

Perbandingan Akurasi Metode Lexicon Based Dan Naive Bayes Classifier Pada Analisis Sentimen Pendapat Masyarakat Terhadap Aplikasi Investasi Pada Media Twitter

Fitrah Amaliah¹, I Kadek Dwi Nuryana²

^{1,2}Jurusan Teknik Informatika, Universitas Negeri Surabaya

fitrah.18007@mhs.unesa.ac.id

dwinuryana@unesa.ac.id

Abstrak — Investasi pada era globalisasi ini menjadi kegiatan yang penting dalam perekonomian dan bisnis. Sudah banyak masyarakat yang memilih untuk menempatkan dana yang mereka miliki dalam bentuk investasi. Dengan adanya perkembangan teknologi para *developer* membuat aplikasi investasi untuk memudahkan proses investasi. Dengan adanya aplikasi investasi, terdapat juga kelebihan dan kekurangan dari aplikasi yang ada, mulai dari aplikasi investasi bodong hingga aplikasi investasi yang terpercaya. Analisis sentimen pada *Twitter* dilakukan agar mengetahui aplikasi investasi yang harus dihindari dan dapat dipercaya. Metode *lexicon based* dan *naive bayes classifier* dipilih agar dapat mengklasifikasikan antara *tweets* yang bersentimen positif, netral, dan negatif agar memudahkan masyarakat dalam menentukan pilihan beserta mengetahui tingkat akurasi antara kedua metode. Dari hasil kedua metode telah didapatkan bahwa sentimen positif memiliki persentase tertinggi terhadap aplikasi investasi. Sedangkan perbandingan akurasi dari kedua metode menghasilkan 67% untuk metode *lexicon based* dan 78% untuk metode *naive bayes classifier*. Dari hasil tersebut dapat menunjukkan bahwa data yang telah dianalisis tentang aplikasi investasi memiliki nilai positif dan hasil akurasi dari metode *naive bayes classifier* memiliki nilai yang lebih tinggi dari pada metode *lexicon based*.

Kata Kunci — *Investasi, Aplikasi Investasi, Analisis Sentimen, Twitter, Lexicon Based, Naive Bayes Classifier*

I. PENDAHULUAN

Pada zaman ini, kegiatan berinvestasi menjadi aktivitas yang penting dalam bidang bisnis dan ekonomi. Penempatan sejumlah dana pada masa sekarang dengan harapan akan mendapatkan keuntungan pada masa yang akan datang disebut investasi [1]. Selain itu, didalam kegiatan investasi terdapat dua jenis yaitu investasi aktif dan investasi pasif, dimana investasi aktif merupakan kegiatan berinvestasi dengan kerjasama secara langsung dengan investor untuk pengelolaannya, sedangkan investasi pasif tidak melakukan kerjasama secara langsung untuk pengelolaannya melainkan akan dikelola oleh manajer investasi dan investasi akan bergerak naik turun dengan

sendirinya [2]. Untuk aplikasi investasi merupakan contoh investasi jenis pasif.

Di Indonesia dalam melakukan kegiatan investasi sudah terlaksana cukup lama, dari masa orde lama sudah disiapkan rancangan perundang-undangan tentang praktik penanaman modal asing di Indonesia [3]. Sebagai negara dalam tahap perkembangan, pemahaman akan kesadaran masyarakat Indonesia terhadap kegiatan investasi masih tergolong rendah. Sedangkan di negara yang sudah maju, masyarakatnya sudah lebih sadar terhadap hal yang berkaitan tentang investasi. Salah satu alasan dari rendahnya pemahaman dan kesadaran akan hal investasi di Indonesia adalah orientasi mayoritas tujuan masyarakat dalam menghadapi kondisi finansial yang masih dalam tujuan jangka pendek contohnya menabung di bank supaya memperoleh bunga dan *saving society* [4]. PT Bursa Efek Indonesia (BEI) telah memberikan pelajaran dan meluaskan industri ke depan yang lebih baik sebagai cara dalam meluaskan industri pasar modal di Indonesia. Tujuan BEI tidak hanya memperbesar kapasitas investor, melainkan juga untuk memasukkan kebutuhan berinvestasi di pasar modal yang secara tidak langsung dapat memperkuat persepsi akan pengetahuan untuk masyarakat Indonesia supaya berinvestasi aktif di pasar modal Indonesia.

Analisa sentimen disini dapat dimanfaatkan oleh masyarakat dalam mengetahui hal yang berkaitan dengan investasi. Analisa sentimen merupakan langkah yang tepat dalam mencari informasi mengenai opini-opini masyarakat dalam menilai aplikasi investasi terpercaya dari media sosial. *Twitter* merupakan salah satu dari banyaknya media sosial yang cukup populer dan banyak diminati oleh masyarakat Indonesia dalam menuangkan berbagai opini-opini, baik opini yang bersifat positif, netral, ataupun negatif. Menurut data Kominfo Indonesia, Tercatat terdapat 19,5 juta masyarakat indonesia yang menggunakan *Twitter* dan menjadikan Indonesia sebagai negara kelima terbesar dalam pemakaian *Twitter* dibawah Inggris dan negara lainnya. Dengan demikian, untuk mengetahui dan memproses data sehingga mengetahui informasi yang tersedia pada kalimat opini, *twitter* dijadikan platform

dalam penelitian ini. [5]. Pengambilan data dari *Twitter* pada penelitian ini menggunakan cara *crawling* data. Hasil dari cara *crawling* ini berupa kumpulan dokumen atau data yang dapat digunakan dalam *sentiment analysis*.

Text mining memiliki beberapa bagian diantaranya adalah *sentiment analysis* yang dalam bahasa indonesia *sentiment analysis* bisa disebut analisis sentimen [6]. Analisis sentimen akan membuat kategori dalam struktur label sentimen positif, netral, atau negatif. *Output* berupa hasil representasi secara visual dari pengumpulan data teks disebut *Wordcloud* [7]. Untuk mengklasifikasikan data dalam melakukan suatu analisis sentimen memerlukan suatu metode dengan tujuan memudahkan penggunaan membedakan antara opini yang termasuk klasifikasi positif, netral, atau negatif serta dapat mengetahui akurasi persentase dari setiap sentimen dan akurasi dari kinerja metode.

