

# IMPLEMENTASI METODE *NAÏVE BAYES* DAN *LEXICON-BASED APPROACH* UNTUK MENGLASIFIKASI SENTIMEN NETIZEN PADA *TWEET* BERBAHASA INDONESIA

Liza Wikarsa<sup>1,\*</sup>, Apriandy Angdresey<sup>1</sup>, Juan D. Kapantow<sup>1</sup>

<sup>1</sup>Program Teknik Informatika, Fakultas Teknik

<sup>1</sup>Universitas Katolik De La Salle Manado

e-mail: <sup>1</sup>lwikarsa@unikadelasalle.ac.id, <sup>1</sup>aangdresey@unikadelasalle.ac.id, <sup>1</sup>16013037@unikadelasalle.ac.id

**Abstrak** – Sosial media, seperti Twitter, banyak digunakan untuk menghubungkan pengguna aplikasi ini di seluruh dunia tanpa dibatasi waktu dan tempat. Tautan Twitter banyak berisi opini dan sentimen dari pengguna terkait topik dan kejadian yang lagi trendi. Analisis sentimen pada teks media sosial telah mendapat banyak perhatian di sektor publik dan komersial. Terdapat banyak penelitian di bidang ini yang mengevaluasi sentimen dari suatu bagian teks, tetapi hasil penelitian sebelumnya belum memadai dan analisis sentimen yang dilakukan mengabaikan konteks. Pengklasifikasian sentimen dari penelitian terkait hanya dibatasi dengan 3 kelas sentimen, yaitu positif, negatif, dan netral sehingga dirasakan belum cukup ekspresif dan informatif. Analisis sentimen menggunakan leksikon berbahasa Indonesia juga terbatas. Oleh karena itu, penelitian ini bertujuan untuk membangun aplikasi pengklasifikasian sentimen netizen pada tweets berbahasa Indonesia dengan menggabungkan metode *Naïve Bayes* dan pendekatan *lexicon-based*. Ada 3 topik utama yang diangkat dalam penelitian ini, yaitu pendidikan, kesehatan, dan teknologi, di mana sentimen pengguna Twitter akan diklasifikasikan ke dalam 7 sentimen, seperti senang, sedih, marah, kaget, takut, jijik, dan netral. Sentimen pengguna ini akan diproses pada 3 level yang berbeda, yakni *word-level*, *tweet-level*, dan *topic-level*, dengan jumlah data sebanyak 1,000 tweets untuk tiap topik. Hasil pengujian menunjukkan bahwa rasio pembagian data latih dan data uji, 70:30, dapat memberikan hasil terbaik. Pada *word-level*, hasil yang mampu dicapai dalam penelitian ini adalah 79,17% akurasi, 89,5% presisi, dan 87,5% recall. Hal ini dipengaruhi oleh jumlah kata unik yang ada di *seed lexicon*. Akurasi, presisi dan recall yang didapat pada *tweet-level* adalah sebanyak 83%, 83,62%, dan 98,36% untuk topik pendidikan *tweet-level*. Sedangkan topik teknologi mendapatkan hasil terendah karena hanya memiliki 69% akurasi, presisi 71,01%, dan recall 93,78% Untuk *topic-level*, sentimen senang mendominasi ketiga topik tersebut.

**Kata Kunci** – *Lexicon*, *Klasifikasi*, *Naive Bayes*, *Sentimen Analisis*, *Twitter*

## I. PENDAHULUAN

Twitter merupakan salah satu media sosial yang paling banyak digunakan sebagai *platform* mediasi yang menghubungkan orang-orang di seluruh dunia dengan membangun jaringan di mana mereka dapat bertukar atau bertransaksi barang, jasa, dan informasi pada waktu tertentu. Di tahun 2022, Twitter mempunyai 220 juta pengguna aktif di dunia dan Indonesia tercatat sebagai negara pengguna Twitter ke-5 terbanyak di dunia dengan jumlah pengguna aplikasi ini

mencapai 18,45 juta orang [1] [2] [3]. Selain itu, terdapat 500 juta *tweets* yang diunggah pada media sosial ini setiap harinya [2]. Tautan *tweet* berisi opini dan sentimen dari pengguna terhadap suatu topik atau kejadian yang menarik perhatiannya. Tidak mengherankan jumlah konten yang dibuat pengguna di situs ini terus meningkat secara eksponensial selama bertahun-tahun.

Konten yang dibuat pengguna di Twitter berisi informasi berharga yang dapat diekstraksi dan dianalisis menggunakan teknik *data mining*, seperti analisis sentimen yang mengukur kecenderungan opini orang melalui teks [4]. Banyak perusahaan maupun individu menggunakan data Twitter untuk ulasan produk dan layanan, perencanaan strategi politik, proyeksi finansial dan lain sebagainya. Analisis sentimen dapat dilakukan dengan menggunakan metode *machine learning* dan pendekatan *lexicon-based*. Salah satu metode klasifikasi dari *machine learning* yang banyak digunakan dalam penelitian adalah *Naïve Bayes* karena memiliki tingkat akurasi tinggi dan mampu mengklasifikasi data dengan jumlah yang besar [5] [6]. *Naïve Bayes* mudah dipahami dan diterapkan karena perhitungannya cepat dan efisien. Selain itu, metode ini hanya membutuhkan jumlah data pelatihan yang kecil untuk mengestimasi parameter yang digunakan dalam proses pengklasifikasian [5] [7]. *Naïve Bayes* juga memiliki kelemahan di mana kualitas data latih tergantung pada hubungan antar atribut dan nilai probabilitasnya tidak boleh nol (0) [6]. Untuk melatih data dibutuhkan waktu dan usaha yang lebih dari pengguna [5].

Pendekatan *lexicon-based* mendeteksi sentimen dengan memanfaatkan daftar kata/frasa (leksikon) di mana setiap kata/frasa dikaitkan dengan sentimen tertentu. Ada 2 metode untuk pendekatan ini, yaitu *dictionary based* dan *corpus-based*. Pendekatan *dictionary-based* biasanya digunakan ketika kata-kata sentimen baru (selanjutnya disebut kata kandidat) diidentifikasi oleh hubungannya dengan *seed lexicon* dan dengan demikian mencari sinonim dan antonimnya dalam kamus [8]. Sedangkan pendekatan *corpus-based* biasanya digunakan ketika kata kandidat baru dikenali berdasarkan hubungannya satu sama lain. Pendekatan ini menggunakan kata penghubung seperti kata ‘dan, atau’ serta kata penghubung lainnya yang memiliki relasi untuk melakukan pencocokan kata [9]. Pendekatan *lexicon-based* berkinerja baik untuk klasifikasi sentimen secara *real-time* karena kemampuan beradaptasinya

terhadap perubahan cepat pada konten di situs media sosial [4] [10]. Pendekatan ini memungkinkan penggunaan penilaian leksikon ketika tidak ada data latih yang memadai. Kekurangan dari pendekatan leksikon adalah sulitnya untuk membangun leksikon sentimen khusus domain karena penggunaan sentimen yang berbeda oleh pengguna untuk menyampaikan pendapat mereka di domain yang berbeda [11]. Oleh karena itu, efektivitas leksikon yang ada dapat menjadi masalah ketika diterapkan pada masalah atau domain baru. Selain itu, pembuatan leksikon sentimen secara manual sangat memakan waktu dan biaya.

