Kecendrungan Sentimen dengan Pendekatan *Support Vector Machine* pada Komunitas yang Berpengaruh di Twitter

Adinda Suci Rezeky Tami Batubara[1], Warih Maharani[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]adindabatubara@students.telkomuniversity.ac.id, [2]wmaharani@telkomuniversity.ac.id,

Abstrak

Analisis sentimen adalah studi komputasional dari opini-opini orang. Analisis sentimen dapat mengelompokkan teks yang ada dalam kalimat atau dokumen tersebut, yang berupa sentimen positif dan sentimen negatif. Penelitian ini bertujuan untuk melihat kecendrungan sentimen masyarakat di twitter terhadap kegiaan investasi di Indonesia. Data yang digunakan berupada data cuitan dengan kata kunci "#investasimilenial", "#investasi" dan "#bursaefekindonesia". Cuitan yang didapat kemudian diolah dengan melakukan *text preprocessing* kemudian melakukan pelabelan pada setiap cuitan serta melakukan pembobotan TF-IDF pada setiap *term* dan selanjutnya dilakukan klasifikasi dengan menggunakan metode *Support Vector Machine.* Parameter yang digunakan dalam metode *Support Vector Machine* ini adalah kernel Linear. Metode *Support Vector Machine* digunakan karena metode ini memiliki tingkat akurasi yang tinggi dibandingkan dengan metode lainnya. Pengujian dilakukan dengan perhitungan *precision, recall, F-Measure* dan akurasi. Hasil klasifikasi yang diperoleh dengan pendekatan *Support Vector Machine* ini memiliki akurasi sebesar 0,87 atau 87%, presisi positif sebesar 0.88 atau 88%, presisi negatif 0,67 atau 67%, *recall* positif sebesar 0,98 atau 98%, *recall* negatif sebesar 0,22 atau 22%, *f1-score* positif sebesar 0,93 atau 93% dan *f1-score* negatif sebesar 0,33 atau 33%. Hasil yang didapatkan dari penelitian ini dapat membantu, mempermudah serta memberikan pertimbangan atau gambaran untuk pihak-pihak yang berkepentingan untuk melihat ketertarikan masyarakat dalam berinvestasi di Indonesia.

Kata kunci : Analisis Sentimen, *Support Vector Machine, Text Mining*.

Abstract

Sentiment analysis is a computational study of people's opinions. The sentiment analysis will cluster text in sentences or documents to find out the opinion in these sentences or documents, either positive sentiment or negative sentiment. The purpose of this research is to see sentiment leaning of people's on twitter about ivestment activities in Indonesia. The data used in this research with the keywords "#investasimilenial", "#investasi" dan "#bursaefekindonesia". Tweets received are then processed by text preprocessing then labeling each tweet and weighing TF-IDF in each term and then classifying using the Support Vector Machine method. The parameter used in the Support Vector Machine method is the Linear kernel. The Support Vector Machine method is used because this method has a high degree of accuracy compared to other methods. Testing is done by calculation precision, recall, F-Measure dan accuracy. The results of this classification with Support Vector Machine have an accuracy of 0.87 or 87%, positive precision of 0.88 or 88%, negative precision of 0.67 or 67%, positive recall of 0.98 or 98%, negative recall of 0.22 or 22%, positive f1-score of 0.93 or 93% and negative f1-score of 0.33 or 33%. The results obtained from this research can be used as a benchmark for interested parties to see at people's interest in investing in Indonesia.

Keywords: Sentiment Analysis, Support Vector Machine, *Text Mining*.

1. Pendahuluan

Latar Belakang

Media sosial saat ini menjadi media komunikasi yang sangat popular dikalangan masyarakat Indonesia. Media sosial yang popular dan banyak diminati antara lain Facebook, Instagram dan Twitter. Twitter merupakan salah satu media sosial yang banyak digunakan saat ini. Twitter merupakan media sosial dengan pertumbuhan tercepat sejak tahun 2006 menurut MIT Technology Review. Pengguna Twitter saat ini memiliki lebih dari 330 juta pengguna aktif. Twitter memungkinkan penggunanya untuk menulis tentang kehidupan mereka, berbagi informasi ataupun menyampaikan opini mengenai suatu hal. Topik yang sedang menjadi perbincangan hangat di Twitter akan menjadi *trending topic.* Penyampaian opini melalui media sosial Twitter dapat menjadi salah satu media untuk menganalisis kecendrungan informasi mengenai suatu topik apakah cenderung ke positif atau negatif. Pihak-pihak yang memerlukan informasi mengenai opini masyarakat terhadap kata kunci tertentu, dapat memanfaatkan Twitter sebagai data analisisnya.

Twitter dapat digunakan sebagai media propaganda bagi kelompok-kelompok yang memiliki tujuan tertentu. Propaganda tersebut dapat berupa masalah politik ataupun masalah sosial. Penelitian ini membahas mengenai propaganda pada masalah sosial yaitu ketertarikan masyarakat atau komunitas investor dalam berinvestasi di Indonesia. Penelitian ini melakukan analisis pada hasil pencarian pada cuitan dengan menggunakan *hashtag* #investasimilenial, #investasi dan #bursaefekindonesia. Kata kunci yang digunakan saling berkaitan mengenai opini masyarakat tentang investasi di Indonesia. Analisis sentimen merupakan serangkaian cara atau teknik yang bertujuan untuk melakukan deteksi dan mengekstrak informasi yang subjektif seperti opini dari suatu tulisan [11]. Analisis sentimen digunakan untuk memproleh informasi dari sebuah tulisan dan mengklasifikasi tulisan tersebut ke bentuk sentiment positif atau negatif. Analisis sentimen digunakan pada penelitian in untuk melihat kecendrungan sentimen masyarakat mengenai investasi di Indonesia.