Pada penelitian analisis sentimen ini menggunakan metode *Lexicon Based* dan metode *Naive Bayes Classifier* (NBC) untuk mengetahui perbandingan kinerja kedua metode. *Vader sentiment* digunakan untuk metode analisis dari metode *Lexicon Based* yang berbasis *rule-based sentiment analysis*. *Naive Bayes Classifier* (NBC) merupakan salah satu jenis algoritma atau metode yang digunakan untuk pengklasifikasian dalam analisis sentimen dengan data yang besar dan menghasilkan tingkat akurasi yang tinggi serta cepat dan tepat [8]. Pada penelitian sebelumnya terdapat penelitian serupa dengan judul “Analisis Sentimen Terhadap Vaksin Covid-19 di Indonesia pada *Twitter* Menggunakan Metode *Lexicon Based*” akan tetapi dalam penelitian tersebut tidak memperhitungkan akurasi dari metode yang digunakan, sehingga masih belum bisa diketahui akurasi dari metode *Lexicon Based*. Kemudian juga terdapat penelitian dengan judul “Analisis Sentimen *Twitter* Bertema Pembelajaran Tatap Muka (PTM) Terbatas Menggunakan Metode *Lexicon Based*” dengan nilai akurasi sebesar 86,60% oleh Muhammmad Nuzul Edit pada tahun 2021. Pada penelitian tersebut untuk tahap *preprocessing* tidak terdapat proses yang cukup detail, seperti *filtering* dan *stemming*. Pada penelitian ini, peneliti akan menambahkan tahap tersebut untuk selanjutnya di proses untuk dibandingkan akurasinya menggunakan metode *Lexicon Based* dan *Naive Bayes Classifier*. Data yang diambil berupa data yang bersumber dari media sosial *Twitter* sebanyak 698 data *tweets* dengan *keyword* “aplikasi investasi” dari tanggal 11 Januari 2022 sampai 19 Januari 2022.

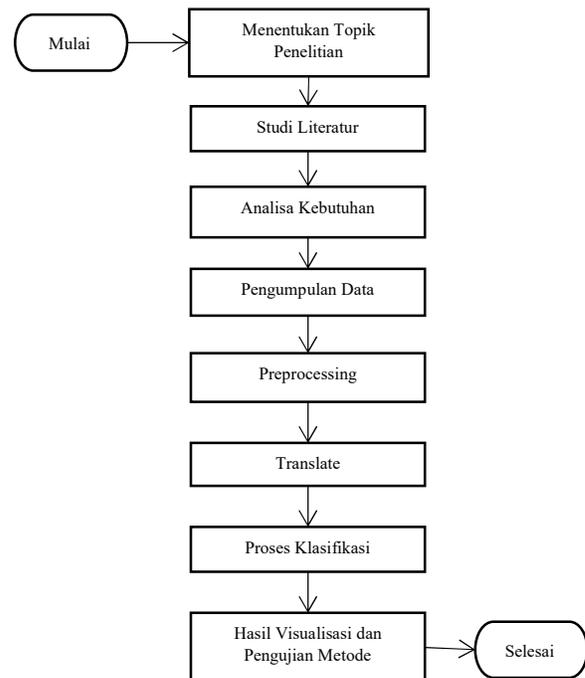
Lexicon Based digunakan dalam penelitian sebagai klasifikasi opini maupun pengklasifikasian [9]. Dokumen atau data yang berupa kata dalam sebuah kalimat akan dibandingkan secara langsung dengan kamus opini yang tersedia dalam *Lexicon* menjadi salah satu kelebihan dari metode *lexicon based*. Jika di dalam kalimat tersebut tersedia kata yang bersifat opini, maka kalimat tersebut akan disebut kalimat opini. Dan jika kata yang tersedia dalam kalimat tersebut tidak tersedia dalam kamus *Lexicon*, maka dianggap bukan kalimat opini merupakan

kekurangan dari metode ini. Sedangkan keunggulan dari algoritma *Naive Bayes Classifier* adalah proses perhitungan yang cepat dan efisien.

Berdasarkan perihal diatas, penulis akan lebih terfokus pada analisis sentimen *twitter* terhadap aplikasi investasi dengan metode *Lexicon Based* dan *Naive Bayes Classifier*. Dengan adanya penelitian ini diharapkan bisa membantu masyarakat mengenai aplikasi investasi yang dipercaya dan menambah kesadaran serta pemahaman akan investasi. Sedangkan untuk *developer* dari aplikasi investasi bisa lebih memperbaiki aplikasi dari *feedback* atau opini yang telah diutarakan masyarakat melalui *twitter* yang digunakan.

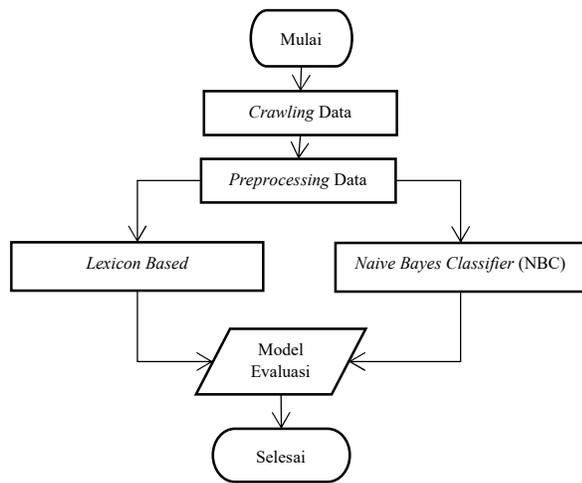
II. METODOLOGI

Metodologi pada suatu penelitian memiliki pedoman berupa alur penelitian atau langkah-langkah agar hasil yang diharapkan sesuai dengan tujuan awal. Metodologi penelitian memiliki rancangan alur yang terstruktur dan sesuai. Rancangan alur pada penelitian ini dapat dilihat pada Gbr. 1 berikut :



GBR. 1 DIAGRAM ALUR PENELITIAN

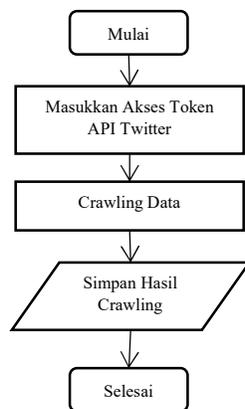
Penelitian ini akan membandingkan tingkat akurasi antara metode *lexicon based* dan *naive bayes classifier* dalam proses pengklasifikasian data yang telah dilakukan. Gbr. 2 berikut merupakan proses pengklasifikasian dalam penelitian ini.



