**Jurnal Informatika: Jurnal pengembangan IT**

*Vol. 10, No. 2, 2025*
*ISSN : 2477-5126, DOI:10.30591/jpit.v10i2.7965*

# Klasifikasi Pertanyaan Quora Menggunakan Metode *Keyword-based* dan Analisis Sentimen dengan ComplementNB

*Alwan Adiuntoro[1], Aria Hendrawan[2]*
*[1,2]Fakutas Teknologi Informasi dan Komunikasi, Universitas Semarang, Tlogosari Kulon, Kota Semarang, 50196, Indonesia*

**Info Artikel**

***Riwayat Artikel:***

Received 2024-11-25
Revised 2025-04-23
Accepted 2025-05-01

***Abstract –*** *Text classification is a fundamental task in Natural Language Processing (NLP) that supports the categorization of data based on predefined labels. This study aims to evaluate the effectiveness of keyword-based labeling and sentiment analysis methods for text classification using the Quora Questions dataset. The dataset comprises 16,921 samples with imbalanced class distribution, where the opinion category dominates, while the hypothetical category is a minority class. The labeling process utilized a keyword-based approach for the fact and hypothetical categories, while the opinion category was labeled using sentiment analysis with the Vader Lexicon library. TF-IDF was employed as the feature representation method, with two approaches explored: n-gram range tuning (1–3) and without tuning. ComplementNB, designed for handling imbalanced datasets, was utilized for classification, with a training-test split of 70:30. The results show that the approach without n-gram tuning achieved the highest accuracy of 93.89%, with zero variance in cross-validation. Evaluation revealed that ComplementNB effectively handles class imbalance, as demonstrated by high precision and recall in the minority class. This study demonstrates that a simple approach combining keyword-based labeling and sentiment analysis can be effectively implemented for category-based text classification tasks, particularly in platforms like Quora. These findings are relevant for similar applications requiring real-time text classification with minimal complexity.*

***Keywords:*** *sentiment analysis, ComplementNB, text classification, keyword-based labeling, Quora*

***Corresponding Author:***

Alwan Adiuntoro

Email: alwanadiuntoro@gmail.com



***Abstrak –*** *Klasifikasi teks merupakan salah satu tugas penting dalam Natural Language Processing (NLP) yang mendukung pengelompokan data berbasis kategori. Penelitian ini bertujuan mengevaluasi efektivitas metode pelabelan berbasis keyword dan analisis sentimen dalam klasifikasi teks pada dataset pertanyaan Quora. Dataset berisi 16.921 sampel dengan distribusi kelas tidak seimbang, di mana kategori opinion mendominasi, sementara kategori hypothetical menjadi kelas minoritas. Metode pelabelan menggunakan pendekatan keyword-based untuk kelas fact dan hypothetical, serta analisis sentimen menggunakan library Vader Lexicon untuk kategori opinion. Penggunaan TF-IDF sebagai representasi fitur teks dieksplorasi melalui dua pendekatan: tuning n-gram range (1–3) dan tanpa tuning. Model ComplementNB, yang dirancang untuk dataset tidak seimbang, digunakan untuk klasifikasi dengan pembagian data latih dan data uji sebesar 70:30. Hasil penelitian menunjukkan bahwa pendekatan tanpa tuning menghasilkan akurasi tertinggi, yaitu 93,89%, dengan variansi nol pada evaluasi cross-validation. Evaluasi menunjukkan bahwa ComplementNB mampu menangani ketidakseimbangan kelas dengan baik, ditunjukkan oleh precision dan recall yang tinggi pada kelas minoritas. Penelitian ini memberikan bukti bahwa pendekatan sederhana berbasis keyword dan analisis sentimen dapat diimplementasikan secara efektif dalam tugas klasifikasi teks berbasis kategori, khususnya pada platform seperti Quora. Temuan ini relevan untuk aplikasi serupa yang membutuhkan klasifikasi teks real-time dengan tingkat kompleksitas rendah.*

***Kata Kunci****: analisis sentimen, ComplementNB, klasifikasi teks, pelabelan berbasis keyword, Quora*

## I. PENDAHULUAN

Peran media sosial sebagai media penyebaran pengetahuan dan pengalaman menjadi hal yang menarik dan semakin dibutuhkan[1]. Media sosial berperan penting dalam penyebaran pengetahuan dengan memfasilitasi berbagi informasi, meningkatkan keterlibatan pengguna, dan mendukung kolaborasi serta diskusi publik[2], [3]. Quora adalah salah satu media sosial yang berperan langsung sebagai platform tanya jawab yang mampu membuat konsensus dari diskusi yang dibuat oleh penggunanya[4].