Terdapat beberapa *general lexicon* (GL) yang sudah siap dipakai, seperti SentiWordNet (SWN), SentiStrength, dan Opinion Lexicon (OL) [5] [10]. Akan tetapi, GL ini berbahasa Inggris dan hanya menggunakan 3 sentimen, seperti positif, netral, dan negatif, sehingga dirasakan kurang informatif dan ekspresif [11]. Jika seseorang menggunakan GL untuk *domain-specific tasks*, kesalahpahaman informasi yang parah dapat terjadi karena ketidakmampuannya untuk mencocokkan dengan benar domain tempat leksikon yang dibangun atau mencoba melakukan tugas tertentu. Banyak GL memiliki kekurangan kata-kata yang terkait dengan topik atau domain tertentu.

Penelitian-penelitian terdahulu hanya menggunakan satu metode saja untuk melakukan analisis sentimen pada media sosial di mana hasil klasifikasi emosi pengguna dirasakan masih belum optimal [5] [12] [13]. Selain itu, GL banyak digunakan pada penelitian terkait di mana sentimen diklasifikasikan menjadi 3 kelas, yaitu positif, negatif, dan netral [10] [14]. Untuk mengatasi kelemahan-kelemahan yang terdapat pada penelitian terdahulu, maka penelitian ini akan membangun suatu aplikasi yang menggabungkan metode *Naive Bayes* dan pendekatan *Lexicon-Based* untuk melakukan analisis sentimen pada media sosial Twitter dengan mengambil *tweets* berbahasa Indonesia sebagai *dataset*. Topik yang dipilih adalah kesehatan, teknologi dan pendidikan. Kesehatan dan teknologi merupakan dua topik besar yang sering dibicarakan oleh pengguna Twitter sedangkan pendidikan merupakan topik yang memiliki pengaruh signifikan terhadap cara berpikir dan tingkah laku seseorang dalam kehidupan sehari-hari. Penelitian ini juga akan menggunakan Eikman klasifikasi emosi yang terdiri dari senang, sedih, marah, kaget, takut, jijik [14] serta menambahkan emosi netral. Basis data yang digunakan adalah MongoDB yang memiliki kecepatan dalam mengeksekusi perintah *Insert*, *Update* dan *Delete Data* dibandingkan dengan *RDBMS (Relational Database Management System)* [5] [15]. Hasil klasifikasi diharapkan dapat ditampilkan pada tiga tingkatan yaitu kata, *tweets* dan topik.

Struktur penulisan dalam penelitian ini terdiri dari Pendahuluan, Tinjauan Pustaka, Metode Penelitian, Hasil dan Pembahasan, dan Kesimpulan dan Saran.

## II. TINJAUAN PUSTAKA

### A. Analisis Sentimen

Media sosial, seperti Twitter, merupakan *platform* di mana netizen saling berbagi informasi dan informasi dengan

bertambahnya pengguna Twitter dan jumlah postingan setiap harinya [1] [2] [3]. Twitter juga digunakan untuk membangun jaringan yang dapat menghubungkan orang di seluruh dunia tanpa dibatasi waktu dan tempat. Maka tidak mengherankan jika media sosial dianggap sebagai tambang emas untuk opini manusia yang menghasilkan peningkatan minat untuk melakukan penelitian mendalam terkait dengan analisis sentimen.

Analisis sentimen adalah deteksi komputasi dan studi opini, sentimen, emosi, dan subjektivitas dalam teks [16]. Analisis sentimen ini dapat diterapkan untuk mendeteksi dan menganalisis permasalahan sosial, ekonomi, politik dan sebagainya. Oleh karena itu, banyak perusahaan dan individu menggunakan teknik analisis sentimen untuk berbagai kepentingan, seperti ulasan produk dan layanan, perencanaan strategi politik, proyeksi finansial dan lain-lain.

Proses ekstraksi opini atau sentimen secara otomatis ini dapat dilakukan dengan menggunakan *machine learning* dan pendekatan *lexicon-based*. Penelitian ini akan menerapkan metode *Naive Bayes* yang merupakan metode *supervised-learning* klasifikasi dari *machine learning*. Selain itu, penelitian ini juga akan menggabungkan metode *Naive Bayes* dan pendekatan *lexicon-based* untuk menganalisis sentimen netizen pad Tweeter yang berkaitan dengan topik kesehatan, teknologi dan pendidikan di tiga level yang berbeda, yaitu *word-level*, *tweet-level*, dan *topic-level*.

### B. Metode *Naive Bayes*

*Naive Bayes* merupakan salah satu metode klasifikasi probabilistik yang berdasarkan penerapan teorema *Bayes* dengan asumsi independensi yang kuat (naif) antar fitur [17]. Langkah-langkah perhitungan metode ini adalah sebagai berikut [18]:

- 1) Menghitung jumlah kelas/label.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

Keterangan :

$P(c|x)$  = peluang kelas  $c$  terhadap kata  $x$  (*posterior probability*)

$P(x|c)$  = peluang kata  $x$  terhadap kelas  $c$  (*likelihood*)

$P(c)$  = peluang kelas  $c$  (*prior*)

$P(x)$  = peluang kemunculan kata  $x$  (*evidence*)

- 2) Menghitung jumlah kasus per kelas.

Pada persamaan (1), terdapat perhitungan *likelihood*, *prior*, dan *evidence*. Perhitungan nilai *likelihood* menggunakan model *Multinomial Naive Bayes*. Model ini dapat bekerja dengan cara menghitung nilai frekuensi kemunculan setiap kata pada sebuah dokumen. Setiap dokumen dapat disebut sebagai "*bag of words*" yang berarti dokumen memuat kata-kata yang akan diolah menggunakan distribusi *multinomial* [18].

Menurut rujukan [18], kelemahan perhitungan model multinomial ialah dimana adanya suatu kata yang tidak pernah muncul akan mengakibatkan perhitungan bernilai nol (*zero frequency problem*). *Laplace smoothing* dapat digunakan untuk memecahkan permasalahan tersebut dengan melakukan penambahan nilai 1 pada kata itu atau pada pembilang

sehingga dianggap pernah muncul sekali dan penambahan kata unik pada penyebut. Rumus *Laplace smoothing* dapat dilihat di bawah ini.

$$P(w_i|c) = \frac{\text{count}(w_i,c)+1}{\text{count}(c)+|V|} \quad (2)$$

Keterangan :

$P(w_i|c)$  = peluang kata ke- $i$  pada kelas  $c$   
 $\text{count}(w_i,c)$  = jumlah kata ke- $i$  pada kelas  $c$   
 $\text{count}(c)$  = jumlah seluruh kata pada kelas  $c$   
 $|V|$  = jumlah kata unik pada semua kelas

3) Menghitung *prior*.

$$P(c) = \frac{Nc}{N} \quad (3)$$

Keterangan :

$P(c)$  = peluang kelas  $c$   
 $Nc$  = jumlah dokumen kelas  $c$   
 $N$  = jumlah seluruh dokumen

4) Menghitung *evidence*.

$$P(x) = \frac{|x|}{S} \quad (5)$$

Keterangan:

$|x|$  = adalah jumlah kata  $x$  yang muncul  
 $S$  = jumlah keseluruhan kata yang muncul dalam seluruh dokumen.

### C. Pendekatan *Lexicon-based*

Pada pendekatan *lexicon-based*, sumber daya leksikal dieksploitasi untuk menetapkan skor polaritas pada kata-kata individual untuk mendeteksi sentimen keseluruhan dari suatu dokumen [4]. Pendekatan ini pada dasarnya bergantung pada leksikon sentimen yang telah ditentukan sebelumnya untuk menentukan polaritas sentimen umum dari dokumen tertentu. Itu juga menyimpulkan sentimen dokumen dengan mengeksploitasi properti statistik dokumen (dalam hal keberadaan kata) dan/atau memanfaatkan leksikon yang ada yang mengandung kata-kata terpolarisasi atau emosional. Meskipun leksikon yang ada berguna untuk berbagai tugas, mereka adalah sumber daya tetap yang perlu ditingkatkan ketika domain berubah.