GBR. 2 FLOWCHART KLASIFIKASI

A. Pengumpulan Data

Analisis sentimen menggunakan cara *crawling* dimana data set berasal dari *tweets* media sosial *Twitter* terkait “aplikasi investasi”. Untuk uraian dari proses *crawling* dapat dilihat pada Gbr. 3 berikut :



GBR. 3 FLOWCHART CRAWLING DATA

Proses *crawling* data *tweets* dari media sosial *Twitter* dari unggahan pengguna *Twitter* dengan menggunakan akses token API *Twitter* dengan kata kunci Aplikasi Investasi dilakukan menggunakan bahasa pemrograman *Python* yang selanjutnya disimpan dalam format *.csv*, dari data yang didapat akan dilanjutkan dengan tahap *preprocessing*.

B. Preprocessing

Pada tahap ini akan dilakukan penyeleksian data dan pembersihan data *tweets* yang telah diambil. Berikut merupakan tahapan-tahapan dari proses *preprocessing* :

1. Cleaning

Pada tahap *cleaning* akan dilakukan proses untuk penghilangan tanda baca dan karakter yang

tidak diperlukan seperti tanda titik, tanda koma, tanda tanya, tanda seru, menghapus HTML dan URL, menghapus *hashtag* dan mention, menghapus emoji, serta menghapus karakter yang tidak relevan.

2. Tokenization

Pada tahap *tokenization* ini akan memisahkan sebuah kalimat dari *tweets* menjadi potongan kata sebelum dianalisis lebih lanjut.

3. Case Folding

Dalam tahap *case folding* akan dilakukan perubahan terhadap kumpulan kalimat *tweets* menjadi huruf kecil semua.

4. Filtering atau Stopwords Removal

Tahap *filtering* akan melakukan proses dalam menghilangkan kata-kata yang tidak memiliki makna atau *stop words* agar terfokus pada kata-kata yang lebih bermakna. Dengan demikian, proses klasifikasi akan lebih cepat dan efisien karena jumlah kata yang diproses akan menjadi lebih sedikit.

5. Stemming

Pada tahap ini akan mengubah kata yang berimbuhan menjadi kata dasar.

C. Translate

Setelah tahap *preprocessing* dilakukan, tahap selanjutnya adalah menerjemahkan data *tweets* berbahasa indonesia menjadi *tweets* berbahasa inggris. Tahap *translate* dilakukan karena pada tahap selanjutnya akan menggunakan *library vader sentiment* yang menggunakan bahasa inggris.

D. Analisis Sentimen atau Pengklasifikasian

Setelah tahap *translate* dilakukan, dilanjutkan tahap pengklasifikasian data *tweets*. di tahap ini proses *labelling sentiment* dilakukan dan akan menghasilkan *score polarity*. Metode yang digunakan pada tahap ini adalah metode *Lexicon Based* dan *Naive Bayes Classifier*.

1. Lexicon Based

Dalam penelitian ini, metode *lexicon based* dengan *library vader sentiment* digunakan dalam analisis sentimen. *Lexicon based* merupakan kamus atau leksikon yang digunakan untuk pemilihan kata pada data atau dokumen [10]. Dalam implementasinya, tersedia dua kamus yaitu kamus dengan kumpulan kata yang bersentimen positif dan kamus dengan kumpulan kata yang bersentimen negatif yang digunakan untuk menjadi *wordlist* [11].

Metode analisis dari metode *lexicon based* adalah VADER (*Valance Dictionary and Sentiment Reasoner*). Vader digunakan untuk menganalisis data berdasarkan *lexicon* (kamus). Hasil dari Vader berupa kelas polaritas positif, netral, dan negatif dengan tambahan *compound score* atau skor total [12]. *Vader Sentiment Lexicon* memiliki 7.500 kata yang didalamnya terdapat sentimen yang terkait

dengan sinonim dan akronim serta kata berbahasa Inggris [13].

Leksikal merupakan kamus yang digunakan sebagai bahasa pokok dalam metode *lexicon based*. Untuk mendeteksi klasifikasi atau sentimen, pada penelitian ini memanfaatkan *library Python* dengan *score polarity* < 0 adalah sentimen negatif, *score polarity* = 0 adalah sentimen netral, dan *score polarity* > 0 adalah sentimen positif. Untuk proses klasifikasi sentimen dapat dilakukan dengan persamaan berikut :

$$\text{Sentence sentiment} \begin{cases} \text{positif if } S_{\text{positif}} > S_{\text{negatif}} \\ \text{netral if } S_{\text{positif}} = S_{\text{negatif}} \\ \text{negatif if } S_{\text{positif}} < S_{\text{negatif}} \end{cases}$$

2. Naive Bayes Classifier

Algoritma yang menggunakan konsep peluang atau yang biasa disebut probabilitas yang digunakan dalam klasifikasi untuk analisis sentimen disebut sebagai *Naive Bayes Classifier*. *Naive Bayes Classifier* juga terhitung dalam algoritma yang mudah digunakan dan sederhana serta bisa memperkirakan suatu kejadian berdasarkan hasil dari klasifikasi dengan baik [14]. Berikut merupakan rumus persamaan perhitungan dari nilai probabilitas metode *Naive Bayes Classifier*:

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)}$$

Dimana :

X = Dugaan sementara data dari suatu *class* spesifik

Y = Data dengan *class* yang belum diketahui

P(X|Y) = Peluang taksiran X dengan syarat Y
(probabilitas *posterior*)

P(X) = Peluang taksiran X (probabilitas *prior*)

P(Y|X) = Peluang taksiran Y dengan X

P(Y) = Peluang Y

Keterangan :

Probabilitas *posterior* : kemungkinan terdapat kelas X

Probabilitas *prior* : kemungkinan sampel awal kelas Y

E. Visualisasi

Setelah setiap tahap dan proses dilakukan, selanjutnya adalah tahap visualisasi. Pada penelitian ini, untuk tahap visualisasi dilakukan dengan menggunakan *library Matplotlib* dan *Wordcloud*. *Output* dari visualisasi ini adalah berupa gambar histogram yang juga akan menampilkan hasil akurasi persentase dari polaritas setiap sentimen yang dihasilkan. Sedangkan untuk visualisasi *wordcloud* menampilkan kata yang sering muncul pada setiap sentimennya.