Dalam beberapa dekade terakhir, perkembangan kecerdasan buatan telah mengubah cara manusia berinteraksi dengan informasi. Salah satu bidang AI yang terus berkembang adalah *Natural Language Processing* (NLP), teknologi inti di balik sistem pencarian, chatbot, dan aplikasi berbasis teks lainnya[5]. Klasifikasi teks merupakan tugas paling mendasar dan penting dalam NLP. Banyak metode, kumpulan data, dan metrik evaluasi telah diusulkan dalam literatur, yang meningkatkan kebutuhan akan survei yang komprehensif dan terkini[6]. Teknik klasifikasi teks memiliki peran penting dalam merancang model data yang disesuaikan, meningkatkan efisiensi operasional, dan mendukung pengambilan keputusan yang lebih efektif di berbagai industri[7].

Dalam konteks media sosial, klasifikasi teks digunakan untuk menganalisis data secara real-time, dengan algoritma seperti *Support Vector Machine* (SVM), Naïve Bayes *classifiers*, dan *decision trees* menjadi yang paling umum diterapkan[8]. Hartmann et al. membandingkan beberapa algoritma klasifikasi teks pada data tidak terstruktur di media sosial. Penelitian mereka menemukan bahwa *Random Forest* dan Naïve Bayes memberikan performa terbaik dalam mengungkap intuisi manusia, menunjukkan kemampuan algoritma ini dalam mengklasifikasikan teks tidak terstruktur dengan akurasi tinggi[9].

Platform seperti Quora menghadapi tantangan besar untuk memahami dan mengelompokkan berbagai jenis pertanyaan yang diajukan oleh pengguna dari seluruh dunia. Pertanyaan di Quora tidak hanya berupa fakta sederhana, tetapi juga melibatkan opini personal dan spekulasi hipotetis. Perbedaan kategori ini bukan hanya sekadar label, ia merepresentasikan cara pertanyaan harus dijawab. Quora sebagai platform yang tidak hanya menyebarkan pengetahuan tetapi juga membangun konsensus berbasis diskusi publik yang berkualitas[4].

Sebagai platform tanya jawab, Quora memanfaatkan interaksi antar penggunanya untuk menghasilkan jawaban yang akurat dan relevan sesuai dengan konteks pertanyaan[3]. Namun untuk memaksimalkan fungsinya, sistem NLP pada Quora harus mampu mengelompokkan pertanyaan ke dalam kategori yang sesuai, seperti *fact*, *opinion*, atau *hypothetical*. Klasifikasi yang tepat bukan hanya mendukung kualitas jawaban, tetapi juga meningkatkan pengalaman pengguna dengan memastikan bahwa diskusi di dalam platform tersebut terarah dan efisien[10].

Penelitian ini menggunakan dataset Quora questions answers yang diperoleh dari Hugging Face. Namun, dataset tidak dirancang secara spesifik untuk klasifikasi kategori seperti ini, sehingga sering kali menghasilkan data yang bias ke kelas mayoritas[11]. Di sinilah dibutuhkan pendekatan yang sederhana tetapi efektif untuk menghadapi keterbatasan ini.

Dalam memproses kasus ini, peneliti menggunakan hanya kolom pertanyaan dan hanya menggunakan satu persen dari dataset. Pertanyaan diklasifikasikan berdasarkan kategori *fact*, *opinion*, dan *hypothetical*. Analisis dilakukan dengan menggunakan analisis sentimen dari sklearn vader lexicon untuk mengidentifikasi pertanyaan yang bersifat opini, dan *keyword-based* untuk mengidentifikasi pertanyaan yang bersifat hipotesis, dan pertanyaan fakta ditentukan dengan *keyword-based* dan dari pertanyaan yang bersifat netral.