Ada 2 metode untuk pendekatan ini, yaitu *dictionary based* dan *corpus-based*. Pendekatan *dictionary based* menggunakan metode pencarian kunci di mana mereka sangat bergantung pada leksikon emosi yang memiliki kumpulan kata dan polaritas sentimen yang sesuai [19]. Aung dan Myo memperluas pandangan ini dengan menguraikan langkah-langkah berikut dalam pendekatan ini [20]:

1. Satu set kecil kata sentimen (*seed lexicon*) dengan sentimen positif atau negatif yang diketahui pertama kali dikumpulkan secara manual.
2. Algoritme kemudian mengembangkan himpunan ini dengan mencari di WordNet atau kamus online lainnya untuk sinonim dan antonimnya.
3. Kata-kata yang baru ditemukan ditambahkan ke daftar *seed* dan iterasi berikutnya dimulai.

4. Proses berulang berakhir ketika tidak ada lagi kata-kata baru yang dapat ditemukan.

Menurut rujukan [4] [19], pendekatan *corpus-based* biasanya digunakan ketika kata kandidat baru dikenali berdasarkan hubungannya satu sama lain. Pendekatan ini juga mengeksploitasi pola kemunculan kata bersama yang ditemukan dalam kumpulan dokumen tekstual yang tidak terstruktur [21]. Seperangkat aturan linguistik dapat digunakan untuk mengidentifikasi kata sentimen dan orientasinya dari corpus, termasuk penggunaan konjungsi 'dan' untuk kata sifat siam yang biasanya memiliki orientasi yang sama (Mladenovic et al., 2016). Misalnya 'rumah ini indah dan luas', 'indah' termasuk sentimen positif, dapat disimpulkan bahwa 'luas' juga positif. Aturan lain dirancang untuk penghubung lain seperti 'atau', 'tetapi', 'baik-atau', dan 'tidak-tidak juga'. Aturan ini akan diterapkan pada dua kata sifat siam yang memiliki polaritas sentimen yang berlawanan. Studi ini mengadopsi dua aturan ini dengan menerapkan konjungsi 'dan' untuk menetapkan polaritas sentimen yang sama pada kata kandidat, 'tetapi' dan 'atau' untuk kata penghubung yang berlawanan.

### D. Penelitian Terkait

Kurniawan dkk dalam penelitiannya melakukan implementasi metode *Naive Bayes* dan *Lexicon Based Features* untuk menganalisis sentimen opini film. Ada beberapa proses yang dilakukan dalam penelitian ini, antara lain pra-pemrosesan teks, pembobotan term, pelatihan dan pengujian data menggunakan metode *Naive Bayes* dengan pembobotan fitur berbasis leksikon menggunakan kamus Barasa. Adapun hasil yang didapat untuk nilai akurasi, presisi, *recall*, dan *f-measure* adalah sama untuk semua hasil pengujian yakni 0,8. Sedangkan hasil yang didapat dari implementasi "... *Naive Bayes* tanpa menggunakan pembobotan *Lexicon Based Features* adalah 0,95, 1, 0,05, dan 0,9474 untuk akurasi, presisi, *recall*, dan *f-measure*" [18, p. 8335]. Dari hasil yang sudah dipaparkan, didapati bahwa integrasi metode *Naive Bayes* dan *Lexicon Based Features* sebelum memberikan hasil yang diharapkan.

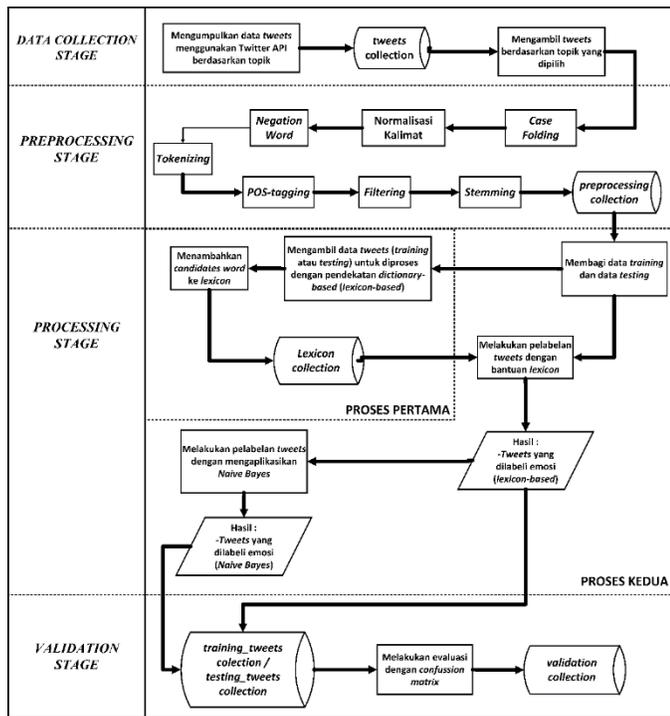
Penelitian berikutnya adalah analisis sentimen pengguna Twitter terhadap Sistem Administrasi Satu Atap (SAMSAT) kota Malang dengan menggunakan metode *Naive Bayes classifier* [13]. Ada sebanyak 859 *tweets* yang diambil dari November 2012 sampai Januari 2017 dengan kata kunci SAMSAT Kota Malang. *Tweets* ini dibagi 80% untuk data latih dan 20% untuk data uji. Pengujian dilakukan sebanyak dua kali dengan skenario yang berbeda-beda. Hasil akurasi dari pengujian skenario pertama menunjukkan bahwa didapati kelas positif (81%), kelas negatif (89%), dan kelas netral (80%). Pada pengujian skenario kedua, hasil akurasi kelas positif (82%) dan kelas negatif (92%) mengalami kenaikan dari skenario pertama. Tetapi untuk kelas netral tidak mengalami perbedaan.

Salah satu kelemahan dari penelitian terdahulu terletak pada penggunaan *dataset* yang cakupan yang lebih spesifik dan tidak cocok digunakan dalam penelitian yang memiliki pembahasan atau topik yang berbeda sehingga dapat mempengaruhi hasil akurasi. Selain itu, penelitian terkait hanya mengklasifikasi sentimen pengguna menjadi 3 kelas, yakni kelas positif, kelas

netral, dan kelas negatif. Sedangkan penelitian ini mengambil *dataset* yang berisi *tweets* berbahasa Indonesia terkait kesehatan, teknologi, dan pendidikan dengan menggunakan *Stream Twitter API (Application Programming Interface)*. Selanjutnya, penelitian ini akan menggabungkan metode *Naive Bayes* dan pendekatan *lexicon-based* untuk melakukan pelabelan pada data *training*. Penggunaan metode pendekatan *lexicon-based* dapat mengurangi pelabelan yang dilakukan secara personal karena pelabelan dilakukan secara otomatis oleh sistem dengan mengacu pada isi dari *lexicon* sedangkan metode *Naive Bayes* dapat mempercepat waktu pemrosesan karena alur perhitungan *Naive Bayes* tidak panjang sehingga dengan menggabungkan kedua metode ini diharapkan dapat meningkatkan hasil akurasi.