F. Performa Metrik

Pada penggunaan suatu metode tentu memiliki gambaran tentang kinerja metode dalam proses pengklasifikasian data. Metode yang digunakan dalam

penelitian ini adalah metode *confusion matrix*. Dimana metode yang digunakan dalam menghitung akurasi dengan membandingkan hasil klasifikasi yang dilakukan secara aktual dengan hasil klasifikasi dari metode [14]. Untuk klasifikasi data yang sesungguhnya, pada penelitian ini melakukan *labelling* secara manual untuk mengetahui polaritasnya.

Tolak ukur hasil perhitungan metode *confusion matrix* yaitu *Precision*, *Recall*, *F1-Score*, *Macro avg*, *Weighted avg* dan *Accuracy*.

1. Precision

Precision yaitu visualisasi dari persentase keakuratan hasil perkiraan oleh metode yang digunakan.

Rumus dari *precision* adalah $precision = \frac{TP}{TP+FP}$

2. Recall

Recall yaitu visualisasi kesesuaian metode dalam mencari ulang sebuah informasi.

Rumus dari *recall* adalah $recall = \frac{TP}{TP+FN}$

3. F1-Score

F1-Score yaitu perbandingan antara rata-rata nilai presisi dan recall dari hasil pengujian.

Rumus dari *f1-score* adalah $f1 - score = \frac{TP}{TP + \frac{1}{2}(FP+FN)}$ atau

$$f1 - score = 2 * \frac{precision * recall}{precision + recall}$$

4. Macro avg

Macro avg yaitu rata-rata tidak tertimbang dari semua *F1-Score* per kelas.

Rumus *macro avg* adalah $macro avg = \frac{\text{jumlah nilai } f1\text{-score}}{\text{jumlah kelas}}$

5. Weighted avg

Weighted avg yaitu rata-rata semua *F1-Score* per kelas dengan mempertimbangkan dukungan masing-masing kelas.

Rumus *weighted avg* adalah

$$weighted avg = \sum f1 \text{ score per kelas} * \text{support proportion}$$

6. Accuracy

Accuracy yaitu visualisasi keakuratan model dalam mengelompokkan dengan benar.

Rumus dari *accuracy* adalah $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

III. HASIL DAN PEMBAHASAN

Pada penelitian analisis sentimen mengenai Aplikasi Investasi ini menggunakan metode *Lexicon Based* yang akan dibandingkan dengan *Naive Bayes Classifier*. Tujuannya untuk mengetahui hasil perbandingan dengan penelitian sebelumnya dengan menambahkan tahap *filtering* dan *stemming* pada proses *preprocessing* dalam keberhasilan meningkatkan akurasi serta untuk memberikan label sentimen *tweets* mengenai Aplikasi Investasi yang digolongkan menjadi label sentimen positif, netral, dan negatif.

Untuk penelitian ini, data didapatkan dengan *crawling* data dari *Twitter*. Untuk *crawling* data dari *Twitter* dibutuhkan akses token API *key Twitter* dengan mengakses <https://developer.twitter.com> dan mendaftar untuk melakukan validasi kepada pihak *developer Twitter* untuk mendapatkan akses token API *key*.

A. Pengumpulan Data

Pengumpulan data *twitter* dengan API Key *Twitter* sejumlah 698 data *tweets* pada tanggal 11 Januari 2022 sampai 19 Januari 2022 dengan *keyword* “aplikasi Investasi”. Setelah data *tweets* terkumpul, data yang diperoleh diubah menjadi data tabel agar mudah untuk diproses pada tahap selanjutnya. Data *frame* berisi tiga atribut, atribut-atributnya adalah:

1. *Username* : *username* mengandung nama pengguna yang berasal dari akun pembuat *tweets* yang diambil.
2. *Tweetcreatedts* : *tweetcreatedts* mengandung waktu saat pengguna membuat *tweets*.
3. *Teks* : *teks* mengandung isi yang berasal dari *tweet* yang telah dibuat.

Representasi secara visual dari proses *crawling* pengambilan data API key *Twitter* dapat dilihat pada Gbr. 4 dan Gbr. 5 adalah hasil data *frame* setelah *crawling* data berikut :

```
Proses Scrapping Selesai Dengan Jumlah Data 18
Proses Scrapping Selesai Dengan Jumlah Data 19
Proses Scrapping Selesai Dengan Jumlah Data 20
Proses Scrapping Selesai Dengan Jumlah Data 21
Proses Scrapping Selesai Dengan Jumlah Data 22
Proses Scrapping Selesai Dengan Jumlah Data 23
Proses Scrapping Selesai Dengan Jumlah Data 24
Proses Scrapping Selesai Dengan Jumlah Data 25
Proses Scrapping Selesai Dengan Jumlah Data 26
Proses Scrapping Selesai Dengan Jumlah Data 27
Proses Scrapping Selesai Dengan Jumlah Data 28
Proses Scrapping Selesai Dengan Jumlah Data 29
Proses Scrapping Selesai Dengan Jumlah Data 30
Proses Scrapping Selesai Dengan Jumlah Data 31
Proses Scrapping Selesai Dengan Jumlah Data 32
Proses Scrapping Selesai Dengan Jumlah Data 33
Proses Scrapping Selesai Dengan Jumlah Data 34
```