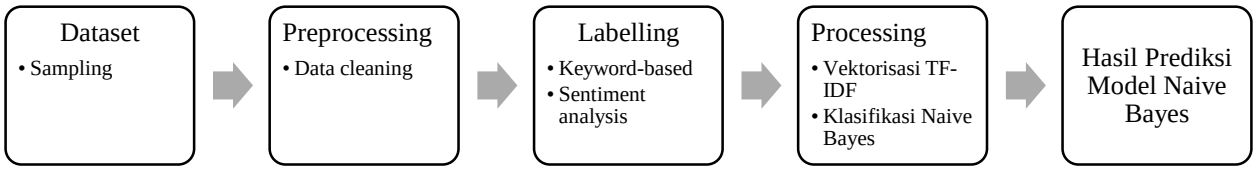
*Literature review* dilakukan oleh Ahamed, N., & Ahangama, S. pada tahun 2023 meninjau mengenai penelitian pada dataset yang sama. Makalah ini mengulas berbagai pendekatan dan algoritma yang digunakan dalam mengklasifikasikan pertanyaan yang tidak tulus di Quora, manfaatnya, dan kekurangannya. Temuan tinjauan menunjukkan bahwa ketidakseimbangan kelas merupakan masalah umum dalam jenis penelitian ini[11]. Sebagian besar algoritma *machine learning* condong ke kelas mayoritas dan kesulitan dalam mengklasifikasi kelas minoritas[17]. Distribusi kelas yang tidak seimbang dalam model pembelajaran mendalam dapat menyebabkan bias frekuensi dan kesulitan belajar membedakan batas antara kelas minoritas dan mayoritas[18].

Menanggapi tantangan ini, penelitian ini menawarkan solusi berbasis model Naïve Bayes, khususnya ComplementNB, yang dikenal mampu menangani dataset dengan distribusi tidak seimbang[12], [13], [14]. Dengan kombinasi preprocessing yang sesuai[15] dan eksplorasi pola frasa menggunakan n-gram (1–3)[16], penelitian ini berusaha membuktikan bahwa framework sederhana dapat mencapai akurasi yang kompetitif, bahkan pada dataset kecil hingga sedang. Penelitian ini menawarkan pendekatan berbasis metode sederhana yang efektif untuk menangani tantangan klasifikasi teks pada dataset tidak seimbang, dengan menggunakan TF-IDF dan model ComplementNB sebagai solusi utama.

Penelitian ini bertujuan untuk mengevaluasi efektivitas metode *keyword-based* dan *sentiment analysis* untuk melabeli data kategori teks, yang kemudian digunakan dalam membangun model klasifikasi menggunakan ComplementNB. Metode ini dirancang untuk menangkap pola linguistik unik dari tiap kategori, seperti fakta yang bersifat deskriptif, opini dengan sentimen tertentu, dan hipotesis yang mengandung elemen spekulatif. Implementasi model pada dataset yang telah dilabeli menunjukkan hasil yang konsisten dan signifikan.

## II. METODE

Penelitian ini melalui beberapa langkah untuk menyelesaikan proyek penelitian yang digambarkan pada Gambar 1.




Gambar 1. Tahapan Penelitian

*A. Dataset*

Peneliti menggunakan 30% dari dataset dan hanya menggunakan satu dimensi kolom question sebagai representasi data sedang-kecil[19]. Data diperoleh dari dataset Quora question answer yang berisi kolom question dan kolom answer sebagai data yang belum diproses. Dataset ini memiliki 56.402 baris yang dipublish pada tanggal 24 Agustus 2023.

*B. Preprocessing*

Preprocessing sangat penting untuk untuk pemodelan dan meningkatkan hasil aplikasi NLP [20]. Dalam menangani data yang semakin kompleks dan besar, preprocessing memastikan efisiensi dan ketahanan[21]. Dalam kasus data yang tidak seimbang, preprocessing membantu mengurangi cacat dan meningkatkan kualitas data yang digunakan dalam studi pembelajaran mesin[22]. Preprocessing pada penelitian ini hanya menggunakan data cleaning untuk vektorisasi TF-IDF dan *sentiment analysis*. Metode *keyword-based* akan menggunakan data yang belum dibersihkan demi mempertahankan konteks.

Proses *data cleaning* mencakup penghapusan *noise*, perbaikan ketidakkonsistenan, serta penghilangan data yang tidak relevan atau berlebihan untuk meningkatkan kualitas data[23]. Penelitian ini menerapkan *case folding* untuk menjaga konsistensi teks. Selain itu, dilakukan juga penghapusan URL, angka, tanda baca, dan *stopwords* guna mengurangi *noise* dan meningkatkan relevansi data.