Penelitian yang terkait adalah pada penelitian Kharde, et al. [9] tentang penerapan analisis sentimen dengan dokumen teks Twitter yang membuktikan bahwa metode *Support Vector Machine* memberikan akurasi yang cukup besar dan unggul dibandingkan dengan metode *Maximum Entropy* dan *Naïve Bayes*. Penelitian lainnya dilakukan oleh Vidya, et al. [10] mengenai analisis sentimen terhadap dokumen Twitter yaitu analisis sentimen terhadap reputasi *provider handphone* dengan menggunakan SVM, *Naïve Bayes, dan Decision Tree.* Hasil yang didapatkan menunjukkan bahwa *Support Vector Machine* mempunyai akurasi paling besar dibandingkan dengan metode *Naïve Bayes* dan metode *Decision Tree*. Berdasarkan penelitian tersebut dapat diketahui bahwa penggunaan metode *Support Vector Machine* menunjukkan akurasi yang tinggi atau lebih baik dibandingkan dengan metode lainnya.

Maka, penelitian ini menggunakan *Support Vector Machine* sebagai metode untuk proses klasifikasi dan menggunaan TF-IDF sebagai ekstraksi fitur. *Support Vector Machine* digunakan untuk proses klasifikasi dari penelitian ini untuk melihat kecendrungan sentimen masyarakat terhadap investasi di Indonesia. *Support Vector Machine* digunakan sebagai fungsi pemisah (klasifier) optimal yang mampu memisahkan dua set data dari dua kelas yang berbeda yaitu sentimen positif atau sentimen negatif.

Topik dan Batasannya

Berdasarkan latar belakang yang telah diuraikan di atas, perumusan masalah dari penelitian ini adalah untuk mengetahui kecendrungan sentimen masyarakat ataupun komunitas investor mengenai investasi di Indonesia dengan menggunakan *hashtag* #investasimilenial, #investasi dan #bursaefekindonesia serta penggunaan metode *Support Vector Machine* untuk proses klasifikasi.

Adapun batasan masalah dari penelitian ini adalah menggunakan data cuitan para pengguna Twitter yang memberikan opini mengenai investasi di Indonesia dengan menggunakan *hashtag* #investasimilenial, #investasi dan #bursaefekindonesia dalam cuitan mereka dengan *Twitterscraper* tanpa harus menggunakan Twitter API. Pengambilan data cuitan dilakukan secara *real-time.* Cuitan yang terkumpul adalah cuitan sejak tanggal 28 Mei 2020 hingga tanggal 30 Mei 2020 sebanyak 300 cuitan. Pada penelitian ini melakukan klasifikasi sentimen menjadi dua kelas yaitu sentimen positif dan sentimennegatif.

Tujuan

Tujuan dari penelitian ini adalah untuk mengetahui kecendrungan sentimen masyarakat atau komunitas investor mengenai masalah investasi di Indonesia dengan menggunakan metode *Support Vector Machine* pada proses klasifikasi. Hasil yang didapatkan dari penelitian ini dapat membantu, mempermudah serta memberikan pertimbangan atau gambaran untuk pihak-pihak yang berkepentingan untuk melihat ketertarikan masyarakat dalam berinvestasi di Indonesia.

Organisasi Tulisan

Pada bagian selanjutnya, akan dijelaskan mengenai penelitian atau studi yang terkait dengan penelitian ini, beserta penulis, judul, dan hasil penelitian tersebut. Pada bagian tiga, dijelaskan sistem yang dibangun pada penelitian ini beserta teori-teori terkait. Bagian empat akan menjelaskan hasil evaluasi dari sistem yang digunakan. Sistem evaluasi yang digunakan pada penelitian ini adalah dengan perhitungan akurasi, presisi, *recall*, dan *f1-score*. Lalu, bagian terakhir adalah kesimpulan dan saran dari hasil penelitian ini.

2. Landasan Teori

2.1 Twitter

Twitter merupakan *microblogging* dan layanan *social network* yang memungkinkan pengguna untuk saling *follow* satu sama lain. Twitter adalah salah satu layanan *microblogging* yang paling terkenal, dimana banyak komunitas menggunakannya. Sebelumnya pengguna Twitter dapat mengirim dan menerima cuitan berupa pesan teks yang dibatasi hanya 140 karakter, namun saat ini jumlah maksimal karakternya ditambah menjadi 280 karakter. Keunggulan Twitter diantara media sosial lainnya adalah kemudahan akses informasi yang sangat cepat.

Twitter menggunakan bit.ly untuk memperpendek otomatis semua URL yang dikirim. Cuitan dapat dilihat secara publik, namun pengirim dapat membatasi pengiriman pesan ke daftar teman-teman mereka saja. Twitter memiliki beberapa fitur yang digunakan secara umum oleh pengguna yaitu *followers*, *following*, *re-tweet*, *hashtag*, *mentions*, *favorite*, *direct message*, *reply*, cuitan dan sebagainya. Twitter dapat digunakan untuk menyebarkan berbagai informasi melalui sebuah cuitan mengenai pikiran dan pengalaman mereka [1].