GBR. 4 HASIL PENGAMBILAN DATA API KEY TWITTER

	username	tweetcreatedts	text
0	pln_jogja	2022-01-19 23:49:05	Gak perlu nabung atau investasi buat menangin ...
1	meileriru	2022-01-19 23:19:49	Temukan peluang pasar sempurna dan investasi d...
2	RonaldDuwiri	2022-01-19 22:53:47	Temukan peluang pasar sempurna dan investasi d...
3	Firman20436201	2022-01-19 22:41:19	Mulai investasi #Bitcoin Anda di Luno - aplika...
4	AdiSuci26	2022-01-19 21:51:53	Mulai investasi #Bitcoin Anda di Luno - aplika...
5	dluo2009	2022-01-19 18:54:02	Mulai investasi #Bitcoin Anda di Luno - aplika...
6	ner_usu	2022-01-19 16:46:09	Mulai investasi #Bitcoin Anda di Luno - aplika...
7	ThecallP	2022-01-19 16:43:01	Mulai investasi #Bitcoin Anda di Luno - aplika...
8	Usmansasmila1	2022-01-19 16:13:39	Mulai investasi #Bitcoin Anda di Luno - aplika...
9	Antara_Kalleng	2022-01-19 15:48:42	Polri bongkar kasus investasi ilegal jual apli...

GBR. 5 HASIL DATA SETELAH DIUBAH MENJADI DATA FRAME

B. Preprocessing

Pada tahap *preprocessing* akan dilakukan beberapa tahapan yaitu *cleaning*, *case folding*, *tokenization*, *filtering* atau *stopwords removal*, dan *stemming*. Tahapan-tahapan tersebut dilakukan untuk

membersihkan dan menghapus data dari tanda baca serta simbol yang tidak diperlukan seperti tanda titik, tanda koma, tanda tanya, tanda seru, menghapus HTML dan URL, menghapus *hashtag* dan mention, menghapus emoji, serta menghilangkan simbol yang tidak relevan. Dalam tahap *preprocessing* juga dapat mengubah seluruh huruf pada data *tweets* menjadi huruf kecil, memotong sebuah kalimat menjadi potongan-potongan kata, menghilangkan *stop word*, dan mengubah kata yang berimbuhan menjadi kata dasar.

1. Cleaning

Pada tahap *cleaning* ini akan dilakukan penghapusan beberapa karakter seperti “@”, link URL, tanda titik, tanda koma, tanda seru, tanda tanya, menghapus emoji, mention dan *hashtag*, serta menghilangkan simbol yang tidak relevan. Tabel I berikut merupakan hasil dari tahap *cleaning* :

TABEL I
HASIL DATA CLEANING

Sebelum	Sesudah
Polri bongkar kasus investasi ilegal jual aplikasi robot Trading Evotrade https://t.co/KNJpHzfsdK	Polri bongkar kasus investasi ilegal jual aplikasi robot Trading Evotrade
Gabung bersama 9 juta pengguna kami dan mulai investasi aset kripto untuk dapat cuan! Semuanya hanya dalam satu aplikasi di Luno.	Gabung bersama juta pengguna kami dan mulai investasi aset kripto untuk dapat cuan Semuanya hanya dalam satu aplikasi di Luno
@semangathayu @karirfess Iyaa ka, tp maksudnya invest di cuddle itu ada aplikasi buat investasi di cuddle? Apa gimana ka maksudnya?	Iyaa ka tp maksudnya invest di cuddle itu ada aplikasi buat investasi di cuddle Apa gimana ka maksudnya

2. Tokenization

Pada tahapan *tokenization*, *library* NLTK dalam penelitian ini digunakan untuk *tokenize tweet*. Tahapan *tokenization* ini digunakan agar dapat memecah suatu kalimat dari *tweets* menjadi potongan-potongan, potongan-potongan tersebut berupa potongan kata. Tabel II berikut merupakan hasil sebelum dan sesudah melewati proses *tokenization*.

TABEL II
HASIL DATA TOKENIZATION

Sebelum	Sesudah
polri bongkar investasi ilegal jual aplikasi robot	['polri', 'bongkar', 'investasi', 'ilegal', 'jual',

trading evotrade	'aplikasi', 'robot', 'trading', 'evotrade']
gabung juta guna investasi aset kripto cuan aplikasi luno	['gabung', 'juta', 'guna', 'investasi', 'aset', 'kripto', 'cuan', 'aplikasi', 'luno']
iyaa ka maksud invest cuddle aplikasi investasi cuddle gimana ka maksud	['iyaa', 'ka', 'maksud', 'invest', 'cuddle', 'aplikasi', 'investasi', 'cuddle', 'gimana', 'ka', 'maksud']

3. Case Folding

Dalam tahapan *case folding* ini akan merubah huruf yang terdapat dalam data *tweets* menjadi huruf kecil atau *lowercase* supaya lebih mudah dibaca oleh komputer.

4. Filtering atau Remove Stopwords

Dalam tahapan *filtering* akan mempergunakan *library NLTK* dalam bahasa indonesia agar memudahkan tahap penghilangan kalimat. Dalam penelitian ini, peneliti juga menambahkan beberapa kata yang sering muncul pada data *tweets* seperti “yg”, “tdk”, “utk” dan lainnya supaya dapat mengurangi terjadinya *noise* dan menjadikan data lebih bersih.

5. Stemming

Pada tahapan *stemming* ini akan menggunakan *library stemmer factory* untuk memudahkan proses. Tujuan penelitian menambahkan tahapan *stemming* adalah agar dapat menjadikan kata dasar dari setiap kata dalam data *tweets* dan menghapus kata imbuhan yang terdapat pada awalan, sisipan, atau akhiran pada kata *tweets*.

Tabel III berikut merupakan perbandingan sebelum dan sesudah data melewati tahap *case folding*, *filtering*, dan *stemming*.

TABEL III
HASIL DATA CASE FOLDING, FILTERING, DAN STEMMING

Sebelum	Sesudah
Polri bongkar kasus investasi ilegal jual aplikasi robot Trading Evotrade	polri bongkar investasi ilegal jual aplikasi robot trading evotrade
Gabung bersama juta pengguna kami dan mulai investasi aset kripto untuk dapat cuan Semuanya hanya dalam satu aplikasi di Luno	gabung juta guna investasi aset kripto cuan aplikasi luno
Iyaa ka tp maksudnya invest di cuddle itu ada aplikasi buat investasi di	iyaa ka maksud invest cuddle aplikasi investasi cuddle gimana ka maksud

cuddle Apa gimana ka maksudnya	
--------------------------------	--

C. Translate

Setelah data *tweets* melakukan tahap *preprocessing*, selanjutnya adalah melakukan proses *translating* data *tweets* dengan *library translator*. Tabel IV merupakan hasil sebelum dan sesudah melalui tahap *translate*.

