

BAB I

PENDAHULUAN

1.1. Latar Belakang

Istilah investasi mengacu pada investasi pada aset atau kekayaan yang dikumpulkan dalam suatu opsi investasi dengan harapan keuntungan di masa depan [1]. Berdasarkan data KSIE per September 2023 berdasarkan single investor identifikasi (SID), jumlah investor pasar modal di Indonesia mencapai 11,72 juta, dengan investor reksa dana naik 14,47%, investor surat berharga negara (SBN) naik 15,45%, dan saham investor naik 13,27%. Rinciannya, investor pasar modal meliputi 10,99 juta investor reksa dana, 5,02 juta investor saham, dan 959 ribu investor surat berharga nasional (SBN). Jika data ini digabungkan dengan SID peserta Tabungan Perumahan Rakyat (TAPERA), maka total SID-nya sebanyak 16 juta. Menurut data KSIE, pertumbuhan pasar modal Indonesia meningkat signifikan, dari 9,2 juta pada tahun 2020 menjadi 28,7 juta pada September 2023, termasuk investasi pada saham, reksadana, dan surat berharga negara (SBN) [2]. Berdasarkan data tersebut, minat masyarakat untuk berinvestasi cukup kuat sehingga banyak bermunculan aplikasi investasi digital. Menurut website CNBC ada beberapa aplikasi investasi terbaik yang telah diawasi oleh OJK yaitu SimInvest, HSB Investasi, Bareksa, Pintu.[1]. Berikut ini merupakan hasil dari visualisasi perbandingan jumlah review di google play.



Gambar 1. 1 Perbandingan Grafik Aplikasi

Dapat dilihat pada gambar 1.1. bahwasanya bareksa memiliki review yang lebih rendah daripada hsb investasi dan pintu tetapi jika dilihat dari rating bareksa unggul sebesar 4.6/5 sedangkan HSB investasi memiliki rating 4.2/5, Pintu memiliki rating 4.1/5. Namun banyaknya aplikasi investasi terbaik membutuhkan informasi yang sangat banyak dan sulit untuk menarik kesimpulan karena banyak pendapat yang berbeda. Untuk memperoleh informasi tersebut, harus dilakukan review data pada aplikasi. Pada pembelajaran *computer science* masalah tersebut bisa diartikan analisis sentimen untuk metode yang bisa digunakan dalam analisis sentimen yakni Naive Bayes Classifier, Support Vector Machine dan Maximum Entropy [2].

Pada paragraf ini terdapat beberapa penelitian yang telah melakukan dengan menggunakan beberapa metode untuk melakukan sebuah pelabelan serta penggunaan metode SVM. Reza Hermansyah *et al* [3] melakukan sebuah penelitian dengan menganalisis produk dan evaluasi pelayanan dalam PT Telekomunikasi Indonesia dengan menggunakan *TextBlob* untuk pelabelan dan *naïve bayers* & K-NN. Penelitian ini menghasilkan akurasi *TextBlob* 54,67%, *naïve bayers* 69,44% dan K-NN sebesar 75% [3]. Sri Lestari *et al*[4], melakukan analisis sentimen terhadap aplikasi saham di *GoolePlay* Store dengan menggunakan algoritma *Support Vector Machine* dengan menggunakan bantuan *Rapid Miner* untuk melakukan pelabelan dengan label positif, negatif dan netral tanpa melakukan ekstraksi fitur. Dalam penelitian ini terdapat 5 aplikasi yang di analisis beserta hasilnya yakni HSB Investasi 88,70%, Ajaib 61,89%, Pluang 68,25%, stockbit 66,95% dan bibit 64,89%. Peneliti menyatakan bahwasanya rapid miner sangat membantu dan mudah untuk digunakan berdasarkan ulasan [4]. Pada tahun 2021, Prasoon Gupt *et al* [5], melakukan sebuah penelitian sentiment analysis terhadap kegiatan Lockdown di India selama covid-19 dengan dataset dari twitter dengan menggunakan *TextBlob* dan Vader Lexicon sebagai pelabelan dengan menggunakan ekstraksi fitur *CountVectorizer* atau biasa dikenal sebagai *Bag Of Words*. Penelitian ini menghasilkan akurasi yang cukup baik yakni 84,4% [5]. Penelitian sama yang dilakukan M. N. Muttaqin *et al* [6], dengan melakukan

perbandingan metode analisis sentimen aplikasi gojek dengan menggunakan lexicon untuk melakukan pelabelan serta menggunakan ekstraksi dengan menggunakan algoritma SVM dan K-NN. Penelitian ini menghasilkan 2 akurasi yang cukup baik K-NN 82,14% dengan K=22 dan SVM 87,98% menggunakan parameter $c=1$. Peneliti menyimpulkan bahwasanya SVM melakukan klasifikasi secara lebih baik di bandingkan K-NN [6]. Dari beberapa penelitian diatas bahwasanya kombinasi metode *Support Vector Machine* dengan penggunaan *vader lexicon* menghasilkan hasil yang cukup baik.