2.2 Text Mining

Text mining adalah penambangan yang dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu yang tidak diketahui sebelumnya atau menemukan kembali informasi yang tersirat secara implisit yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda [7]. *Text mining* merupakan teknik yang digunakan untuk menyelesaikan masalah klasifikasi, *clustering*, *informationextraction* dan *information retrieval* [8]. *Text mining* bertujuan untuk mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan Analisa keterhubungan antara dokumen.

Text mining dapat memberikan solusi dari permasalahan seperti pemrosesan, pengorganisasian dan menganalisa *unstructured text* (teks tidak terstuktur) dalam jumlah besar. *Text mining* memiliki tujuan dan menggunakan proses yang sama dengan *data mining*, hanya saja memiliki masukan yang berbeda. Masukan untuk *text mining* adalah *unstructured data* (data tidak terstuktur) yaitu data yang tidak memiliki bentuk atau struktur khusus seperti PDF, kutipan teks dan sebagainya atau *semistructured* (semi terstuktur) yaitu data yang memiliki struktur namun belum sepenuhnya terstruktur misalnya daftar riwayat hidup (CV), sedangkan masukan *data mining* adalah *structured data* (data terstuktur) atau basis data sebagai masukan. Permasalahan pada *text mining* sama dengan permasalahan pada *data mining*, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah dan data *noise*. Tahap awal pada *text mining* disebut dengan *text preprocessing* yaitu yang terdiri dari *case folding*, normalisasi, tokenisasi dan *stopwords.*

2.2.1 Pembobotan TF-IDF

Pembobotan TF-IDF adalah suatu teknik yang digunakan untuk memberikan bobot terhadap suatu kata yang sudah diolah atau diekstrak. TF-IDF merupakan suatu metode yang sering digunakan dalam melakukan pembobotan [12]. Tahapan pembobotan kata dengan TF-IDF yaitu,

a. Menghitung *weight term frequency* dengan persamaan sebagai berikut,

$$W_{tft,d} = \begin{cases} 1 + \log_{10} tf\ t, d\ , if tf_{t,d} > 0 \\ 0, \quad otherwise \end{cases} \quad (1)$$

b. Menghitung IDF dengan persamaan sebagai berikut,

$$idf_t = \log_{10} \left(\frac{N}{df_t}\right) \quad (2)$$

c. Menghitung TF-IDF dengan persamaan sebagai berikut,

$$W_{t,d} = W_{tft,d} \times idf_t \quad (3)$$

Dengan $W_{tft,d}$ adalah bobot frekuensi *term*, $tf_{t,d}$ adalah frkuensi term, idf_t adalah nilai *inverse document frequency*, N adalah banyaknya dokumen, df_t adalah banyak dokumen yang mengandung suatu *term* dan $W_{t,d}$ adalah nilai bobot TF-IDF.

2.3 Analisis Sentimen

Keterkaitan antara *text mining*, *data mining* dan *opion mining* yaitu *text mining* merupakan proses ekstraksi pola (informasi dan pengetahuan yang berguna) dari sejumlah data tak terstruktur yang nantinya akan diperoleh pola-pola data, tren dan ekstraksi pengetahuan yang potensial dari data teks [14]. Masukan untuk penambangan teks adalah data yang tidak (atau semi) terstruktur, seperti dokumen, *word*, PDF, kutipan teks dll sedangkan masukan untuk penambangan data adalah data yang terstruktur [15]. Salah satu tujuan penggunaan *text mining* adalah analisis sentimen (*opinion mining*). Analisis sentimen adalah riset komputasional dari opini, sentimen dan emosi yang diekpresikan secara tekstual [2].

Analisis sentimen atau *opinion mining* merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecendrungan opini terhadap sebuah masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif [2]. Analisis sentimen adalah kegiatan dengan melakukan analisa terhadap pendapat, opini, sikap atau emosi seseorang mengenai suatu produk,

topik atau permasalahan tertentu sehingga bisa diketahui hal tersebut masuk kedalam sentimen positif atau negatif. Sentimen positif atau negatif ini yang dapat dijadikan sebagai parameter pengambilan keputusan. Untuk melakukan analisis sentimen ada beberapa algoritme yang dapat digunakan salah satunya adalah algoritme *Support vector machine* (SVM).

2.4 *Support Vector Machine* (SVM)

Support vector machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi [13]. *Support Vector Machine* dalam penerapannya terdapat tahapan pelatihan kemudian dilanjutkan tahapan pengujian. Metode pembelajaran ini ialah *supervised* yang membutuhkan membutuhkan ketersediaan data berlabel. Fungsi pemetaan ini bisa berupa fungsi klasifikasi atau fungsi regresi [4]. *Support Vector Machine* dikembangkan untuk memecahkan masalah klasifikasi karena *support vector machine* memiliki kemampuan yang lebih baik dalam menggeneralisasi data dibandingkan denan teknik yang sudah ada sebelumnya [5]. *Support Vector Machine* merupakan salah satu *machine learning* yang melakukan pelatihan dengan menggunakan *training dataset* dan melakukan generalisasi dan membuat prediksi dai data baru. Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. *Support Vector Machine* juga dapat mengatasi masalah klasifikasi dan regresi dengan *linear* maupun *non-linear*.

Support Vector Machine bertujuan untuk menemukan fungsi pemisah (*classifier/hyperplane*) terbaik untuk memisahkan dua buah *class* pada *input space*. *Hyperplane* terbaik anatara kedua *class* dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* dengan *pattern* terdekat dari masing-masing *class*. *Pattern* yang paling dekat disebut sebagai *support vector*. Pencarian lokasi *hyperplane* merupakan inti dari proses pembelaran pada *Support Vector Machine.* Kelebihan dari metode SVM adalah memiliki konsep *structural risk minimization* (SRM) dimana konsep tersebut mampu mengatasi permasalahan *overfitting*.