### III. METODE PENELITIAN

Penelitian ini dibagi menjadi 4 tahapan yang terdiri dari *data collection*, *pre-processing*, *processing*, dan *validation* seperti yang diperlihatkan di Gambar 1.



Gambar 1. Kerangka Penelitian

#### A. Tahap I: Data Collection

*Dataset* yang digunakan dalam penelitian ini berupa *tweets* yang ditarik menggunakan *Stream Twitter API* berdasarkan kata kunci dari setiap topik. *Tweets* yang diambil disimpan di *tweets collection*. Jumlah data untuk tiap topik ada sebanyak 1,000 *tweets*. Semua data ini akan disimpan ke dalam basis data MongoDB.

#### B. Tahap II: Pre-processing

Pada tahapan *preprocessing* akan dilakukan penerapan dari tujuh teknik *preprocessing* yang dimulai dari *case folding*, normalisasi kalimat, *negation word*, *tokenizing*, *POS-tagging*,

*filtering*, dan *stemming*. Hasil penerapan dari tujuh teknik *preprocessing* kemudian disimpan ke dalam *pre-processing collection*.

#### C. Tahap III: Processing

*Dataset* yang diambil dari *pre-processing collection* dibagi menjadi 70% data latih dan 30% data uji untuk setiap topik. Jumlah *seed lexicon* yang disediakan ada sebanyak 353 untuk ketujuh sentimen.

Tabel 1. Jumlah *Seed Lexicon*

Emosi	Jumlah Kata
Senang	83
Sedih	50
Marah	27
Kaget	27
Takut	64
Jijik	36
Netral	66
Total:	353

Ketujuh sentimen juga memiliki intensitas untuk menentukan tingkatan dari emosi apabila nantinya terdapat 2 emosi yang memiliki kemunculan tertinggi.

Tabel 2. Intensitas Sentimen

Emosi	Jumlah Kata
Senang	7
Sedih	6
Marah	5
Kaget	4
Takut	3
Jijik	2
Netral	1

Untuk menghitung skor sentimen keseluruhan dari *tweet* atau topik, skor sentimen individu dari kata-kata yang muncul di *tweet* digabungkan menjadi satu angka yang paling baik memperkirakan sentimen. Dalam penelitian ini, skor sentimen digunakan untuk menekankan kekuatan/intensitas sentimen untuk masing-masing sentimen serta untuk mendapatkan sentimen dominan pada *topic-level*. Setelah dominan sentimen ditentukan, skor dapat diabaikan untuk penggunaan. Namun, untuk memperbarui sentimen untuk penilaian khusus domain lainnya, seseorang perlu menyimpan skor itu secara internal. Sebagai catatan, penilaian tidak diperlukan untuk pengguna akhir. Namun, penyimpanan skor internal diperlukan untuk pembaruan skor sentimen selanjutnya.

Berikut ini merupakan penjelasan penerapan metode *Naive Bayes* dan pendekatan *lexicon-based* seperti yang diilustrasikan di Gambar 1.

- 1) Proses pertama menambahkan *candidate words* ke *lexicon*
  - a. Mengambil data *tweets* untuk *training* (500 data).
  - b. Mengaplikasikan *dictionary-based* dan *corpus-based* dari pendekatan *lexicon-based*.
  - c. Menambahkan *candidate words* (CW) ke *lexicon*.
- 2) Proses kedua melakukan pelabelan *tweets*
  - a. Mengambil data *tweets* untuk *training* (700 data).

- b. Membagi *tweets* yang akan dilakukan pelabelan *tweets* dengan *lexicon-based* dan *Naive Bayes*. *Lexicon-based* sebanyak 500 data dan *Naive bayes* sebanyak 200 data.
- c. Melakukan pelabelan *tweets* dengan bantuan *lexicon*.
- d. Mengaplikasikan *Naive Bayes* untuk melakukan pelabelan *tweets* dengan bantuan hasil pelabelan *tweets* (*lexicon-based*).
- e. Hasil pelabelan *tweets* dengan *lexicon-based* dan *Naive Bayes* disimpan pada *collection training*.

D. Tahap IV: *Validation*

*Confussion matrix* digunakan untuk menghitung *accuracy*, *precision*, dan *recall*. Hasil dari tahap *validation* akan disimpan di *validation collection*.

**IV. HASIL DAN PEMBAHASAN**

Tahap 1, *data collection*, dan tahap 2, *pre-processing*, untuk membangun aplikasi pengklasifikasian sentimen netizen pada Twitter sudah dipaparkan di bagian Metode Penelitian. Pada bagian ini akan dijelaskan pembangunan aplikasi tersebut untuk tahap 3, *processing*, dan tahap 4, *validation*.

A. Daftar Spesifikasi Persyaratan

Spesifikasi persyaratan untuk membangun aplikasi pengklasifikasian ini adalah sebagai berikut.

1. Aplikasi dapat mengumpulkan data *tweets* berdasarkan topik dengan menggunakan *Stream Twitter API* dan menyimpannya di *collection tweets* pada basis data MongoDB.
2. Aplikasi dapat menampilkan data *tweets* yang disimpan di *collection tweets*.
3. Aplikasi dapat melakukan *pre-processing* terhadap data *tweets* berdasarkan topik yang dipilih lalu menyimpannya di *collection preprocessing*.
4. Aplikasi dapat menampilkan hasil *pre-processing* yang disimpan di *collection preprocessing*.
5. Aplikasi dapat melakukan *processing* untuk mencari *candidate words* dengan metode pendekatan *lexicon-based* yaitu pendekatan *dictionary-based*.
6. Aplikasi dapat menampilkan hasil *CW* dan dapat menyimpannya di *collection lexicon*.
7. Aplikasi dapat melakukan *processing* untuk mengklasifikasi data *tweets* berdasarkan topik yang sudah dibagi dalam data *training* (70% data) dan data *testing* (30% data) ke tujuh kelas sentimen (senang, sedih, marah, kaget, takut, jijik dan netral) dengan menghitung jumlah kemunculan emosi data *tweets* serta dapat mengklasifikasi data *tweets* dengan metode *Naive Bayes*.
8. Aplikasi juga dapat menyimpan hasil klasifikasi di *collection 'training'* untuk data *training* dan *collection 'testing'* untuk data *testing*.
9. Aplikasi dapat menampilkan hasil klasifikasi dari data *training* maupun data *testing* yang disimpan di *collection 'training'* atau *collection 'testing'*.
10. Aplikasi dapat melakukan *validation* dengan perhitungan *confussion matrix* dan disimpan di *collection 'validation'*.

11. Aplikasi dapat menampilkan hasil *validation* yang disimpan di *collection 'validation'*.

B. Lingkungan Implementasi

Berikut ini adalah daftar perangkat keras dan perangkat lunak yang digunakan untuk membangun aplikasi pengklasifikasian sentimen netizen Twitter berbahasa Indonesia.

Tabel 3. Daftar Perangkat Keras dan Perangkat Lunak

I. Perangkat Keras	
Spesifikasi <i>Laptop</i>	CPU Ryzen 5-2500U, SSD 128 GB, RAM 8 GB, Windows 10
II. Perangkat Lunak	
1. Python 3.8	Bahasa pemrograman
2. Pustaka Python	Penunjang dalam membangun aplikasi.
3. Robo 3T 1.4.4	DBMS
4. Mozilla Firefox 77.0.1	Menampilkan aplikasi yang akan dibuat
5. PyCharm	<i>Text-editor</i> dalam membangun aplikasi

C. Tahap III: *Processing*

Tahap ini akan menjelaskan aturan implementasi dan alur kerja dari metode yang digunakan, contoh kasus, basis data aplikasi, dan tampilan antarmuka aplikasi.