TABEL 1
HASIL DATA CLEANING UNTUK PERTANYAAN QUORA

| **Data Mentah** | ***Clean Data*** |
|---|---|
| *Is it true that Lord Shiva is not mentioned in the Vedas?* | *true lord shiva mentioned vedas* |
| *What is the most dangerous place on Earth?* | *dangerous place earth* |
| *Do all entrepreneurs need an outside investor?* | *entrepreneurs need outside investor* |
| *CNN reported that 1,000 US health experts have demanded that businesses not be allowed to reopen, but protestors should not be criticized for violating suggestions and increasing infection risks. Does science not matter anymore?* | *cnn reported us health experts demanded businesses allowed reopen protestors criticized violating suggestions increasing infection risks science matter anymore* |
| *What are some things that only electrical and electronics engineers know, but most people don't?* | *things electrical electronics engineers know people dont* |

*C. Labelling*

Dataset pertanyaan yang telah disampling diberi label *fact*, *hypothetical*, dan *opinion*. Proses pelabelan untuk kategori *fact* dan *hypothetical* dilakukan menggunakan pendekatan *keyword-based*, yaitu dengan mengidentifikasi kata-kata yang secara spesifik merepresentasikan karakteristik fakta atau hipotesis. Preprocessing diterapkan setelah tahap pelabelan *keyword-based* agar tidak menghilangkan elemen esensial dari teks yang mendukung pelabelan tersebut. Selanjutnya, analisis sentimen dilakukan dengan library Vader Lexicon untuk mendeteksi pola emosional yang menjadi indikator utama teks berlabel *opinion*.

Penelitian ini mengurutkan prioritas pelabelan dengan kategori *hypothetical* sebagai prioritas utama, diikuti oleh *opinion* dan *fact*. Kategori *fact* dilabeli berdasarkan karakteristik deskriptif pertanyaan yang tidak mengandung sentimen, biasanya diawali dengan frasa seperti *"what is," "what are," "who is," "who are," "where is," "define," "explain," "how does,"* dan *"why does."* Jika pertanyaan tidak dapat teridentifikasi ke dalam kategori tersebut, maka dilabeli sebagai *others*.

TABEL 2
HASIL PELABELAN KATEGORI *FACT* DENGAN *KEYWORD-BASED* LABELING

| **Data Mentah** | **Clean Data** | **Label** |
|---|---|---|
| *Is it true that Lord Shiva is not mentioned in the Vedas?* | *true lord shiva mentioned vedas* | *others* |
| *What is the most dangerous place on Earth?* | *dangerous place earth* | *others* |
| *Do all entrepreneurs need an outside investor?* | *entrepreneurs need outside investor* | *fact* |
| *CNN reported that 1,000 US health experts have demanded that businesses not be allowed to reopen, but protestors should not be criticized for violating* | *cnn reported us health experts demanded businesses allowed reopen protestors criticized violating suggestions increasing infection risks science matter anymore* | *others* |

| | | |
|---|---|---|
| *suggestions and increasing infection risks. Does science not matter anymore?* | | |
| *What are some things that only electrical and electronics engineers know, but most people don't?* | *things electrical electronics engineers know people dont* | *fact* |

Metode selanjutnya adalah menerapkan analisis sentimen pada setiap pertanyaan. Analisis ini menghasilkan skor dalam rentang -1 hingga 1, yang merepresentasikan sentimen negatif hingga positif. Dalam penelitian ini, skor tersebut tidak digunakan untuk mengevaluasi sentimen, melainkan untuk melabeli pertanyaan dengan sentimen sebagai kategori *opinion*.

TABEL 3
HASIL PELABELAN KATEGORI *OPINION* DENGAN *SENTIMENT ANALYSIS*

| Data Mentah | *Clean Data* | Label | *Question Sentiment* |
|---|---|---|---|
| *Is it true that Lord Shiva is not mentioned in the Vedas?* | *true lord shiva mentioned vedas* | *opinion* | 0.4215 |
| *What is the most dangerous place on Earth?* | *dangerous place earth* | *opinion* | -0.4767 |
| *Do all entrepreneurs need an outside investor?* | *entrepreneurs need outside investor* | *fact* | 0.0 |
| *CNN reported that 1,000 US health experts have demanded that businesses not be allowed to reopen, but protestors should not be criticized for violating suggestions and increasing infection risks. Does science not matter anymore?* | *cnn reported us health experts demanded businesses allowed reopen protestors criticized violating suggestions increasing infection risks science matter anymore* | *opinion* | -0.836 |
| *What are some things that only electrical and electronics engineers know, but most people don't?* | *things electrical electronics engineers know people dont* | *fact* | 0.0 |

Kategori *hypothetical* dilabeli berdasarkan kata kunci yang mencerminkan sifat hipotesis, seperti *"suppose," "imagine," "what if," "assume," "assuming," "could," "would,"* dan sejenisnya. Pelabelan dilakukan secara sistematis untuk memastikan konsistensi antar data. Pertanyaan yang bersifat netral dan tidak masuk dalam kategori hipotesis diklasifikasikan sebagai *fact*.