Metode *Support Vector Machine* mempunyai beberapa model pendekatan atau biasa disebut kernel seperti *linear, polynomial* dan *gaussian/ radial basis* function (RBF). Persamaan kernel tersebut dapat dilihat sebagai berikut,

a. *Linear*

$$K(x_i, x_j) = x_i^T x_j + C \quad (4)$$

Dengan x_i dan x_j merupakan vektor dari data set dan C merupakan constant.

b. *Polynomial*

$$K(x_i, x_j) = (\gamma . x_i^T x_j + c)^d, \gamma > 0 \quad (5)$$

Dengan x_i dan dan x_j merupakan vektor dari data set, γ adalah parameter untuk mengkontrol kecepatan proses learning, C merupakan constant dan d merupakan pangkat *polynomial* yang digunakan

c. *Gaussian/Radial Basis Function* (RBF)

$$K(x_i, x_j) = exp(-\gamma||X_i - X_j||^2), \gamma > 0 \quad (6)$$

Dengan x_i dan x_j merupakan vektor dari data set, γ adalah parameter untuk mengkontrol kecepatan proses learning dan exp merupakan basis dari logaritma alami.

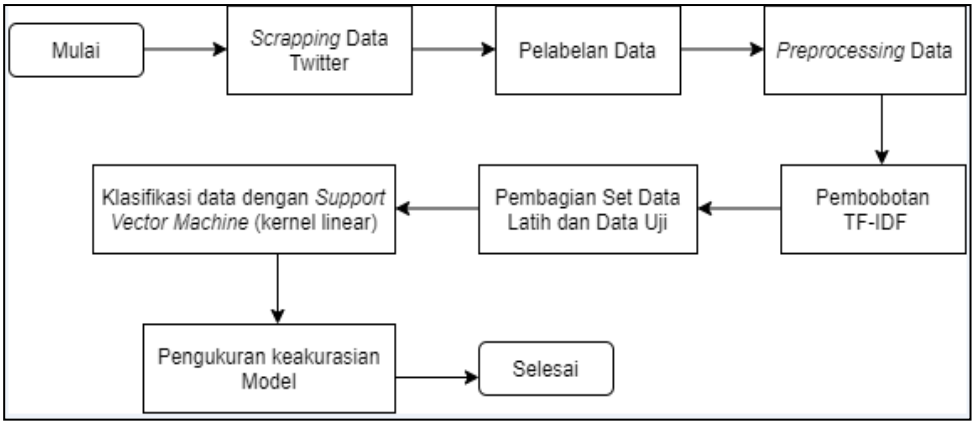
Tujuan dari penggunaan kernel ini ialah untuk mengimplementasikan suatu model pada ruang dimensi yang lebih tinggi (*feature space*) tanpa harus mendefinisikan fungsi pemetaan dari ruang input ke *feature space*, sehingga untuk kasus yang *non lineary separable* pada ruang input diharapkan menjadi l*ineary separable* pada *feature space*. Selanjutnya dapat digunakan *hyperplane* sebagai *decision boundary* secara efisien. Pada penelitian ini hanya menggunakan kernel *linear* dalam pengaplikasiannya karena kernel *linear* memiliki tingkat akurasi yang tinggi dibandingkan yang lainnya.

Kernel *linear* merupakan fungsi kernel yang baik digunakan ketika data sudah terpisah secara *linear*. Kernel ini digunakan ketika data yang akan diklasifikasikan dapat terpisah dengan sebuah garis/hyperplane. Kernel linear mungkin merupakan kernel yang paling sederhana, Diasumsikan bahwa x $\in$ R^n dan mendefinisikan k(x,x') = {x,x') = $\sum_i$xi xi'. Jika x dan x' padat maka perhitungan waktu pada kernel memerlukan waktu O(n), untuk sparse vectors dapat di reduksi ke O (|nnz ∩ nnz(x')|), dimana nnz(.) menunjukkan himpunan indeks tidak nol dari sebuah vector dan |.| menunjukkan ukuran dari sebuah himpunan. Linear kernel digunakan untuk representasi data berupa vector. Kernel ini juga digunakan untuk *text mining* dimana dokumen di representasikan oleh vector yang terdiri dari frekuensi kata–kata.

3. Sistem yang Dibangun

3.1 Gambaran Umum Sistem

Gambaran umum sistem adalah proses sistem secara garis besar dengan menggunakan *flowchart* atau diagram. Gambaran umum sistem pada penelitian ini adalah sebagai berikut.



Gambar 1. ***Flowchart* gambaran umum sistem**

3.2 Tahapan Sistem

Sistem yang dibangun dalam penelitian ini memiliki beberapa tahapan pengerjaan. Tahapan yang dilakukan beserta penjelasannya adalah sebagai berikut.