1) Aturan Implementasi dan Alur Kerja

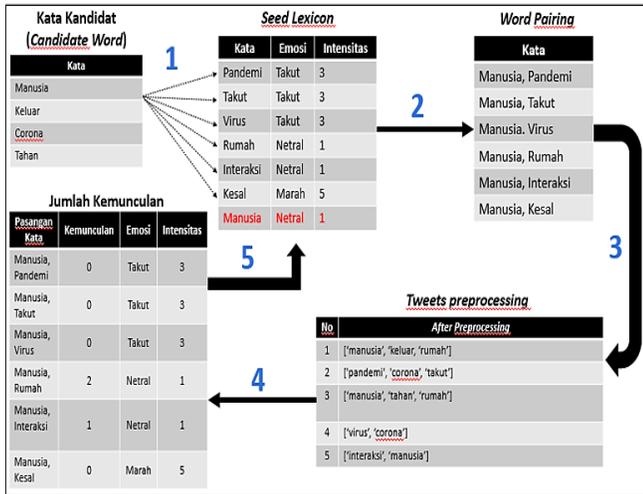
Ada beberapa aturan implementasi yang harus diperhatikan pada tahap *processing ini*. Tiga aturan untuk penerapan pendekatan *dictionary-based* di *word-level* pada penelitian ini adalah:

1. Jika kata yang ada di *CW* terdapat pada *Seed Lexicon*, maka kata tersebut akan dihapus dari *CW*. Kemudian, dilakukan pencarian pada kata berikutnya pada *CW*.
2. Jika semua pasangan kata yang dicari tidak didapati, maka kata tersebut dihapus dari *CW* dan tidak ditambahkan pada *Seed Lexicon*.
3. Jika terdapat 2 pasangan kata yang jumlah kemunculannya terbanyak dari semua kata dari sebuah *tweet* dan dalam jumlah yang sama, maka pasangan kata yang diambil adalah pasangan kata yang emosinya memiliki intensitas yang lebih tinggi. Intensitas sentimen dapat dilihat di Tabel 2.

Berdasarkan aturan yang dibuat, berikut ini adalah alur kerja pendekatan *dictionary-based* di *word-level* yang ada pada Gambar 2.

1. Kata-kata yang ada pada *CW* adalah kata-kata yang sudah melewati tahap *pre-processing*. Contoh, *CW* yang akan dicari adalah 'Manusia'. Kata 'Manusia' ini akan dicek keberadaan di *Seed Lexicon*. Jika kata ini ada di *Seed Lexicon*, maka akan diterapkan aturan pertama.
2. Kata yang dicari kemudian dipasangkan dengan kata yang ada di *Seed Lexicon*.

- Menghitung berapa kemunculan pasangan kata (*Word Pairing*) dengan *tweet* yang sudah melewati tahapan *pre-processing*.
- Mencari jumlah kemunculan tertinggi pasangan kata di *tweet* yang sudah melewati tahap *pre-processing*. Jika pasangan kata yang dicari tidak memiliki kemunculan, maka akan diterapkan aturan kedua. Selain itu, jika terdapat 2 kemunculan tertinggi namun memiliki emosi yang berbeda, maka diterapkan aturan ketiga.
- Menambahkan *CW* dan emosi dari pasangan katanya ke dalam *Seed Lexicon*.



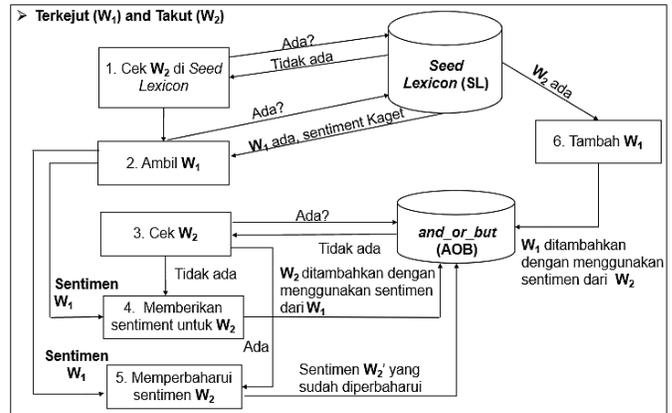
Gambar 2. Penerapan Pendekatan Dictionary-based

Aturan untuk dua CW yang menggunakan kata penghubung *dan* atau *tetapi* di *tweet-level*:

- Kata penghubung 'dan'**  
Jika salah satu kata sudah ada di *Seed Lexicon*, maka secara otomatis akan menetapkan sentimen dan skor kata tersebut ke CW. Misalnya, kata *kagum* dan *gembira* dimana *gembira* sudah ada di *Seed Lexicon* dan mempunyai sentimen *senang*. Oleh karena kata penghubung *dan* yang digunakan, kata *kagum* (CW) akan secara otomatis diberikan ke sentimen dan skor yang sama dengan kata *gembira*.
- Kata penghubung 'tetapi'**  
Ketika dua kata digabungkan menggunakan *tetapi*, sentimen dan skor kandidat akan mengambil sentimen yang berlawanan dari kata yang ada. Misalnya, kata *bagus* tetapi *mahal* dan kata *bagus* sudah ada dalam leksikon benih dan termasuk dalam polaritas kebahagiaan. Oleh karena itu, kata *mahal* akan memiliki sentimen *sedih*. Proses adopsi leksikon untuk konjungsi *atau* dan *tetapi* mirip dengan proses *dan*.

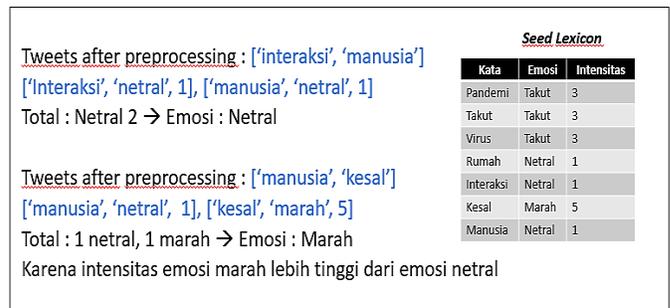
Berikut ini adalah alur kerja dari pendekatan *corpus-based*. Gambar 3 mendemonstrasikan proses adopsi leksikon untuk kata-kata siam menggunakan 'dan'. Poin penting yang perlu diperhatikan di sini adalah bahwa kata penghubung yang menggunakan 'atau' atau 'tetapi' memiliki proses yang sama dengan kata penghubung yang menggunakan 'dan'. Satu-satunya perbedaan adalah sentimen yang ditetapkan untuk kata-

kata. Kata-kata penghubung yang menggunakan 'dan' akan menetapkan sentimen yang sama untuk kata lain secara berpasangan. Kata-kata siam menggunakan 'atau' atau 'tetapi', di sisi lain, akan memberikan polaritas yang berlawanan untuk kata lain secara berpasangan.



Gambar 3. Penerapan Pendekatan Corpus-based

Pada *tweet-level*, setiap *tweet* akan dihitung berapa banyak kata yang mengandung sentimen didalamnya dan jumlah kemunculannya. Sebagai contoh dapat dilihat pada Gambar 4.



Gambar 4. Penerapan Intensitas Sentimen untuk Pendekatan Corpus-based

## 2) Contoh Kasus

Penerapan metode *Naive Bayes* mengikuti langkah-langkah yang sudah dijelaskan di bagian Tinjauan Pustaka. Contoh kasus terdapat 5 buah data yang sudah dilabeli sentimen dengan menggunakan pendekatan *lexicon-based*.