TABEL 4
HASIL PELABELAN KATEGORI *HYPOTHESIS* DENGAN *KEYWORD-BASED* LABELING

| Data Mentah | *Clean Data* | Label | *Question Sentiment* |
|---|---|---|---|
| *Is it true that Lord Shiva is not mentioned in the Vedas?* | *true lord shiva mentioned vedas* | *opinion* | 0.4215 |
| *What is the most dangerous place on Earth?* | *dangerous place earth* | *opinion* | -0.4767 |
| *Do all entrepreneurs need an outside investor?* | *entrepreneurs need outside investor* | *fact* | 0.0 |
| *CNN reported that 1,000 US health experts have demanded that businesses not be allowed to reopen, but protestors should not be criticized for violating suggestions and increasing infection risks. Does science not matter anymore?* | *cnn reported us health experts demanded businesses allowed reopen protestors criticized violating suggestions increasing infection risks science matter anymore* | *hypothetical* | -0.836 |
| *What are some things that only electrical and electronics engineers know, but most people don't?* | *things electrical electronics engineers know people dont* | *fact* | 0.0 |

Proses pelabelan menghasilkan data yang telah terstruktur dengan label pada setiap pertanyaan. Visualisasi hasil pelabelan disajikan pada Gambar 2.

| label | |
|---|---|
| opinion | 8218 |
| fact | 5716 |
| hypothetical | 2987 |

dtype: int64

Gambar 2. *Count* Label *Opinion, Fact,* dan *Hypothetical*

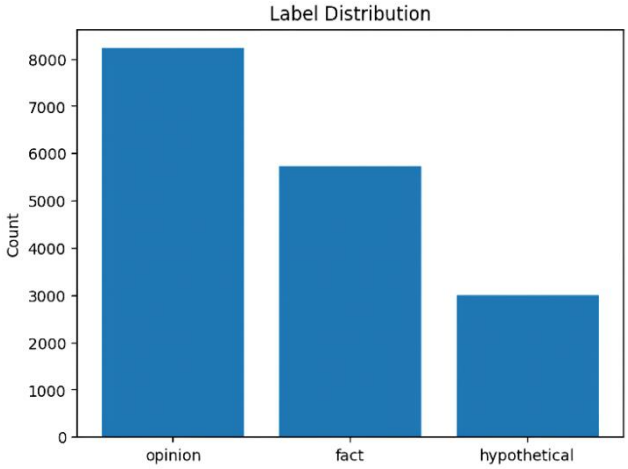
*D. Processing*

Dalam penelitian ini, representasi teks menggunakan TF-IDF diimplementasikan dengan dua pendekatan. Tuning n-gram range (1–3) untuk menangkap pola frasa unik, dan TF-IDF tanpa tuning yang hanya fokus pada *unigrams*. Proses klasifikasi dilakukan menggunakan algoritma Naïve Bayes, yang cocok untuk tugas klasifikasi teks karena kesederhanaannya sebagai pengklasifikasi probabilistik dengan asumsi independensi antar fitur[24]. Model ComplementNB dipilih karena dirancang untuk menangani data dengan kategori tidak seimbang, memastikan performa yang lebih baik pada kelas minoritas[12], [14].

Data dilatih dan diuji menggunakan komposisi 70:30, memungkinkan evaluasi akurasi model secara menyeluruh[25]. Evaluasi dilakukan dengan menggunakan confusion matrix dan laporan metrik seperti *precision*, *recall*, dan F1-score melalui library scikit-learn untuk menganalisis performa klasifikasi di setiap kelas. *Confusion matrix* digunakan untuk menganalisis kemampuan model dalam mengenali distribusi kelas, baik mayoritas maupun minoritas, pada setiap kategori.

## III. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 16.921 sampel dari dataset Quora Questions. Kategori dalam dataset menunjukkan distribusi yang tidak seimbang. Kelas *opinion* mendominasi hampir 50% total sampel, sementara kelas *hypothetical* hanya mencapai kurang dari 20%.




Gambar 3. *Bar Chart* Kategori *Opinion, Fact,* dan *Hypothetical*

Setelah menerapkan tuning *n-gram range* 1–3 kata dan menetapkan ukuran data uji sebesar 30% dari sampel, klasifikasi menggunakan model ComplementNB menghasilkan akurasi sebesar 88.48%. Proses evaluasi lanjutan menggunakan 5-fold cross validation menunjukkan akurasi sebesar 89%. Variansi sebesar 0 pada hasil evaluasi ini merepresentasikan bahwa model memberikan performa yang sangat konsisten di berbagai split data uji.