TABEL IV
HASIL DATA TRANSLATE

Sebelum	Sesudah
polri bongkar investasi ilegal jual aplikasi robot trading evotrade	The Indonesian National Police unloads illegal investments by selling the evotrade trading robot application
gabung juta guna investasi aset kripto cuan aplikasi luno	join million to invest in luno app crypto assets
iyaa ka maksud invest cuddle aplikasi investasi cuddle gimana ka maksud	does to purpose invest cuddle investment application cuddle how to purpose

D. Analisis Sentimen atau Pengklasifikasian

Setelah tahap *translate* dilakukan pada data *tweets* yang sudah bersih, selanjutnya adalah tahap pengklasifikasian. Pada penelitian ini, tahap pengklasifikasian dilakukan dengan menggunakan metode *Lexicon Based* dan *Naive Bayes Classifier*.

1. Lexicon Based

Metode *lexicon based* ini mempergunakan kamus atau *leksikon* sebagai pokok bahasa. Dasar penentuan dalam penelitian ini, untuk mengklasifikasikan sentimen dan *score polarity* adalah dengan memanfaatkan *library Vader Sentiment*. Sentimen dengan *score polarity* < 0 adalah sentimen negatif, *score polarity* = 0 adalah sentimen netral, dan *score polarity* > 0 adalah sentimen positif. Tabel V merupakan hasil dari sentimen dengan metode *lexicon based*.

TABEL V
HASIL SENTIMEN DENGAN LEXICON BASED

Tweet	Score Polarity	Sentimen
The Indonesian National Police unloads illegal investments by selling the evotrade trading	-0.5574	Negatif

robot application		
join million to invest in luno app crypto assets	0.4404	Positif
does to purpose invest cuddle investment application cuddle how to purpose	0.0	Netral

2. Naive Bayes Classifier

Proses klasifikasi dengan algoritma *Naive Bayes Classifier* di sini yaitu untuk menetapkan sebuah kalimat sebagai himpunan bersentimen positif, netral, atau negatif berlandaskan nilai perhitungan probabilitas dari rumus Bayes yang lebih besar. Jika hasil peluang kalimat tersebut untuk klasifikasi positif lebih besar dari pada klasifikasi negatif, maka kalimat tersebut termasuk ke dalam klasifikasi positif. Jika hasil peluang kalimat tersebut untuk klasifikasi positif sama dengan klasifikasi negatif maka termasuk kedalam klasifikasi netral. Sedangkan peluang untuk klasifikasi positif lebih kecil dari pada klasifikasi negatif, maka kalimat tersebut termasuk ke dalam klasifikasi negatif [15]. Tabel VI berikut merupakan hasil sentimen dari metode *Naive Bayes Classifier*.

TABEL VI
HASIL SENTIMEN DENGAN NAIVE BAYES CLASSIFIER

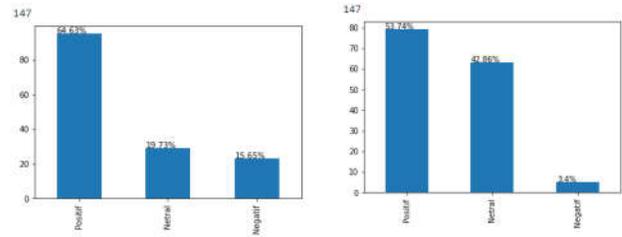
Tweet	Klasifikasi	Klasifikasi Naive Bayes
the indonesian national police unloads illegal investments by selling the evotrade trading robot appl	Negatif	Negatif
join million to invest in luno app crypto asset	Netral	Netral
does to purpose invest cuddle investment application cuddle how to purpos	Netral	Netral

E. Visualisasi

Setelah semua tahap telah dilakukan, selanjutnya adalah tahap visualisasi. *Output* dari tahap visualisasi dalam penelitian ini berupa *histogram* dan *wordcloud*. Tampilan *histogram* akan menampilkan persentase dari kelas polaritas masing-masing sedangkan *wordcloud* akan menampilkan gambar berupa kata yang sering muncul di setiap sentimen.

Gbr. 6 merupakan perbandingan tampilan *histogram* beserta nilai persentase tiap sentimen dari metode *lexicon based* dan metode *naive bayes classifier*,

sedangkan Gbr. 7, Gbr. 8, dan Gbr. 9 merupakan tampilan *wordcloud* dari sentimen positif, netral, dan negatif perbandingan antara kedua metode.



GBR. 6 TAMPILAN HISTOGRAM DAN PERSENTASE SETIAP SENTIMEN METODE LEXICON BASED DAN NAIVE BAYES CLASSIFIER



GBR. 7 TAMPILAN WORDCLOUD DARI SENTIMEN POSITIF METODE LEXICON BASED DAN NAIVE BAYES CLASSIFIER



GBR. 8 TAMPILAN WORDCLOUD DARI SENTIMEN NETRAL METODE LEXICON BASED DAN NAIVE BAYES CLASSIFIER



GBR. 9 TAMPILAN WORDCLOUD DARI SENTIMEN NEGATIF METODE LEXICON BASED DAN NAIVE BAYES CLASSIFIER

F. Performa Metrik

Pada tahapan pengujian performa metrik, pengujian dilakukan dengan metode *confusion matrix*, dimana peneliti akan membandingkan hasil klasifikasi yang dilakukan secara manual dengan hasil klasifikasi yang

telah dihasilkan oleh model dengan metode *Lexicon Based* dan *Naive Bayes Classifier*. Tabel VII merupakan hasil klasifikasi yang dilakukan secara manual yang akan dibandingkan dengan kedua metode. *Library sklearn* digunakan untuk membantu proses dalam menampilkan *classification report* dari perbandingan tersebut. Dalam membangun model *machine learning* ini, peneliti menggunakan sebuah ukuran atau yang biasa disebut *metric* untuk mengetahui seberapa baik kinerja model yang telah dibuat. *Precision*, *recall*, *f1-score*, *macro avg*, dan *weighted avg* digunakan untuk evaluasi *metric* dalam model klasifikasi ini. Model klasifikasi tersebut digunakan karena dapat memperlihatkan bagaimana model yang telah dibuat dapat mengambil suatu keputusan di dunia nyata.