Tabel 4. Hasil Pelabelan Menggunakan *Lexicon-based*

Tweets after pre-processing	Emosi
['manusia', 'keluar', 'rumah']	Netral
['pandemi', 'corona', 'takut']	Takut
['manusia', 'tahan', 'rumah']	Netral
['virus', 'corona']	Takut
['interaksi', 'manusia']	Netral

Data *tweets* untuk penerapan metode *Naive Bayes*:

- Tweets 1:** ['krisis', 'ekonomi', 'corona']
- Tweets 2:** ['masyarakat', 'tidakbetah', 'rumah']

Berikut ini adalah tahapan dari metode *Naive Bayes*:

1. Jumlah *tweets* sebanyak 5, terdiri dari 3 *tweets* yang dilabeli netral dan 2 *tweets* yang dilabeli takut.
2. Menghitung peluang kelas sentimen (label sentimen) dengan rumus: “jumlah *tweets* pada setiap kelas sentimen / jumlah seluruh *tweets*”, sehingga hasil yang didapatkan yaitu *Netral* 3/5, dan *Takut* 2/5.
3. Menghitung jumlah kata pada setiap kelas sentimen, hasilnya kelas emosi netral memiliki jumlah kata sebanyak 8 dan kelas emosi takut memiliki jumlah kata sebanyak 5
4. Menghitung total kata unik pada 5 *tweets* yang di atas. Rincian kata-katanya adalah sebagai berikut: {'tahan', 'manusia', 'keluar', 'interaksi', 'virus', 'pandemi', 'corona', 'takut', 'rumah'}. Total kata unik adalah 9.
5. Menghitung peluang setiap kata pada *tweets* yang belum diberi label sentimen terhadap setiap kelas.

Tabel 5. Menghitung Peluang Setiap Kata (Tweets 1)

Tweets 1 : ['krisis', 'ekonomi', 'corona']	
Peluang kata “krisis” pada kelas netral	$= \frac{0+1}{8+9} = 0.0588$
Peluang kata “krisis” pada kelas sedih	$= \frac{0+1}{5+9} = 0.0714$
Peluang kata “ekonomi” pada kelas netral	$= \frac{0+1}{8+9} = 0.0588$
Peluang kata “ekonomi” pada kelas sedih	$= \frac{0+1}{5+9} = 0.0714$
Peluang kata “corona” pada kelas netral	$= \frac{0+1}{8+9} = 0.0588$
Peluang kata “corona” pada kelas sedih	$= \frac{2+1}{5+9} = 0.2142$

Tabel 6. Menghitung Peluang Setiap Kata (Tweets 2)

Tweets 2 : ['masyarakat', 'tidakbetah', 'rumah']	
Peluang kata “masyarakat” pada kelas netral	$= \frac{0+1}{8+9} = 0.0588$
Peluang kata “masyarakat” pada kelas sedih	$= \frac{0+1}{5+9} = 0.0714$
Peluang kata “tidakbetah” pada kelas netral	$= \frac{0+1}{8+9} = 0.0588$
Peluang kata “tidakbetah” pada kelas sedih	$= \frac{0+1}{5+9} = 0.0714$
Peluang kata “rumah” pada kelas netral	$= \frac{2+1}{8+9} = 0.1764$
Peluang kata “rumah” pada kelas sedih	$= \frac{0+1}{5+9} = 0.0714$

6. Hitung probabilitas peluang kelas terhadap kata dengan rumus peluang kelas \* peluang setiap kata pada *tweets* terhadap kelas.

Tabel 7. Menghitung Peluang Kelas (Tweets 1)

Kelas Netral	Kelas Takut
$3/5 * 0.0588 * 0.588 *$	$2/5 * 0.0714 * 0.0714 *$
$0.0588 = 0.00012$	$0.2142 = 0.00043$

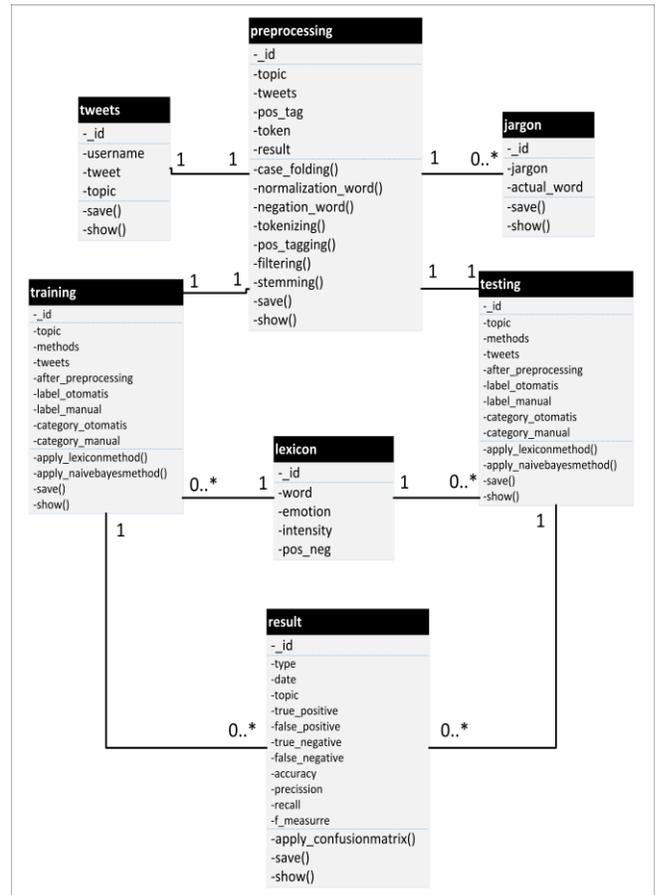
Tabel 8. Menghitung Peluang Kelas (Tweets 2)

Kelas Netral	Kelas Takut
$3/5 * 0.0588 * 0.0588 *$	$2/5 * 0.0714 * 0.0714 *$
$0.1764 = 0.00036$	$0.0714 = 0.00014$

7. Nilai maksimum pada hasil peluang kelas terhadap kata disetiap *tweets* akan menjadi label emosi untuk *tweets* tersebut. Nilai maksimum untuk *tweets* 1 ada pada kelas takut dan nilai maksimum untuk *tweets* 2 ada pada kelas netral.

### 3) Basis Data Aplikasi Pengklasifikasian

Gambar 5 mengilustrasikan implementasi basis data untuk penelitian ini.



Gambar 5. Basis Data Aplikasi Pengklasifikasian

1. *Collection preprocessing* berisi *tweets* yang sudah melewati tahapan *preprocessing*.
2. *Collection lexicon* berisi kata-kata beserta pasangan emosi yang akan digunakan pada tahapan *processing* untuk metode *lexicon-based*.
3. *Collection training* berisi *tweets* yang akan digunakan untuk *training*.
4. *Collection testing* berisi *tweets* yang akan digunakan untuk *testing*.
5. *Collection validation* berisi hasil validasi untuk *tweets-level* berdasarkan data *training* atau data *testing* dan topik.

### 4) Implementasi Tampilan Antarmuka Aplikasi

Berikut ini adalah beberapa tampilan antarmuka aplikasi.



Gambar 6. Pembagian Data Latih dan Data Uji

Pengguna aplikasi harus memilih jumlah data latih dan data uji berdasarkan topik.