Pada beberapa penelitian terkait dibawah ini memuat tentang penggunaan metode seleksi fitur antaralain yakni Gientry Rachma Ditami *et al* [7], dalam penelitian ini menggunakan TI-IDF sebagai ekstraksi fitur menggunakan algoritma *Support Machina Vector* (SVM) dengan bantuan grid search untuk menentukan parameter terbaik dalam pengujian. Dalam penelitian ini membandingkan event yang dimiliki oleh aplikasi Tokopedia dan Shopee dan menghasilkan akurasi yakni tokopedia 66,23% naik menjadi 67,67%, shopee 66,93% menjadi 67,47 kenaikan tersebut karena menggunakan grid search [7]. Ruba Obiedat *et al* [8], melakukan penelitian menggunakan teknik *crowdsourcing* untuk pelabelan dan menggunakan 2 ekstraksi fitur yakni N-gram dan *Bag Of Words*. Peneliti juga menambahkan SVM PSO dan SMOTE untuk mendapatkan hasil terbaik. Peneliti juga membandingkan beberapa metode yakni SVM *default*, KNN, Naïve Bayes dan 4 lainnya. Penelitian ini menghasilkan akurasi yang paling bagus yakni SVM dengan bantuan PSO dan Smote sebesar 89% [8]. M. D. Purbolaksono *et al* [9], membandingkan seleksi fitur yang berbeda yakni gini index dan chi square dengan menggunakan ekstraksi fitur TF-IDF dan menghasilkan akurasi sebesar 85.8% untuk Gini Index dan 89.2% untuk Chi Square [9]. Maka dari itu kombinasi yang terbaik yakni SVM dengan Chi Square karena SVM memiliki kelemahan dalam pemilihan fitur. Pada tahun yang sama, P. R. B. Putra *et al* , melakukan penelitian dengan menggunakan TF-IDF sebagai ekstraksi fiturnya dan memberikan hasil yang terbaik yakni akurasi 93,06%, presisi 92,11%, *recall* 93,06%, dan *f1 score* 93,04% [10]. V. Nurcahyawati *et al* [11] Peneliti membandingkan analisis sentimen dengan manual notasi SVM

tanpa bantuan vader lexicon mendapatkan akurasi 86,% presisi 86.06%, recal 95.43% dan *f1 score* 90.51%. Sedangkan hasil dari analisis sentimen dengan menggunakan bantuan vader lexicon meningkat menjadi akurasi 86.57%, presisi 89.71%, recal 96.32%, dan *f1 score* 92.89% [11]. Dari penelitian terkait yang telah dijelaskan bahwasanya penggunaan kombinasi seleksi fitur chi square memiliki hasil yang cukup baik.

Dari hasil penelitian tersebut dapat peneliti simpulkan bahwasanya metode *Support Vector Machine* merupakan metode yang paling cocok digunakan untuk penelitian ini dengan menggunakan *Vader Lexicon Based* dengan ekstraksi fitur TF-IDF serta seleksi fitur menggunakan fitur *Chi Square* dapat menghasilkan performa yang optimal jika dibandingkan dengan metode klasifikasi lainnya dalam pengklasifikasian data teks.

1.2. Rumusan Masalah

1.2.1. Permasalahan

Bareksa merupakan salah satu aplikasi investasi yang terbanyak dipakai oleh masyarakat indonesia dan sudah terverifikasi aman oleh pemerintah. Hal ini tentunya banyak menuai tanggapan oleh masyarakat Indonesia khususnya pengguna aplikasi bareksa. Ada banyak komentar yang didapatkan aplikasi bareksa baik itu positif maupun negatif. Komentar dan tanggapan tersebut bisa menjadi bahan evaluasi untuk bareksa sendiri, tetapi tidak adanya sistem yang bisa menganalisis hal tersebut. Terkait permasalahan ini dibutuhkan suatu sistem sentiment analisis guna memberikan referensi untuk menjadi bahan evaluasi. Pemilihan algoritma *support vector machine* memiliki akurasi yang cukup baik. Akan tetapi, algoritma tersebut memiliki kekurangan dalam pemilihan fitur, oleh karena itu dibutuhkan *chisquare* untuk mendapatkan performa yang optimal.