Analisis confusion matrix menunjukkan kemampuan model dalam mengenali data dari kelas mayoritas dan minoritas secara seimbang. Berdasarkan evaluasi *precision*, *recall*, dan F1-score, model menunjukkan performa stabil di semua kelas, dengan skor keseluruhan di atas 80%. Pada kelas *hypothetical* yang merupakan

kelas minoritas, *precision* dan *recall* memiliki nilai tinggi dengan selisih minimal, menandakan model dapat mengenali pola pada kelas ini dengan baik. Sementara itu, kelas *opinion* sebagai kelas mayoritas tetap menghasilkan performa yang seimbang dengan kelas lainnya.

Pada eksperimen setelahnya dengan menerapkan TF-IDF tanpa tuning, model ComplementNB mengalami kenaikan signifikan pada akurasi dari 88.48% menjadi 93.89%. TF-IDF tanpa tuning cenderung hanya fokus pada kata-kata tunggal (*unigrams*) yang lebih general. Ini sejalan dengan pola pada ComplementNB yang asumsinya sederhana yaitu independen antar fitur. Peningkatan akurasi pada TF-IDF tanpa tuning menunjukkan bahwa pendekatan berbasis unigram mampu menangkap pola kata yang lebih general tanpa memicu overfitting dari fitur n-gram yang jarang muncul.

```
test_size =  0.3
model:  ComplementNB()
CV Accuracy: 0.89 ± 0.00
Confusion Matrix:
 [[1509   76  145]
 [  62  742   65]
 [ 142   95 2241]]
Classification Report:
              precision    recall  f1-score   support

        fact       0.88      0.87      0.88      1730
hypothetical       0.81      0.85      0.83       869
     opinion       0.91      0.90      0.91      2478

    accuracy                           0.88      5077
   macro avg       0.87      0.88      0.87      5077
weighted avg       0.89      0.88      0.89      5077

Accuracy: 88.48%
==========================
```

(a)

```
test_size =  0.3
model:  ComplementNB()
CV Accuracy: 0.94 ± 0.00
Confusion Matrix:
 [[1610   34   86]
 [  32  809   28]
 [  66   64 2348]]
Classification Report:
              precision    recall  f1-score   support

        fact       0.94      0.93      0.94      1730
hypothetical       0.89      0.93      0.91       869
     opinion       0.95      0.95      0.95      2478

    accuracy                           0.94      5077
   macro avg       0.93      0.94      0.93      5077
weighted avg       0.94      0.94      0.94      5077

Accuracy: 93.89%
```

(b)

Gambar 4. Perbandingan Akurasi Model ComplementNB (a) dengan TF-IDF Tuning Range 1-3 Kata dan (b) tanpa TF-IDF Tuning.

Hasil evaluasi menunjukkan bahwa ComplementNB memberikan performa stabil pada dataset dengan distribusi kelas tidak seimbang. Model ini memitigasi bias terhadap kelas mayoritas (*opinion*), sehingga kelas minoritas seperti *hypothetical* tetap dikenali dengan baik. *Precision* dan *recall* untuk kelas *hypothetical* mencapai nilai tinggi dengan nilai masing-masing 89% dan 93%, menunjukkan bahwa model mampu menangkap pola linguistik unik, seperti penggunaan frasa *"what if"* atau *"suppose"*.

Meskipun hasilnya positif, terdapat tantangan pada kelas *hypothetical* karena konteks hipotesis sering kali ambigu atau tumpang tindih dengan kategori lainnya. Pertanyaan "*CNN reported that 1,000 US health experts have demanded that businesses not be allowed to reopen, but protestors should not be criticized for violating suggestions and increasing infection risks. Does science not matter anymore?"* memiliki nilai sentimen sebesar -0.836 sehingga dapat dikategorikan sebagai opini, namun karena pertanyaan ini terdapat kata yang mengandung hipotesis maka dikategorikan sebagai kelas *hypothetical*. Sementara *"What if we stopped using fossil fuels?"*

jelas bersifat hipotetis. Ambiguitas ini menunjukkan bahwa metode sederhana seperti *keyword-based* belum sepenuhnya menangkap konteks kompleks yang diperlukan untuk mengidentifikasi kategori ini secara akurat. Penelitian lanjutan dapat mengeksplorasi model berbasis transformer seperti BERT (*Bidirectional Encoder Representations from Transformers*) untuk mengatasi masalah ini dengan memahami konteks secara mendalam.