3.2.1 *Scraping* Data Twitter

Dalam penelitian ini, data yang digunakan diambil menggunakan *Twitterscraper*, tanpa menggunakan Twitter API. Data yang diambil adalah data cuitan para pengguna Twitter yang menggunakan *hashtag* #investasimilenial, #investasi dan #bursaefekindonesia yang di *scrap* secara *real-time* pada tanggal 30 Mei 2020. Cuitan yang terkumpul adalah cuitan sejak tanggal 28 Mei 2020 hingga tanggal 30 Mei 2020 sebanyak 300 cuitan. Berikut adalah sampel data hasil scraping menggunakan *Twitterscraper*. Berikut adalah contoh sampel dataset yang digunakan beserta pelabelan kelas sentimen yang dilabelin secara manual oleh penulis,

Tabel 1. **Sampel Dataset**

No.	Cuitan	Nama Akun	Label
1	Jgn memikirkan untungnya aja, tp jg resikonya. Tdk pernah ada jaminan bhw #Reksadana selalu akan profit (bahkan dlm jk panjang sekalipun).	@donioptions	Negatif
2	IHSG anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.	@KontanNews	Negatif
3	Ada pandemi corona Danareksa Mawar Fokus 10 jadi pilihan #investasi #reksadana #saham	@KontanNews	Positif
4	Wahai generasi milenial, kini saatnya dirimu berinvestasi 😎😎 Masa Muda Saatnya Pilih Investasi Terbaik untuk Masa Depanmu	@IidYanie	Positif

3.2.2 Preprocessing

Preprocessing data adalah proses pembersihan dan mempersiapkan teks untuk klasifikasi [6]. *Preprocessing* juga dapat diartikan juga sebagai tahapan dimana data mentah akan diolah menjadi data yang berkualitas. Tahapan ini dibutuhkan untuk mendapatkan data yang berkualitas dan juga untuk meningkatkan efisiensi pada proses pencarian informasi. Serta tahapan ini digunakan sebagai pembersihan pada teks dengan menghilangkan bagian-bagian yang tidak diperlukan, yang bertujuan untuk mengurangi *noise* dan *missing value*, sehingga dapat memudahkan proses selanjutnya. Tahapan-tahapan *preprocessing* yang dilakukan pada penelitian ini adalah sebagai berikut,

1. *Case Folding*

Case folding yaitu merubah semua karakter huruf dalam dokumen menjadi huruf kecil, hanya huruf "a" sampai "z" yang diterima. Karakter yang dianggap tidak valid yaitu karakter selain huruf akan dihilangkan seperti angka, tanda baca dan karakter tersebut akan dianggap delimeter. Berikut contoh hasil proses *case folding* pada penelitian ini beserta penghilangan karakter emoji atau karakter special dalam data,

Tabel 2. **Proses *Case Folding***

No.	Cuitan	*Case folding*
1	Jgn memikirkan untungnya aja, tp jg resikonya. Tdk pernah ada jaminan bhw #Reksadana selalu akan profit (bahkan dlm jk panjang sekalipun).	jgn memikirkan untungnya aja, tp jg resikonya. tdk pernah ada jaminan bhw #reksadana selalu akan profit (bahkan dlm jk panjang sekalipun).
2	IHSG anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.	ihsg anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.
3	Ada pandemi corona Danareksa Mawar Fokus 10 jadi pilihan #investasi #reksadana #saham	ada pandemi corona danareksa mawar fokus 10 jadi pilihan #investasi #reksadana #saham
4	Wahai generasi milenial, kini saatnya dirimu berinvestasi 😎 😎 Masa Muda Saatnya Pilih Investasi Terbaik untuk Masa Depanmu	wahai generasi milenial, kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu

2. Normalisasi

Normalisasi kata dilakukan untuk penggantian kata yang tidak baku menjadi baku, karena kata yang sudah baku akan cenderung lebih kecil keambiguitasnya dalam pelafalan dibandingkan dengan kata yang tidak baku. Contoh kata tidak baku seperti "jd", "bs", "dlm", "hrs", "sbg" dan lain sebagainya kalau dijadikan kata baku menjadi "jadi", "bisa", "dalam", "harus", "sebagai" dan lain sebagainya.

Untuk itu perlu dilakukan normalisasi kata dengan cara mengganti kata yang tidak baku dengan kata yang sesuai konteknya. Pada tahapan normalisasi, kata terlebih dahulu dibuat kamus kata baku dan tidak baku. Pada penelitian ini proses pengantian kata tidak baku menjadi baku menggunakan dataset. Dataset yang digunakan adalah sebuah kamus yang berisi kumpulan data tidak baku dengan kata bakunya. Hal ini dilakukan untuk memudahkan proses penggantian kata. Dataset kata tidak baku dan kata baku yang digunakan sebanyak 283 kata. Hasil dari normalisasi kata ini adalah berupa kumpulan tweet yang berisi kata-kata yang sudah baku. Berikut contoh proses normalisasi pada dataset,

Tabel 3. **Proses Normalisasi**

No.	Cuitan	Normalisasi
1	jgn memikirkan untungnya aja, tp jg resikonya. tdk pernah ada jaminan bhw #reksadana selalu akan profit (bahkan dlm jk panjang sekalipun).	jangan memikirkan untungnya aja, tapi juga resikonya. tidak pernah ada jaminan bahwa #reksadana selalu akan profit (bahkan dalam jk panjang sekalipun).
2	ihsg anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.	ihsg anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.
3	ada pandemi corona danareksa mawar fokus 10 jadi pilihan #investasi #reksadana #saham	ada pandemi corona danareksa mawar fokus 10 jadi pilihan #investasi #reksadana #saham
4	wahai generasi milenial, kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu	wahai generasi milenial, kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu

3. Tokenisasi

Tokenisasi merupakan proses pemotongan teks menjadi bagian-bagian yang disebut token, yaitu

sebuah *instance* dari urutan karakter dalam beberapa dokumen tertentu yang dikelompokkan bersama sebagai unit semantik yang berguna untuk diproses. Token bisa berupa paragraf, kalimat, frasa kata tunggal, sederhana, dan konsep. Teknik yang digunakan dalam proses tokenisasi adalah segmentasi dan memilah. Pada penelitian ini token yang dihasilkan berupa kata tunggal yang nantinya akan menjadi *term* yang akan digunakan sebagai penciri untuk klasifikasi sentimen pada Twitter. Berikut contoh tokenisasi pada data,

Tabel 4. **Proses Tokenisasi**

No.	Cuitan	Tokenisasi
1	jagan memikirkan untungnya aja, tapi juga resikonya. tidak pernah ada jaminan bahwa #reksadana selalu akan profit (bahkan dalam jk panjang sekalipun).	jangan memikirkan untungnya aja tapi juga resikonya tidak pernah ada jaminan bahwa reksadana selalu akan profit bahkan dalam jk panjang sekalipun
2	ihsg anjlok dalam sepekan lalu, hampir semua jenis #reksadana catatkan kinerja negatif.	ihsg anjlok dalam sepekan lalu hampir semua jenis reksadana catatkan kinerja negatif
3	ada pandemi corona danareksa mawar fokus 10 jadi pilihan #investasi #reksadana #saham	ada pandemi corona danareksa mawar fokus 10 jadi pilihan investasi

		reksadana saham
4	wahai generasi milenial, kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu	wahai generasi milenial kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu

4. Stopwords

Stopwords adalah metode untuk menyaring kata dalam dokumen untuk mendapatkan data yang berkualitas. *Stopwords* dapat berupa kata tanpa arti atau tidak mengandung informasi. Kata-kata pada *stopwords* dianggap terlalu sering berada diantara dokumen dan kata yang muncul dalam dokumen tersebut tidak dapat membantu pemahaman suatu dokumen dan tentunya harus dihilangkan. Contoh *stopwords* dalam Bahasa Indonesia seperti "yang", "juga", "dari", "dia", "kami", "kamu", "aku", "saya, "ini", "itu", "atau" dan lain sebagainya. *Stopwords* yang digunakan dapat melakukan sedikit perubahan pada list-nya seperti simbol tanda baca atau angka. Hasil dari penghapusan *stopwords* adalah kumpulan kata-kata yang mempengaruhi makna dari sebuah cuitan. Berikut adalah contoh *stopwords* pada data,

Tabel 5. **Proses *Stopwords***

No.	**Cuitan**	***Stopwords***
1	jangan memikirkan untungnya aja tapi juga resikonya tidak pernah ada jaminan bahwa reksadana selalu akan profit bahkan dalam jk panjang sekalipun	memikirkan untungnya aja tapi resikonya tidak ada jaminan reksadana akan profit jk panjang
2	ihsg anjlok dalam	ihsg anjlok sepekan

	sepekan lalu hampir semua jenis reksadana catatkan kinerja negatif	lalu jenis reksadana catatkan kinerja negatif
3	ada pandemi corona danareksa mawar fokus 10 jadi pilihan investasi reksadana saham	ada pandemi corona danareksa mawar fokus pilihan investasi reksadana saham
4	wahai generasi milenial kini saatnya dirimu berinvestasi masa muda saatnya pilih investasi terbaik untuk masa depanmu	wahai generasi milenial kini dirimu berinvestasi muda pilih investasi terbaik depanmu

3.2.3 Pembobotan TF-IDF

Pada proses ini dilakukan pengitungan banyaknya *term* atau kata yang muncul pada cuitan (tf), menghitung banyaknya cuitan yang mengandung *term* tersebut (df), menghitung *inverse* dokumen *frequency* (idf), dan mengalikan tf dengan idf sebagai bobot dari *term* pada setiap cuitan. Berikut contoh dari TF-IDF sesuai dengan contoh dataset sebelumnya,

Tabel 6. **Proses Pembobotan TF-IDF**

Term	**TF**				**DF**	**N/DF**	**IDF**	**TF-IDF**			
	d_1	d_2	d_3	d_4				d_1	d_2	d_3	d_4
memikirkan	1	0	0	0	1	4	0,60	0,60	0	0	0
untungnya	1	0	0	0	1	4	0,60	0,60	0	0	0

aja	1	0	0	0	1	4	0,60	0,60	0	0	0
tapi	1	0	0	0	1	4	0,60	0,60	0	0	0
resikonya	1	0	0	0	1	4	0,60	0,60	0	0	0
tidak	1	0	0	0	1	4	0,60	0,60	0	0	0
ada	1	0	1	0	2	2	0,30	0,30	0	0,30	0
jaminan	1	0	0	0	1	4	0,60	0,60	0	0	0
reksadana	1	1	1	0	3	1,3	0,11	0,11	0,11	0,11	0
akan	1	0	0	0	1	4	0,60	0,60	0,60	0	0
profit	1	0	0	0	1	4	0,60	0,60	0,60	0	0
jk	1	0	0	0	1	4	0,60	0,60	0,60	0	0
panjang	1	0	0	0	1	4	0,60	0,60	0,60	0	0
ihsg	0	1	0	0	1	4	0,60	0	0,60	0	0
anjlok	0	1	0	0	1	4	0,60	0	0,60	0	0
sepekan	0	1	0	0	1	4	0,60	0	0,60	0	0
lalu	0	1	0	0	1	4	0,60	0	0,60	0	0
jenis	0	1	0	0	1	4	0,60	0	0,60	0	0
catatkan	0	1	0	0	1	4	0,60	0	0,60	0	0
kinerja	0	1	0	0	1	4	0,60	0	0,60	0	0
negatif	0	1	0	0	1	4	0,60	0	0,60	0	0
pandemi	0	0	1	0	1	4	0,60	0	0	0,60	0
corona	0	0	1	0	1	4	0,60	0	0	0,60	0
danareksa	0	0	1	0	1	4	0,60	0	0	0,60	0
mawar	0	0	1	0	1	4	0,60	0	0	0,60	0
fokus	0	0	1	0	1	4	0,60	0	0	0,60	0
pilihan	0	0	1	0	1	4	0,60	0	0	0,60	0
investasi	0	0	1	1	2	2	0,30	0	0	0.30	0.30
saham	0	0	1	0	1	4	0,60	0	0	0,60	0
wahai	0	0	0	1	1	4	0,60	0	0	0	0,60
generasi	0	0	0	1	1	4	0,60	0	0	0	0,60
milenial	0	0	0	1	1	4	0,60	0	0	0	0,60
kini	0	0	0	1	1	4	0,60	0	0	0	0,60
dirimu	0	0	0	1	1	4	0,60	0	0	0	0,60
berinvestasi	0	0	0	1	1	4	0,60	0	0	0	0,60
muda	0	0	0	1	1	4	0,60	0	0	0	0,60