TABEL VII
HASIL KLASIFIKASI DENGAN LABELLING MANUAL

Tweet	Klasifikasi
the indonesian national police unloads illegal investments by selling the evotrade trading robot appl	Negatif
join million to invest in luno app crypto asset	Netral
does to purpose invest cuddle investment application cuddle how to purpos	Netral

Pengujian hasil akurasi pada penelitian ini dilakukan dengan cara membandingkan klasifikasi secara aktual dengan hasil klasifikasi yang dihasilkan oleh model. Pengujian dilakukan dengan cara menampilkan *classification report* yaitu menghitung *accuracy*, *precision*, *recall*, *f1-score*, *macro avg*, dan *weighted avg*. Tabel VIII merupakan hasil *multiclass confusion matrix* dari metode *lexicon based* yang merupakan hasil dari data opini yang telah dirubah menjadi data numerik dan tabel IX merupakan hasil *classification report* metode *lexicon based*. Sedangkan tabel X merupakan hasil *multiclass confusion matrix* dari metode *naive bayes classifier* yang merupakan hasil dari data opini yang telah dirubah menjadi data numerik dan tabel XI merupakan hasil *classification report* metode *naive bayes classifier*.

TABEL VIII
HASIL MULTICLASS CONFUSION MATRIX LEXICON BASED

Klasifikasi	TP (True Positive)	FP (False Positive)	FN (False Negative)
Positif	70	24	15
Netral	21	8	26

Negatif	9	15	6
---------	---	----	---

TABEL IX
HASIL CLASSIFICATION REPORT METODE LEXICON BASED

	Precision	Recall	F1-Score	Support
Positif	73%	82%	78%	84
Netral	72%	44%	55%	48
Negatif	38%	60%	46%	15
Micro avg (accuracy)	67%	67%	67%	147
Macro avg	61%	62%	59%	147
Weighted avg	69%	67%	67%	147

Berdasarkan hasil *classification report* metode *lexicon based*, telah didapatkan nilai *accuracy* sebesar 67%, nilai *precision*, *recall*, dan *f1-score* pada sentimen positif secara berurutan adalah 73%, 82%, dan 78% dengan *support* 84, sedangkan sentimen netral adalah dengan nilai sebesar 72%, 44%, dan 55% dengan *support* 48, dan untuk sentimen negatif adalah dengan nilai sebesar 38%, 60%, dan 46% dengan *support* 15. Dari data diatas maka diperoleh *micro average* untuk *precision* 67%, *recall* 67%, *f1-score* 67%, dan *support* 147, kemudian untuk *macro average* untuk *precision* 61%, *recall* 62%, *f1-score* 59%, dan *support* 147, sedangkan *weighted average* untuk *precision* 69%, *recall* 67%, *f1-score* 67%, dan *support* 147. Dengan demikian, hal tersebut menunjukkan bahwa analisis sentimen dengan menggunakan *lexicon based* mempunyai tingkat akurasi kinerja yang cukup tinggi karena memiliki nilai akurasi diatas 50%.

TABEL X
HASIL MULTICLASS CONFUSION MATRIX NAIVE BAYES CLASSIFIER

Klasifikasi	TP (True Positive)	FP (False Positive)	FN (False Negative)
Positif	71	11	13
Netral	41	18	7
Negatif	6	0	9

TABEL XI
HASIL CLASSIFICATION REPORT METODE NAIVE BAYES CLASSIFIER

	Precision	Recall	F1-Score	Support
Positif	88%	82%	85%	84
Netral	65%	88%	74%	48
Negatif	100%	27%	42%	15
Micro avg (accuracy)	78%	78%	78%	147
Macro avg	84%	65%	67%	147
Weighted avg	82%	78%	77%	147

UCAPAN TERIMA KASIH

Sedangkan berdasarkan hasil *classification report* metode *naive bayes classifier*, telah didapatkan nilai *accuracy* sebesar 78%, nilai *precision*, *recall*, dan *f1-score* pada sentimen positif secara berurutan adalah 88%, 82%, dan 85% dengan *support* 84, sedangkan sentimen netral adalah dengan nilai sebesar 65%, 88%, dan 74% dengan *support* 48, dan untuk sentimen negatif adalah dengan nilai sebesar 100%, 27%, dan 42% dengan *support* 15. Dari data diatas maka diperoleh *micro average* untuk *precision* 78%, *recall* 78%, *f1-score* 78%, dan *support* 147, kemudian untuk *macro average* untuk *precision* 84%, *recall* 65%, *f1-score* 67%, dan *support* 147, sedangkan *weighted average* untuk *precision* 82%, *recall* 78%, *f1-score* 77%, dan *support* 147.

IV. KESIMPULAN

Penelitian ini merupakan analisis sentimen mengenai pendapat masyarakat terhadap aplikasi investasi melalui media sosial *twitter* dengan metode *lexicon based* dan *naive bayes classifier*. Data *crawling* dari media sosial *twitter* disimpan dan selanjutnya akan diklasifikasikan menjadi tiga polaritas, yaitu positif, netral, dan negatif. Dari total 698 data *tweets* yang diperoleh, setelah melakukan penghilangan atau penghapusan terhadap *tweets* dan melalui tahapan *preprocessing*, *translate*, dan selanjutnya masuk pada metode *lexicon based* atau *naive bayes classifier* maka didapatkan total keseluruhan data tersebut adalah 147 data *tweets*. Hasil akurasi kinerja dari metode *lexicon based* adalah sebesar 67% dengan polaritas positif merupakan persentase tertinggi dari proses klasifikasi sebesar 64,63%. Sedangkan hasil akurasi kinerja metode *naive bayes classifier* sebesar 78% dengan polaritas positif merupakan persentase tertinggi dalam pengklasifikasian sebesar 53,74%. Dari hasil penelitian, perbandingan antara hasil akurasi klasifikasi yang dilakukan antara metode *lexicon based* dan *naive bayes classifier* adalah 67% dan 78%. Dengan akurasi yang berbeda, menunjukkan bahwa analisis sentimen dengan metode *naive bayes classifier* memiliki akurasi yang lebih tinggi dari pada metode *lexicon based*.