Pendekatan yang digunakan dalam penelitian ini memiliki relevansi tinggi dalam dunia nyata, terutama untuk platform yang berbasis teks seperti Quora. Kemampuan untuk mengklasifikasikan pertanyaan menjadi kategori *fact*, *opinion*, dan *hypothetical* memungkinkan sistem untuk memberikan respons yang lebih relevan dan kontekstual. Pada penerapannya kelas *fact* dapat diarahkan ke jawaban yang bersifat deskriptif, kelas *opinion* dapat digunakan untuk memfasilitasi diskusi komunitas atau memberikan saran personal, dan kelas *hypothetical* cocok untuk mendorong debat spekulatif atau pengembangan ide kreatif. Selain itu, metode ini dapat diterapkan pada sistem chatbot, di mana pemahaman konteks pertanyaan menjadi kunci untuk meningkatkan kualitas respons. Dengan hasil yang stabil dan akurasi tinggi, metode ini berpotensi diterapkan dalam sistem berbasis AI lainnya seperti chatbot layanan pelanggan atau analisis data tanya-jawab real-time pada platform edukasi.

Eksperimen menunjukkan bahwa metode *keyword-based* dan analisis sentimen memberikan hasil klasifikasi teks yang kompetitif ketika digunakan bersama TF-IDF tanpa tuning, dengan akurasi mencapai 93.89%. Pendekatan tanpa tuning ini memanfaatkan pola kata individual secara efektif, tanpa memicu kompleksitas tambahan dari n-gram yang berpotensi memperbesar fitur yang jarang muncul. Hasil ini mengindikasikan bahwa pendekatan sederhana dapat lebih efektif dalam dataset dengan pola kata individual yang kuat. Hasil ini juga menegaskan bahwa fitur sederhana berbasis frekuensi kata tetap menjadi representasi yang kuat untuk klasifikasi teks.

## IV. SIMPULAN

Hasil evaluasi menunjukkan bahwa metode pelabelan dengan *keyword-based* dan analisis sentimen sangat efektif untuk mendukung klasifikasi teks menggunakan model ComplementNB, dengan akurasi mencapai 93,89%. Model menunjukkan stabilitas performa pada dataset tidak seimbang, dengan *precision* dan *recall* yang konsisten untuk kelas mayoritas maupun minoritas. Selain itu, variansi akurasi yang bernilai nol pada evaluasi cross-validation menegaskan konsistensi model bahkan ketika diterapkan pada data dengan distribusi label yang tidak merata. Pendekatan TF-IDF tanpa tuning menghasilkan akurasi yang lebih tinggi dibandingkan dengan tuning n-gram range 1–3. Meskipun terdapat tantangan pada kelas *hypothetical*, model tetap mampu untuk mengenali kelas minoritas dengan *precision* dan *recall* masing-masing sebesar 89% dan 93%. Representasi sederhana ini terbukti lebih optimal untuk tugas klasifikasi berbasis kategori teks, khususnya pada dataset yang menggunakan pelabelan dengan *keyword-based* dan analisis sentimen. Kombinasi metode ini memberikan keunggulan dalam menyederhanakan kompleksitas model sekaligus mempertahankan performa tinggi pada dataset tanya-jawab Quora. Meskipun penelitian ini membuktikan efektivitas pendekatan sederhana pada dataset kecil hingga sedang, eksplorasi lebih lanjut dapat dilakukan pada dataset berukuran besar atau dengan menggunakan model yang lebih kompleks seperti *transformer-based* untuk menguji generalisasi metode. Dengan demikian, penelitian ini memberikan bukti empiris bahwa pendekatan *keyword-based* dan analisis sentimen dapat digunakan sebagai fondasi yang efektif dalam klasifikasi teks berbasis kategori, khususnya pada platform tanya-jawab seperti Quora, sekaligus membuka peluang eksplorasi lebih lanjut dengan dataset yang lebih besar atau metode berbasis *Machine learning*.

## DAFTAR PUSTAKA

[1] H. Ihsaniyati, S. Sarwoprasodjo, P. Muljono, and D. Gandasari, "The Use of Social Media for Development Communication and Social Change: A Review," 2023. doi: 10.3390/su15032283.

[2] Y. A. Ahmed, M. N. Ahmad, N. Ahmad, and N. H. Zakaria, "Social media for knowledge-sharing: A systematic literature review," 2019. doi: 10.1016/j.tele.2018.01.015.

[3] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: An analysis of Quora," in *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[4] S. D. Anggraeni, "Quora: Situs Komunitas Tanya Jawab Sebagai Medium Diskursus Ruang Publik," *Jurnal Sosia Logica*, vol. 2, no. 1, pp. 1–14, 2023.