Gambar 7. Pencarian CW Baru

Gambar 7 menampilkan jumlah tweets per topik dan pengguna aplikasi diminta untuk mengisi jumlah data latih atau data uji berdasarkan topik. Kemudian akan dilakukan pencarian CW menggunakan pendekatan *dictionary-based* pada *word-level*. Selain itu, pengguna dapat melihat daftar leksikon yang sudah tersimpan dalam basis data.

Gambar 8 mendemonstrasikan hasil klasifikasi berdasarkan topik *Pendidikan* dengan menggunakan 7 kelas sentimen. Pada gambar ini juga ditunjukkan tabel yang berisi tweets, hasil *pre-processing*, label manual yang diisi oleh pengguna, dan label otomatis yang diberikan oleh algoritma.

#### D. Tahap IV: Validation

Validasi dilakukan dengan menggunakan *confusion matrix* untuk menghitung *accuracy*, *precision* and *recall* dari data tweets tersebut. Pada penelitian ini terdapat tujuh sentimen sebagai hasil klasifikasi yang kemudian akan digabungkan menjadi 2 kelas sentimen utama, yaitu positif dan negatif. Sentimen senang, kaget dan netral akan menjadi kelas positif sedangkan sentimen sedih, marah, takut, dan jijik akan menjadi kelas negatif.

##### 1) Word-level

Hasil validasi pada *word-level* diambil dari *lexicon* yang awalnya berjumlah 417 kata dan setelah diperluas menjadi 4,772 kata. *Confusion matrix* digunakan untuk menghitung jumlah dari nilai yang diprediksi dan nilai aktual. Aplikasi ini mempunyai 79,17% akurasi, 89,5% presisi, dan 87,5% *recall*.

Dari hasil pengujian ini didapati bahwa jumlah kata yang diprediksi sebagai kelas positif adalah sama dengan jumlah kata dari hasil aktual kelas positif (*true positive* =3,340) lebih banyak daripada jumlah kata yang diprediksi sebagai kelas positif tetapi hasil aktual menunjukkan kumpulan kata tersebut dari kelas negatif (*false positive* =392). Sementara, jumlah kata yang diprediksi sebagai kelas negatif dan jumlah data dari hasil aktualnya yang juga kelas negative (*true negative*=438) lebih sedikit dibanding dengan jumlah kata yang salah diprediksi sebagai kelas negatif tetapi hasil aktualnya adalah kelas positif (*false negative*=602).

	Senang	Sedih	Marah	Kaget	Takut	Jijik	Netral
Lexicon Based	2	0	0	0	0	0	4
Naive Bayes	0	0	0	0	0	0	4

Tweets	Hasil Pre-prosesan	Label Manual	Label Otomatis
ada maksudnya aing semakin yakin kuliah online sampe desainer	['kuliah']	netral	netral
ada video call sama dosening buat bimbingan skripsi setelah hampir bulan email gak dibatas akhirnya vidcall juga semangattt ya kalian yang kesulitan kuliah online semua ada jalannya	['Paba', 'video', 'call', 'doubling', 'bimbingan', 'skripsi', 'email', 'tidakdibatas', 'vidcall', 'semangattt', 'sult', 'kuliah', 'jalan']	netral	senang
absen ae dikorupsi pomeneh duwit kuliah online	['ae', 'korupsi', 'pomeneh', 'duwit', 'kuliah online']	sedih	netral
ad g kmpus yg bta kuliah a online kyknya prnh baca	['kuliah', 'baca']	netral	netral
ad kuliah online kk	[]	netral	netral
ada aturan kampus soal bimbingan online ada aturan kampus soal kelanjutan penelitian selama covid nggak ada kan kuliah atasannya kalian akan diuruk masuk kuliah wala covid belum mereda dan harus menyelesaikan pi skripsi dim waktu bulan terbit sult bukan	['atur', 'kampus', 'bimbing', 'atur', 'kampus', 'telit', 'covid', 'alas', 'kuliah', 'covid', 'veda', 'selesa', 'pi', 'skripsi', 'simo', 'waktu', 'sult']	netral	senang
ada dosen yg pas kuliah online klo gaada yg tanya gaada yg kasih feedback trn kng jdi gimana minustma dftin apa yg kamu dapat sejauh mana tugasmu jdi yg maunya aku rebahan jdi lrgg diuruk sambil mikiiiiir	['kuliah', 'gaada', 'gaada', 'kasih', 'jdi', 'gimana', 'minustma', 'dftin', 'kamu', 'tugas', 'jdi', 'jdi', 'diuruk', 'mikiirir']	netral	netral

Gambar 8. Hasil Pengklasifikasian Sentimen Sesuai Topik

##### 2) Tweet-level

Pada level ini, dilakukan pengujian terhadap 700 data latih dan 300 data uji yang hasilnya dapat dilihat pada Tabel 8.

Tabel 9. Hasil Pengujian pada Tweet Level untuk Tiap Topik

Topik	Accuracy	Precision	Recall
<b>I. Training</b>			
Pendidikan	81.86%	80.14%	97.70%
Kesehatan	72.57%	87.30%	77.87%
Teknologi	73.00%	73.44%	96.11%

Topik	Accuracy	Precision	Recall
<b>II. Testing</b>			
Pendidikan	83.00%	83.62%	98.36%
Kesehatan	76.67%	84.94%	85.65%
Teknologi	69.00%	71.01%	93.78%

Rasio data latih dan data uji yang digunakan dalam pengujian di Tabel 8 adalah 70:30. Dari hasil validasi *testing* didapati bahwa tingkat akurasi tertinggi terdapat pada topik *pendidikan* yang diambil (83%) sedangkan topik *teknologi* (69%) memiliki nilai akurasi terkecil.

##### 3) Topic-level

Hasil validasi untuk *topic-level* diambil dari label otomatis yang paling banyak didapatkan. Ketiga topik tersebut

mempunyai sentimen senang, seperti yang ditunjukkan pada Tabel 10.

Tabel 10. Hasil Validasi pada Topic-level

Topik	Label Otomatis
Pendidikan	Senang
Kesehatan	Senang
Teknologi	Senang

Berdasarkan pengujian yang telah dilakukan, didapati bahwa hasil akurasi untuk *word-level* hanya sebesar 79.17% dikarenakan jumlah kata yang diprediksi sebagai kelas negatif lebih besar dibandingkan jumlah kata dari hasil aktual yang berkelas negatif. Penelitian ini juga melakukan pengujian pada *tweets-level* sebanyak tiga kali dengan rasio data latih dan data uji yang berbeda untuk ketiga topik, yaitu: a) 60:40, b) 70:30, dan c) 80:20.

Tabel 11. Perbandingan Data Latih dan Data Uji

Rasio	Accuracy	Precision	Recall
<b>I. Training</b>			
60:40	79.33%	77.34%	97.49%
70:30	81.86%	80.14%	97.70%
80:20	79.75%	79.35%	96.07%
<b>II. Testing</b>			
60:40	82.50%	83.25%	98.15%
70:30	83.00%	83.62%	98.36%
80:20	80.00%	83.24%	94.48%

Dari hasil pengujian di atas dengan menggunakan rasio perbandingan data latih dan data uji yang berbeda, didapati bahwa rasio 70:30 (*testing*) memberikan hasil terbaik untuk akurasi, presisi dan *recall* sebanyak 83%, 83,62%, dan 98,36% terkait topik *pendidikan* (lihat Tabel 8). Akan tetapi, topik *teknologi* mempunyai tingkat akurasi (69%), presisi (71%), dan *recall* (93,78) terendah dari ketiga topik. Selanjutnya, pengujian terhadap topik yang berbeda mempengaruhi hasil akurasi untuk topik *kesehatan* dan *teknologi* karena *lexicon* yang digunakan sebagai acuan untuk melakukan proses dengan pendekatan *lexicon-based* sudah diperluas dari topik *pendidikan*. Validasi terhadap *topic-level* juga didominasi oleh sentimen *senang* pada tiga topik tersebut.