1.2.2. Metode Usulan

Berdasarkan hal tersebut dalam penelitian ini melakukan analisis sentimen dengan lexicon based sebagai alat untuk pelabelan serta menggunakan SVM dengan seleksi fitur *Chi Square*.

1.2.3. Pertanyaan Penelitian

Berdasarkan permasalahan dan metode usulan dalam penelitian ini didapatkan pertanyaan penelitian, seperti berikut :

- a. Berapa nilai peningkatan performa *accuracy* pada metode *Support Vector Machine* dengan menggunakan seleksi fitur *chi square* ?
- b. Berapa nilai peningkatan performa *accuracy* pada metode *Support Vector Machine* dengan menggunakan metode *smote*?
- c. Berapa nilai peningkatan performa *accuracy* pada metode *Support Vector Machine* dengan menggunakan metode *smote* serta seleksi fitur *chi square*?

1.3. Tujuan dan Manfaat

1.3.1. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut :

1. Mengetahui nilai peningkatan performa *accuracy* pada metode *support vector machine* dengan menggunakan seleksi fitur *square*
2. Mengetahui nilai peningkatan performa *accuracy* pada metode *support vector machine* dengan menggunakan metode *smote*
3. Mengetahui nilai peningkatan performa *accuracy* pada metode *support vector machine* dengan menggunakan metode *smote* serta seleksi fitur *chi square*

1.3.2. Manfaat Penelitian

Adapun manfaat dilakukannya penelitian ini adalah :

1. Dapat mengetahui peningkatan performa metode *support vector machine* dengan menggunakan metode *smote* dan seleksi fitur *chi square*
2. Dapat menjadi bahan referensi untuk bahan evaluasi aplikasi bareksa dalam mempertahankan dan meningkatkan pelayanan aplikasi bareksa
3. Dapat menjadi referensi dan sebagai rujukan untuk penelitian serupa.

1.4. Batasan Masalah

Berikut adalah batasan-batasan penelitian yang dilakukan dalam penelitian ini :

1. Data dalam penelitian ini adalah data ulasan aplikasi bareksa yakni PT. Bareksa Marketplace Indonesia dengan jumlah 2008 ulasan yang diambil pada tahun 2023 melalui google play.
2. Pengambilan data dengan menggunakan library *google-play-scapper*
3. Pelabelan dilakukan dengan menggunakan *Vader Lexicon* dengan bantuan kamus lexicon Indonesia.
4. Metode yang digunakan pada penelitian ini yaitu metode *Support Vector Machine* dengan seleksi fitur *chi square* dan tidak dengan algoritma lain.
5. Pengklasifikasian sentiment dijadikan ke dalam 2 Kategori yaitu sentiment positif sebanyak 1240 ulasan dan sentiment negatif sebanyak 977 ulasan.
6. Bahasa Pemrograman menggunakan bahasa pemrograman *Python*.

1.5. Sistematika Penulisan

Berikut merupakan beberapa aturan dalam penulisan yang digunakan dalam penyusunan laporan. Adapun sistematika proposal yang digunakan terdiri dari beberapa bagian yaitu :

BAB I PENDAHULUAN

Bab ini membahas latar belakang usulan, rumusan masalah, tujuan dan manfaat, batasan masalah, dan sistematika.

BAB II TINJAUAN PUSTAKA

Bab ini membahas mengenai beberapa teori untuk menunjang penelitian, menguraikan tinjauan pustaka, penelitian terdahulu, serta teori penunjang lainnya sebagai bahan acuan.

BAB III METODE USULAN

Pada bab ini akan menguraikan arsitektur sistem yang digunakan dalam penelitian ini, yang melibatkan beberapa proses penting. Dimulai dari pengumpulan data melalui proses scrapping, lalu proses *text preprocessing*, serta menguraikan tahapan pada seleksi fitur menggunakan *chi square* dan metode smote serta proses klasifikasi menggunakan *support vector machine*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil penelitian yang telah dilakukan dan analisis terhadap hasil penelitian yang didapatkan.

BAB V PENUTUP

Bab ini berisi kesimpulan yang merangkum dan menjawab keseluruhan penelitian serta saran yang memberikan pandangan lebih lanjut.