[5] I. Dergaa, K. Chamari, P. Zmijewski, and H. Ben Saad, "From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing," *Biol Sport*, vol. 40, no. 2, 2023, doi: 10.5114/BIOLSPORT.2023.125623.

[6] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," 2022. doi: 10.1145/3495162.

[7] M. Thangaraj and M. Sivakami, "Text classification techniques: A literature review," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 13, 2018, doi: 10.28945/4066.

[8] D. Rogers, A. Preece, M. Innes, and I. Spasić, "Real-Time Text Classification of User-Generated Content on Social Media: Systematic Review," 2022. doi: 10.1109/TCSS.2021.3120138.

[9] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[10] M. Chandra, A. Rodrigues, and J. George, "An Enhanced Deep Learning Model for Duplicate Question Detection on Quora Question pairs using Siamese LSTM," in *IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, ICDCECE 2022*, 2022. doi: 10.1109/ICDCECE53908.2022.9792906.

[11] N. Ahamed and S. Ahangama, "A Review of Classification of Insincere Questions in Quora Using Deep Learning Approaches," in *2023 IEEE 17th International Conference on Industrial and Information Systems, ICIIS 2023 - Proceedings*, 2023. doi: 10.1109/ICIIS58898.2023.10253595.

[12] B. Marapelli, S. Kadiyala, and C. S. Potluri, "Performance Analysis and Classification of Class Imbalanced Dataset Using Complement Naive Bayes Approach," in *Proceedings of the ACCTHPA 2023 - Conference on Advanced Computing and Communication Technologies for High Performance Applications*, 2023. doi: 10.1109/ACCTHPA57160.2023.10083369.

[13] H. Florenci Tapikap, B. S. Djahi, and T. Widiastuti, "MAIL MENGGUNAKAN METODE TRANSFORMED COMPLEMENT NAÏVE BAYES (TCNB)," *J-ICON*, vol. 7, no. 1, 2019.

[14] J. D. M. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," in *Proceedings, Twentieth International Conference on Machine learning*, 2003.

[15] MR ADEPU RAJESH and DR TRYAMBAK HIWARKAR, "Exploring Preprocessing Techniques for Natural LanguageText: A Comprehensive Study Using Python Code," *international journal of engineering technology and management sciences*, vol. 7, no. 5, 2023, doi: 10.46647/ijetms.2023.v07i05.047.

[16] S. S. M. M. Rahman, K. B. M. B. Biplob, M. H. Rahman, K. Sarker, and T. Islam, "An investigation and evaluation of N-gram, TF-IDF and ensemble methods in sentiment classification," in *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, 2020. doi: 10.1007/978-3-030-52856-0_31.

[17] M. Bader-El-Den, E. Teitei, and T. Perry, "Biased Random Forest for Dealing with the Class Imbalance Problem," *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 7, 2019, doi: 10.1109/TNNLS.2018.2878400.

[18] K. R. M. Fernando and C. P. Tsokos, "Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 7, 2022, doi: 10.1109/TNNLS.2020.3047335.

[19] J. A. Prenner and R. Robbes, "Making the Most of Small Software Engineering Datasets with Modern *Machine learning*," *IEEE Transactions on Software Engineering*, vol. 48, no. 12, 2022, doi: 10.1109/TSE.2021.3135465.

[20] C. P. Chai, "Comparison of text preprocessing methods," *Nat Lang Eng*, vol. 29, no. 3, 2023, doi: 10.1017/S1351324922000213.

[21] A. W. Blocker and X. L. Meng, "The potential and perils of preprocessing: Building new foundations," *Bernoulli*, vol. 19, no. 4, 2013, doi: 10.3150/13-BEJSP16.

[22] E. A. Felix and S. P. Lee, "Systematic literature review of preprocessing techniques for imbalanced data," 2019. doi: 10.1049/iet-sen.2018.5193.

[23] *Encyclopedia of Machine learning and Data Mining*. 2017. doi: 10.1007/978-1-4899-7687-1.

[24] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J Inf Sci*, vol. 44, no. 1, 2018, doi: 10.1177/0165551516677946.

[25] H. A. Parhusip, B. Susanto, L. Linawati, S. Trihandaru, Y. Sardjono, and A. S. Mugirahayu, "Classification Breast Cancer Revisited with *Machine learning*," *International Journal of Data Science*, vol. 1, no. 1, 2020, doi: 10.18517/ijods.1.1.42-50.2020.