## V. KESIMPULAN DAN SARAN

### A. Kesimpulan

Kesimpulan dari penelitian ini adalah sebagai berikut:

1. Aplikasi berhasil menerapkan proses klasifikasi sentimen netizen pada Twitter dengan menggunakan metode *Naive Bayes* dan pendekatan *lexicon-based*.
2. Teknik-teknik *pre-processing* yang diaplikasikan dapat mempengaruhi hasil *processing*.
3. Semakin banyak kata unik pada *seed lexicon* awal semakin baik pula hasil *processing* yang diterapkan dengan menggunakan pendekatan *lexicon-based*.
4. Aplikasi bisa memperluas *lexicon* dari tiga topik yang terdiri dari pendidikan, kesehatan, dan teknologi.

5. Aplikasi bisa menampilkan hasil validasi pada 3 level yang berbeda, yakni *word-level*, *tweets-level* dan *topic-level*.
6. Penggunaan *dataset* dari topik yang berbeda memiliki hasil akurasi yang baik pada topik yang diproses pertama dibandingkan dengan topik setelahnya.
7. Pada pengujian 70:30, hasil akurasi pada *word-level* lebih baik dibandingkan dengan hasil akurasi pada *tweets-level* dikarenakan pada *word-level* hanya menentukan emosi pada setiap kata bukan kalimat (*tweets*).

### B. Saran

Saran untuk pengembangan dari penelitian ini adalah :

1. *Tweets* yang digunakan sebaiknya disaring kembali untuk menghapus *tweets* yang memiliki duplikasi meskipun terdapat fitur untuk menyaring *tweets* yang di *re-tweets* oleh pengguna lain.
2. Lebih banyak mendaftarkan kata-kata yang termasuk kedalam *stopwords* dan kata-kata singkatan pada tahap *preprocessing* agar data dapat lebih bersih untuk diproses.
3. Menghapus kata bahasa Inggris dengan bantuan daftar kata bahasa Inggris.
4. Menggunakan dua penerapan *lexicon-based* yaitu *dictionary-based* dan *corpus-based* untuk lebih meningkatkan hasil akurasi.
5. Menambahkan jumlah kata awal pada *lexicon* untuk meningkatkan hasil akurasi.

## UCAPAN TERIMA KASIH

Penulis ingin menyampaikan terima kasih kepada Universitas Katolik De La Salle Manado, terutama Lembaga Penelitian dan Pengabdian Masyarakat, yang terus memotivasi dan mendukung proyek penelitian ini dari awal hingga akhir.

## DAFTAR PUSTAKA

- [1] Statista Research Department, 19 April 2022, Statista, <https://www.statista.com/statistics/273172/twitter-accounts-with-the-most-followers-worldwide/>, diakses tgl 6 Juni 2022.
- [2] M. Ahlgren, 22 Mei 2022, Websiterating, <https://www.websiterating.com/research/twitter-statistics/#chapter-1>, diakses tgl 6 Juni 2022.
- [3] Kementerian Komunikasi dan Informatika, 12 April 2022, Indonesia Peringkat Lima Pengguna Twitter, [https://kominfo.go.id/content/detail/2366/%20indonesia-peringkat-lima-penggunatwitter/0/sorotan\\_media](https://kominfo.go.id/content/detail/2366/%20indonesia-peringkat-lima-penggunatwitter/0/sorotan_media), diakses tgl 6 Juni 2022.
- [4] M. G. Giatsoglou, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas. Sentiment Analysis Leveraging Emotions And Word Embeddings. *Expert Systems With Applications*, vol. 69, 214-224, 2017.

- [5] L. Wikarsa, M. S. Kim. Automatic Generation Of Word-Emotion Lexicon For Multiple Sentiment Polarities On Social Media Texts. *ICIC Express Letters*, vol.13, no. 4, 317-324, 2019.
- [6] A. P. Wibawa, A. C. Kurniawan, D. M. Murti, R. P. Adiperkasa, S. M. Putra, S. A. Kurniawan, Y. R. Nugraha. Naïve Bayes Classifier for Journal Quartile Classification. *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, vol.7, no.2 91-97, 2019.
- [7] T. Rosandy. Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4.5) Untuk Menganalisa Kelancaran Pembiayaan (Studi Kasus: KSPPS / BMT AL-FADHILA). *Jurnal TIM Darmajaya*, vol.2, no.1, 52-62, 2016.
- [8] S. Deng, A. P. Sinha, and H. Zhao. Adapting Sentiment Lexicons to Domain-specific Social Media Texts. *Decision Support Systems*, vol.94, 65-76, 2017.
- [9] M. Mladenovi, J. Mitrovi, C. Krstev, and D. Vitas. Hybrid sentiment Analysis Framework for a Morphologically Rich Language. *Journal of Intelligent Information Systems*, vol.46, no.3 599-620, 2016.
- [10] E. Kim, and R. Klinger. *A Survey on Sentiment and Emotion Analysis for Computational Literary Studies*. New York: Cornell University, 2018.
- [11] F. B. Marquez, E. Frank, and B. Pfahringer. Building a Twitter Opinion Lexicon From Automatically-annotated Tweets. *Knowledge-Based Systems*, vol.108, 65-78, 2016.
- [12] V. Fitri, R. Andreswari, and M. A. Hasibuan. Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm, *The Fifth Information Systems International Conference 2019*, Bandung, 24 Juli.
- [13] I. F. Rozi, E. N. Hamdana, and M. B. Alfahmi. Pengembangan Aplikasi Analisis Sentimen Twitter Menggunakan Metode Naive Bayes Classifier (Studi Kasus SAMSAT Kota Malang). *Informatika Polinema*, vol.4, no.2 149-154, 2018.
- [14] H. Nida, K. Mahira, M. Mudasir, M. M. Ahmed, and M. Mohsin. Automatic Emotion Classifier, *Proceedings of International Conference on Advanced Computing and Intelligent Engineering (ICACIE) 2017*, Ajmer, India, 23-25 November.
- [15] MongoDB, 16 Oktober 2018, MongoDB, <https://www.mongodb.com/what-is-mongodb>, diakses tgl 6 Juni 2022.
- [16] F. A. Pozzi, E. Messina, and B. Liu. *Challenges of Sentiment Analysis in Social Networks*, San Fransisco: Morgan Kaufmann Publisher, 2017.
- [17] B. Purnama. *Pengantar Machine Learning*, Bandung: Informatika Bandung, 2019.
- [18] A. Kurniawan, Indriati, and S. Adinugroho. Analisis Sentimen Opini Film Menggunakan Metode Naïve Bayes dan Lexicon Based Features. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol.3, no.9, 8335-8342, 2019.
- [19] S. Goyal, and N. Tiwari. Emotion Recognition: A Literature Survey. *International Journal For Technological Research In Engineering*, No.9, vol. 4, no.9, 1502-1524, 2017.
- [20] K. Z. Aung, and N. N. Myo. Sentiment Analysis of Students' Comment Using Lexicon Based Approach, *IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS) 2017*, Wuhan, China, 24-26 Mei.
- [21] Q. Wang, Y. Jin, T. Yang, and S. Cheng. An Emotion-Based Independent Cascade Model For Sentiment Spreading. *Knowledge-Based Systems*, vol. 116, 86-93, 2017.