklasifikasikan sebagai "Radikal". Hasil *confusion matrix* algoritma *Naive Bayes Multinomial* terlihat pada gambar 14.



Gambar 14. Confusion Matrix Algoritma Naive Bayes Multinomial Teks

Berdasarkan matriks tersebut, dari total 194 tweet yang secara aktual termasuk kategori 'Moderat', model berhasil mengklasifikasikannya dengan benar. Sementara itu, 50 tweet yang sebenarnya 'Moderat' keliru diklasifikasikan sebagai 'Radikal' (*False Positives*). Di sisi lain, untuk kategori 'Radikal', model mengidentifikasi 352 *tweet* dengan benar sebagai 'Radikal' (*True Negatives*), sedangkan hanya 3 *tweet* yang secara aktual 'Radikal' salah diklasifikasikan sebagai 'Moderat' (*False Negatives*). Hasil ini mengindikasikan bahwa model dengan algoritman *Naïve Bayes Multinomial* memiliki kemampuan yang baik di dalam melakukan identifikasi konten cuitan 'Radikal' dengan tingkat kesalahan yang rendah, meskipun masih terdapat ruang untuk peningkatan dalam membedakan konten 'Moderat' agar tidak salah diklasifikasi sebagai 'Radikal'.

```

Accuracy Multinomial : 0.911519

Classification Report:
              precision    recall  f1-score   support

   moderat      0.98      0.80      0.88       244
   radikal      0.88      0.99      0.93       355

 accuracy
macro avg      0.93      0.89      0.90       599
weighted avg   0.92      0.91      0.91       599

Confusion Matrix:
[[194  50]
 [   3 352]]
    
```

Gambar 15. Precision, Recall, Accuracy dan F1-Score

Model menunjukkan performa keseluruhan yang kuat dengan akurasi 0.91 (atau 91.15%). Lebih rinci, untuk kelas 'moderat', model mencapai *precision* 0.98 dan *recall* 0.80, menunjukkan kemampuannya yang sangat baik dalam memastikan kebenaran prediksi 'moderat' (tingkat *false positive* yang rendah) meskipun ada beberapa *tweet* moderat yang tidak terdeteksi. Sebaliknya, pada kelas 'radikal', model menunjukkan *precision* 0.88 dan *recall* yang sangat tinggi yaitu 0.99, mengindikasikan kemampuannya untuk mengidentifikasi hampir semua konten 'radikal' yang sebenarnya (tingkat *false negative* yang sangat rendah), meskipun dengan sedikit lebih banyak *false positive* dibandingkan kelas moderat. Keseluruhan nilai *F1-score* untuk kelas 'moderat' adalah 0.88 dan 'radikal' adalah 0.93, sementara *weighted average* untuk *precision*, *recall*, dan *F1-score* masing-masing adalah 0.92, 0.91, dan 0.91, menegaskan konsistensi kinerja model yang baik pada kedua kategori dengan mempertimbangkan jumlah sampel (*support*) masing-masing kelas (244 untuk moderat dan 355 untuk radikal).

4. KESIMPULAN

Penggalian *datasets* moderasi beragama yang terdapat dalam media sosial menjadi aspek yang sangat penting. Menentukan keyword menjadi aspek penentu dalam memperoleh *datasets* cuitan moderasi beragama. Setelah *datasets* didapatkan maka diperlukan validasi sehingga *datasets* yang didapat merupakan *datasets* valid. Dalam penelitian ini telah berhasil membuat sebuah sistem atau model yang dapat melakukan pelabelan *datasets* menggunakan validator. Dengan sistem tersebut *datasets* dapat terlabeli secara valid. Selanjutnya *datasets* tersebut dilakukan analisis dengan menggunakan algoritma *Naive Bayes*.