V. SARAN

Pada penelitian selanjutnya disarankan untuk menggunakan data yang lebih banyak lagi serta menggunakan metode selain *lexicon based* dan *naive bayes classifier* serta menggunakan metode selain yang digunakan dalam perhitungan kinerja model. Selain itu untuk pengembangan lebih lanjut dapat diimplementasikan dalam bentuk sistem, baik berupa web atau aplikasi yang dapat dengan lebih mudah untuk dipahami dan digunakan oleh masyarakat umum.

Puji syukur penulis panjatkan kepada Allah SWT. atas rahmat, ridho, serta hidayahNya sehingga penulis dapat menyelesaikan penelitian ini dengan hasil yang InsyaAllah penuh barokah ini dan bermanfaat untuk setiap kalangan. Kedua kalinya sholawat dan salam tercurahkan kepada Nabi besar Muhammad SAW. yang telah menuntun penulis agar bisa menjadi manusia yang lebih baik, InsyaAllah.

Penulis juga berterimakasih kepada semua pihak yang telah memberikan dukungan terbaiknya kepada penulis sehingga penulis dapat menyelesaikan tugas yang seharusnya diselesaikan. Kepada keluarga terutama kedua orangtua penulis, penulis ucapkan terimakasih sebanyak-banyaknya karena dengan adanya mereka penulis dapat merasakan indahnya dunia perkuliahan beserta suka duka sebagai pembelajaran hidup. Selain itu terimakasih juga atas waktu yang mereka relakan demi menghidupi anak-anaknya sehingga menjadi anak yang InsyaAllah bermanfaat di dunia dan akhirat kelak, Aamiin.

Untuk orang disekitar termasuk dosen pembimbing, dosen penguji, dan teman-teman penulis, terimakasih, telah menjadi yang terbaik.

REFERENSI

- [1] H. Manuel "Pengaruh Kemudahan, Keamanan, Kepercayaan dan Kualitas Informasi Pada Aplikasi Investasi Online Terhadap Minat Investasi Saham," *J. Ilmiah Mahasiswa FEB*, Vol. 7, No. 2, 2019.
- [2] Sawidji. (2008). *Professional Investing*. Jakarta: Elex Media Komputindo.
- [3] G. R. Siri, And D. Meirini, "Pengaruh Modal Investasi, Teknologi yang Memadai, dan Motivasi Terhadap Minat Mahasiswa Berinvestasi Di Pasar Modal," *J. Penelit. Mahasiswa Ilmu Sos., Eko., dan Bis. Islam.*, Vol. 1, Isu 1, 2021.
- [4] M. A. F. Habib, "Kajian Teoritis Pemberdayaan Masyarakat dan Ekonomi Kreatif," *J. of Islamic Tourism, Halal Food, Islamic Traveling, and Creative Economy*, Vol. 1, No. 2, 2021.
- [5] P. A. Sumitro, Rasiban, D. I. Mulyana, And W. Saputro, "Sentimen Terhadap Vaksin Covid-19 di Indonesia pada Twitter Menggunakan Metode Lexicon Based," *J. Informatika dan teknologi Komputer.*, Vol. 2, No. 2, Pp. 50-56, 2021.
- [6] N. Rochmawati And S. C. Wibawa, "Opinion Analysis On Rohingya Using Twitter Data," In *Iop Conference Series: Materials Science And Engineering*, 2018, Vol. 336, No. 1, P. 012013.
- [7] B. Gunawan, H. Sastypratiwi, And E. E. Pratama, "Sistem Analisis Sentimen Pada Ulasan Produk Menggunakan Metode Naive Bayes," *Jepin J. Edukasi Dan Penelit. Inform.*, Vol. 4, No. 2, Pp. 113-118, 2018.
- [8] F. Ratnawati, "Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter", *Inovtek Polbeng-Seri Inform.*, Vol. 3, No. 1, Pp. 50-59, 2018.
- [9] M. N. Edit, "Analisis Sentimen *Twitter* Bertema Pembelajaran Tatap Muka (PTM) Terbatas Menggunakan Metode *Lexicon Based*," 2021.
- [10] M. Taboada, J. Brooke, M. Tofiloski, K. Volf, And M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, Vol. 37, No. 2, Pp. 267-307, 2011.
- [11] N. S. Fathullah, Y. A. Sari, And P. P. Adikara, "Analisis Sentimen Terhadap Rating dan Ulasan Film dengan menggunakan Metode Klasifikasi Naive Bayes dengan Fitur *Lexicon-Based*," *J. Pengemb. Teknol. Inf. dan Ilmu Komputer*, Vol. 4, No. 2, Pp. 590-593, 2020.

- [12] N. Anggraini, And H. Suroyo, "Comparison of Sentiment Analysis against Digital Payment 'T-cash and Go-Pay' in Social Media Using Orange Data Mining," *J. Of Information Systems and Informatics*, Vol. 1, No. 2, Pp. 152-153, 2019.
- [13] B. Y. Melani, S. R. Wardhana, And D. Puspita, "Analisa Kualitas Fitur Aplikasi Mobile Dengan Menggunakan Pendekatan Sentimen Grey," *Semin. Nas. Sains dan Teknol. Terap.*, Vol. VII, 2019.
- [14] P. P. E. Indarbensyah, And N. Rochmawati, "Penerapan *N-Gram* menggunakan Algoritma *Random Forest* dan *Naive Bayes Classifier* pada Analisis Sentimen Kebijakan PPKM 2021," *J. of Informatics and Computer Science*, Vol. 2, No. 4, 2021.
- [15] D. A. Muthia, "Analisis Sentimen Pada Review Buku Menggunakan Algoritma Naive Bayes," *J. Paradigma*, Vol. 16, No. 1, 2014.