pilih	0	0	0	1	1	4	0,60	0	0	0	0,60
terbaik	0	0	0	1	1	4	0,60	0	0	0	0,60
depanmu	0	0	0	1	1	4	0,60	0	0	0	0,60

3.2.4 Klasifikasi dengan pendekatan *Support Vector Machine*

Setelah dilakukan pembobotan maka tahap selanjutnya adalah melakukan pembagian set data *training* dan *testing* untuk selanjutnya dilakukan proses klasifikasi. Pembagian set data *training* dan *testing* dibagi sebesar 20% data *testing dan 80% data training*. Pada penelitian ini *Support Vector Machine* dan pembobotan TF-IDF diimplementasikan dengan menggunakan *Scikit-Learn. Scikit-Learn* sudah teruji dan memiliki dokumentasi yang super lengkap. Bahkan kontributornya pun banyak. *Scikit-Learn* pun menyediakan ekstensi untuk *fuzzy logic* dan *computer vision.* Hasil Analisa sentiment menggunakan metode *Support Vector Machine* akan di evaluasi dengan dihitung nilai akurasi , presisi, *recall* serta *f1-score*nya.

3.2.5 Evaluasi Performansi

Setelah pengguna Twitter diklasifikasi, selanjutnya dilakukan evaluasi dari performa sistem atau metode yang digunakan. Pada penelitian ini, evaluasi performansi diukur menggunakan *confusion matrix. Penggunaan confusion matrix* untuk pengukuran performansi pada penelitian ini karena *confusion matrix* menunjukkan bagaimana model ketika membuat prediksi. Tidak hanya memberi informasi tentang kesalahan yang dibuat oleh model tetapi juga jenis kesalahan yang dibuat. Setiap kolom dari *confusion matrix* merepresentasikan *instance* dari kelas prediksi. Setiap baris dari *confusion matrix* mewakili *instance* dari kelas aktual. *Confusion matrix* pada penelitian ini menggunakan *library scikit-learn. Performance metrics* yang umum atau sering digunakan pada *confusion matrix* adalah *accuracy* (akurasi), *presicion* (presisi), *recall*, dan *f-measure* dengan penjelasan mengacu pada tabel *confusion matrix* sebagai berikut,

Tabel 7. ***Confussion Matrix***

Kelas Sebenarnya	**Kelas Prediksi**	
	True	***False***
True	*True Positive* (TP)	*False Positive* (FP)
False	*True Negative* (TN)	*False Negative* (FN)

dimana, *True Positive* (TP) adalah jumlah data positif yang terklasifikasi benar oleh sistem. *True Negative* (TN) adalah jumlah data negatif yang terklasifikasi benar oleh sistem. *False Positive* (FP) adalah jumlah data positif namun terklasifikasi salah oleh sistem. Terakhir, *False Negative* (FN) adalah jumlah data negatif namun terklasifikasi salah oleh sistem.

1. *Accuracy* (Akurasi)
   Akurasi adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual. Rumus akurasi adalah sebagai berikut,

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

2. *Presicion* (Presisi)

Presisi adalah tingkat ketepatan antara informasi yang diminta dan jawaban yang diberikan oleh sistem. Rumus presisi adalah sebagai berikut.

$$precision = \frac{TP}{TP + FP} \quad (8)$$

3. *Recall*

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali suatu informasi. Rumus *recall* adalah sebagai berikut

$$recall = \frac{TP}{TP+FN} \quad (9)$$

4. *F1-Measure* (*F1 Score*)

F1-Measure (*F1 Score*) adalah penyetaraan nilai *precision* dan *recall* karena kedua nilai tersebut biasanya memiliki perbedaan yang cukup tinggi. Rumusnya adalah sebagai berikut.