Berdasarkan evaluasi kinerja model *Naïve Bayes Multinomial* dalam klasifikasi konten moderasi beragama pada media sosial, dapat disimpulkan bahwa model ini menunjukkan performa yang kuat dan konsisten. Metode evaluasi menggunakan *confusion_matrix*, *accuracy*, *precision*, *recall*, dan *F1-score* secara komprehensif menggambarkan kapabilitas model. Secara keseluruhan, model mencapai akurasi sebesar 0.91 (91.15%), menandakan tingkat kebenaran prediksi yang tinggi. Meskipun model sangat tangguh dalam mengklasifikasikan kelas "Radikal" dengan *recall* mencapai 0.99 dan *False Negative* yang sangat rendah (hanya 3), masih terdapat tantangan dalam mengurangi *False Positive*, di mana 50 *tweet* "Moderat" salah diklasifikasikan sebagai "Radikal". Hal ini terefleksi dari *recall* kelas "Moderat" sebesar 0.80, meskipun *precision* untuk kelas "Moderat" sangat tinggi (0.98). Kendati demikian, *F1-score* yang tinggi untuk kedua kelas (0.88 untuk "Moderat" dan 0.93 untuk "Radikal"), serta nilai *weighted average precision*, *recall*, dan *F1-score* yang konsisten di angka 0.91-0.92, menegaskan bahwa algoritma *Naive Bayes Multinomial* efektif dan dapat diandalkan untuk tugas klasifikasi ini, terutama dalam mengidentifikasi konten yang bersifat radikal.

UCAPAN TERIMAKASIH

Penulis mengucapkan terimakasih kepada Lembaga Penelitian Pengabdian Masyarakat (LP2M) Universitas Islam Negeri Walisongo yang telah mendanai dengan anggaran BOPTN UIN Walisongo Semarang Tahun 2024 dengan kontrak dengan SK Nomor 877/Un.10.0/R/HK.01.14/4/2024 dan nomor kontrak 372/Un.10.0/L.1/TA.00.01/04/2024

REFERENCE

- [1] S. Hamdi, M. Munawarah, and H. Hamidah, "Revitalisasi Syiar Moderasi Beragama di Media Sosial: Gaungkan Konten Moderasi untuk Membangun Harmonisasi," *intizar*, vol. 27, no. 1, pp. 1–15, Jul. 2021, doi: 10.19109/intizar.v27i1.8191.
- [2] R. W. Wibowo and A. S. Nurjanah, "Aktualisasi Moderasi Beragama Abad 21 Melalui Media Sosial".
- [3] A. S. Anwar, K. Leo, U. Ruswandi, and M. Erihadiana, "Internalisasi Nilai-Nilai Moderasi Beragama Abad 21 melalui Media Sosial," *jiiip*, vol. 5, no. 8, pp. 3044–3052, Aug. 2022, doi: 10.54371/jiiip.v5i8.795.
- [4] C. S. D. Kusuma, "DAMPAK MEDIA SOSIAL DALAM GAYA HIDUP SOSIAL (STUDI KASUS PADA MAHASISWI PENDIDIKAN ADMINISTRASI PERKANTORAN FE UNY)," *EFISIENSI*, vol. 17, no. 1, pp. 15–33, Feb. 2020, doi: 10.21831/efisiensi.v17i1.30436.
- [5] A. H. Jakiyudin, "Urgensi Literasi Moderasi Beragama Dalam Ruang Bermedia Sosial Perspektif Al Qur'an," *JP*, vol. 17, no. 1, p. 105, Sep. 2023, doi: 10.21043/jp.v17i1.17148.
- [6] Pusat Bahasa and Indonesia (Departemen Pendidikan Nasional), *Kamus besar bahasa Indonesia : edisi ketiga*, Ed. 3, cet. 3. [Jakarta]: Pusat Bahasa : Departemen Pendidikan Nasional : Balai Pustaka [Jakarta], 2005.
- [7] E. Kosasih, "Literasi Media Sosial Dalam Pemasyarakatan Sikap Moderasi Beragama," *Jurnal Bimas Islam*, vol. 12, no. 2, pp. 263–296, 2019, doi: 10.37302/jbi.v12i2.118.
- [8] C. Manning, P. Raghavan, and H. Schuetze, "Introduction to Information Retrieval," 2009.
- [9] N. Lestari, O. S. Riza, and R. Ardinal, "Implementation Of Text mining And Pattern Discovery With Naive Bayes Algorithm For Classification Of Text Documents".
- [10] A. Hermawan, "Implementasi Text-Mining Untuk Analisis Sentimen Pada Twitter Dengan Algoritma Support Vector Machine," *JST (Jurnal Sains Dan Teknologi)*, vol. 12, no. 1, pp. 129–137, 2023, doi: 10.23887/jstundiksha.v12i1.52358.
- [11] C. J. E. Munthe, N. A. Hasibuan, and H. Hutabarat, "Penerapan Algoritma Text mining Dan TF-RF Dalam Menentukan Promo Produk Pada Marketplace," *Resolusi Rekayasa Teknik Informatika Dan Informasi*, vol. 2, no. 3, pp. 110–115, 2022, doi: 10.30865/resolusi.v2i3.309.
- [12] S. N. Arafah and F. Fathoni, "Sentiment Analysis Pada Masyarakat Terhadap LRT Kota Palembang Menggunakan Metode Improved K-Nearest Neighbor," *Jurnal Media Informatika Budidarma*, vol. 6, no. 3, p. 1554, 2022, doi: 10.30865/mib.v6i3.4434.
- [13] D. F. SURIANTO, "Clustering Tweets Data on Twitter Social Media Using K-Means Method," *J. Of Security, Computer, Inf., Embedded, Netw., and Intelligen*, pp. 44–51, 2023, doi: 10.61220/scientist.v1i2.20232.
- [14] M. Afdal, "Penerapan Text mining Pada Aplikasi Tokopedia Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi*, vol. 8, no. 1, p. 78, 2022, doi: 10.24014/rmsi.v8i1.16595.
- [15] M. B. Priyantono, "Optimasi Sistem Pelabelan Topik Skripsi Menggunakan Algoritma Naive Bayes Dengan Pendekatan Design Thinking," *Jurnal Edukasi Dan Penelitian Informatika (Jepin)*, vol. 8, no. 1, p. 168, 2022, doi: 10.26418/jp.v8i1.50702.
- [16] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, "A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews," in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, Lucknow, India: IEEE, Feb. 2020, pp. 217–220. doi: 10.1109/IC3A48958.2020.233300.
- [17] M. R. Fais Sya' Bani, U. Enri, and T. N. Padilah, "Analisis Sentimen Terhadap Bakal Calon Presiden 2024 Dengan Algoritme Naive Bayes," *Jur. Ris. Kom.*, vol. 9, no. 2, p. 265, Apr. 2022, doi: 10.30865/jurikom.v9i2.3989.
- [18] A. P. Nardilasari, A. L. Hananto, S. S. Hilabi, T. Tukino, and B. Priyatna, "Analisis Sentimen Calon Presiden 2024 Menggunakan Algoritma SVM Pada Media Sosial Twitter," *JOINTECS*, vol. 8, no. 1, p. 11, Mar. 2023, doi: 10.31328/jointecs.v8i1.4265.
- [19] C. C. Aggarwal and C. X. Zhai, "A Survey of Text Classification Algorithms," pp. 163–222, 2012, doi: 10.1007/978-1-4614-3223-4_6.
- [20] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter BMKG Nasional," *Jurnal Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.
- [21] A. Falasari and M. A. Muslim, "Optimize Naive Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency for Amazon Review Sentiment Analysis," *Journal of Soft Computing Exploration*, vol. 3, no. 1, pp. 31–36, 2022, doi: 10.52465/josce.v3i1.68.
- [22] A. B. P. Negara, H. Muhardi, and I. M. Putri, "Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes Dan Seleksi Fitur Information Gain," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, vol. 7, no. 3, p. 599, 2020, doi: 10.25126/jtiik.2020711947.
- [23] Han, Jiawei, Kamber, Michelin, and Pei, Jian, *Data Mining: Concepts and Techniques*. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.