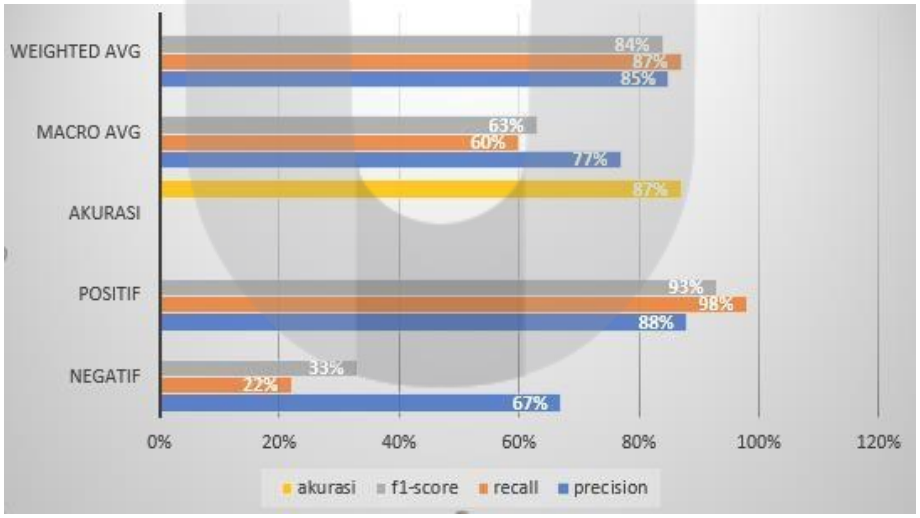
$$F1\text{-}Measure = \frac{2 \times (recall \times precission)}{(recall + precission)} \quad (10)$$

4. Evaluasi

Setelah dilakukan evaluasi dengan akurasi, presisi, *recall*, dan *f-measure* (*f-score*), hasil dan analisis dari evaluasi penelitian ini adalah sebagai berikut.

4.1 Hasil Pengujian

Pengujian dari penelitian ini dilakukan dengan menggunakan *confusion matrix* yaitu terdiri dari akurasi, *recall*, presisi dan f1-*score*. Berikut adalah grafik hasil dari evaluasi pada penelitan ini,



Gambar 2. **Hasil Evaluasi**

Hasil klasifikasi yang diperoleh dengan pendekatan *Support Vector Machine* ini memiliki akurasi sebesar 0,87 atau 87%, presisi positif sebesar 0.88 atau 88%, presisi negatif 0,67 atau 67%, *recall* positif sebesar 0,98 atau 98%, *recall* negatif sebesar 0,22 atau 22%, *f1-score* positif sebesar 0,93 atau 93% dan *f1-score* negatif sebesar 0,33 atau 33%. Hasil akurasi yang didapatkan cukup tinggi sehingg dapat dikatakan bahwa performansi dari model klasifikasi yang telah dibangun dapat dikatakan cukup baik.

5. Kesimpulan

Penelitian ini dilakukan untuk mengetahui kecendrungan sentimen masyarakat atau komunitas investor mengenai masalah investasi di Indonesia dengan metode *Support Vector Machine*. Penelitian ini menggunakan data cuitan yang mengandung *hashtag* #investasimilenial, #investasi dan #bursaefekindonesia dengan total data 300 cuitan. Berdasarkan hasil yang diproleh dapat dilihat bahwa sentimen masyarakat mengenai masalah investasi di Indonesia cenderung positif. Penelitian ini menghasilkan tingkat akurasi yang cukup baik yaitu sebesar 87% tetapi akan lebih baik lagi jika akurasi diatas 90%, namun hasil yang di dapat bisa dikatakan bahwa klasifikasi sentimen menggunakan *Support Vector Machine* mampu mengklasifikasi sentimen dengan baik.

Saran untuk penelitian selanjutnya adalah dengan menambahkan jumlah data *training* untuk menghasilkan model *classifier* yang lebih baik. Penelitian berikutnya juga diharapkan dapat menggunakan dua atau lebih model klasifikasi sebagai pembandinng untuk mengetahui metode mana yang lebih baik dari metode *Support Vector Machine.*

Daftar Pustaka

[1] T. J. Bristol. Twitter: Consider the possibilities for continuing nursing education. *The Journal of Continuing Education in Nursing*, pages 199–200, 2010.

[2] B. Liu, "Sentiment Analysis and Subjectivity," *Chicago: University*, 2010.

[3] N. C. a. J. Taylor, "An Introduction to Support Vector Machine and Other Kernel-based Learning Methods," *Cambridge:Cambridge University Press*, 2000.

[4] W. Lippo, "Support Vector Machines," *Theory and Application.Singapore*, 2005.

[5] S. E. G. S. Vladimir Vapnik, "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing," 1997.

[6] E. L. X. d. S. Y. Haddi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science* , 2013.

[7] Feldman, R. dan James, S. (2007). *The Text Mining Handbook*. New York: Cambridge.

[8] Berry, M. W., dan J. Kogan. *Text Mining Application and Theory*. WILEY. United Kingdom. 2010.

[9] Kharde, V.A. & Sonawane, S., 2016. *Sentiment Analysis of Twitter Data: A Survey of Techniques*. International Journal of Computer Applications.

[10] Vidya, N.A., Fanany, M.I., & Budi, I., 2015. *Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers*. Procedia Computer Science.

[11] Mantyla, M.V., Graziotin, D., Kuutila, M., 2018. *The evolution of sentiment analysis*—A review of research topics, venues, and top cited papers. Computer Science Review.

[12] Luqyana, W., Cholissodin, I., & Perdana, R., 2018. *Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine*. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.

[13] B. Santosa, Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[14] Turban, E.; et.al. Decision Support and Business Intelligence Systems (edisi ke-9). New Jersey: Pearson Education, Inc. 2011.

[15] R. Feldman, and J. Sanger, “The Text Mining Handbook Addvances Approaches in Analyzing Unstruktured Data.” Cambridge University Press, New York, 2007.