ISSN: 2339-2541
JURNAL GAUSSIAN Volume 11, Nomor 4, Tahun 2022, Halaman 542 – 553
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

ANALISIS SENTIMEN PADA ULASAN APLIKASI INVESTASI *ONLINE* AJAIB PADA *GOOGLE PLAY* MENGGUNAKAN METODE *SUPPORT VECTOR MACHINE* DAN *MAXIMUM ENTROPY*

Fath Ezzati Kavabilla[1*], Tatik Widiharih[2], Budi Warsito[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
***e-mail**: fathezzatikavabilla@students.undip.ac.id

DOI: 10.14710/j.gauss.11.4.542-553

Article Info:
Received: 2022-10-08
Accepted: 2022-12-09
Available Online: 2023-02-25

Keywords:
Investment; Ajaib; Sentiment Analysis; Support Vector Machine (SVM); Maximum Entropy.

Abstract: Investment is money or asset to earn profits in the future. Online investment applications are already available, one of which is Ajaib. A review of Ajaib's application is needed to find out reviews given are positive or negative. Sentiment analysis in Ajaib is used to see the user's response to Ajaib's performance which is divided into positive and negative classes. Sentiment analysis of the Ajaib's reviews classification can be used with the Support Vector Machine and Maximum Entropy methods. Support Vector Machine on non-linear problems inserts the kernel into a high-dimensional space, to find a hyperplane that can maximize the distance between classes. The kernel used in SVM is the Radial Basis Function (RBF) kernel with gamma parameters of 0.002 and Cost (C) of 0.1; 1; 10. Maximum Entropy is a classification technique that uses the entropy value to classify data with the evaluation model used, namely 5-fold cross-validation. The algorithm which has the highest accuracy and kappa statistics is the best algorithm for classifying the sentiments of Ajaib users. The results using the Support Vector Machine algorithm show the overall accuracy is 85.75% and the kappa accuracy is 58.07%. The results using the Maximum Entropy algorithm show an overall accuracy of 83% and kappa accuracy of 50.5%. This shows that sentiment using the Support Vector Machine has a better performance than Maximum Entropy.

1. PENDAHULUAN

Pemahaman tentang investasi pada penduduk Indonesia masih sangat minim, dengan dijelaskan sekitar 2,4% dari penduduk Indonesia merupakan investor dalam pasar modal dan akan terus bertambah (BeritaSatu, 2021). Angka 2,4% dari penduduk Indonesia ini dapat dijadikan peluang bagi pasar modal untuk membantu membangun perekonomian negara Indonesia. Angka 2,4% juga menunjukkan bahwa sebagian besar masyarakat Indonesia kurang memahami pentingnya investasi untuk masa depan, dan beberapa orang masih khawatir untuk melakukan investasi. Beberapa instrumen pada pasar modal memiliki cara sendiri untuk memberikan pengaruh lebih pada kemajuan ekonomi suatu negara.

Salah satu memaksimalkan perkembangan teknologi, dengan adanya fasilitas yang mempermudah investasi dalam bertransaksi dan mengambil keputusan calon investor atau disebut *online trading*. Menunjang *online trading*, banyak bermunculan aplikasi investasi *online* di Indonesia. Salah satu aplikasi investasi *online* adalah Ajaib, Ajaib merupakan pemain baru dalam aplikasi investasi *online* yang bukan merupakan akuisisi dari perusahaan lain. Setiap pengguna memiliki respon yang berbeda terhadap performa Ajaib. Berdasarkan pernyataan sebelumnya, penulis ingin mengetahui bagaimana respon pengguna Ajaib terhadap performa Ajaib dengan menganalisis ulasan pengguna Ajaib di *Google Play*.

Penelitian ini difokuskan pada klasifikasi sentimen yang mengacu pada ulasan pengguna Ajaib pada *Google play* ke dalam dua kelas yaitu sentiment positif dan negatif. Analisis sentimen dilakukan dengan 2 metode yaitu *Support Vector Machine* (SVM) dan *Maximum Entropy*. Algoritma SVM ini sudah dilakukan pada banyak penelitian, dan disimpulkan memiliki tingkat akurasi terbaik dari algoritma yang lain pada *machine learning*. Algoritma SVM dimodifikasi dengan memasukkan fungsi *kernel* RBF, yang membuat SVM dapat beroperasi dalam ruang berdimensi tinggi. Selanjutnya Algoritma *Maximum Entropy* merupakan algoritma yang menggunakan probabilitas kata spesifik untuk mendapatkan nilai probabilitas.

2. TINJAUAN PUSTAKA

Menurut Tandelilin (2010), investasi sebagai keharusan terhadap beberapa sumber dana yang terdapat pada saat ini dengan tujuan untuk mendapatkan keuntungan di masa yang akan datang. Banyaknya keinginan investasi secara mudah, sekarang bermunculan aplikasi investasi secara *online* atau daring. Ajaib merupakan sebuah *platform* yang dirancang untuk melakukan aktivitas investasi secara *online* seperti jual beli saham dan reksadana yang saat ini sedang popular di berbagai kalangan. *Platform* Ajaib ini telah dirancang dan dijalankan oleh Ajaib Reksadana atau PT. Takjub Teknologi Indonesia yang telah berdiri pada tahun 2019, dan merupakan aplikasi investasi *online* yang tergolong baru. Ajaib dapat diunduh melalui toko aplikasi seperti *Google play*. *Google play* merupakan salah satu fasilitas yang dimiliki secara digital oleh *Google*, yang di dalamnya berupa toko dengan layanannya untuk memasarkan produk berupa aplikasi, permainan, buku hingga hiburan seperti musik, film, dan televisi.

Analisis sentimen merupakan salah satu studi komputasi yang bertujuan untuk menganalisis pendapat serta emosi seseorang yang dapat diekspresikan ke dalam teks (Liu, 2010). Analisis sentimen digunakan untuk mengetahui sikap dari individu maupun kelompok terhadap suatu topik bahasan. Sikap tersebut dapat berupa komentar serta penilaian yang dapat dijadikan bahan evaluasi. Analisis sentimen juga digunakan untuk mengidentifikasi keluhan, opini terhadap produk atau layanan, dan opini terhadap suatu merek. Analisis sentimen cenderung berfokus pada pendapat atau opini yang menyatakan suatu sentiment ke dalam positif dan negatif. Analisis sentimen dilakukan dengan beberapa tahapan untuk mendapatkan hasil pengujian yang terbaik, berikut tahapan untuk mempersiapkan data sebelum tahapan klasifikasi.

a. *Text Pre-processing*

   Text pre-processing merupakan salah satu tahap dalam *text mining* yang bertujuan untuk mempersiapkan data pada proses utama dengan menyeragamkan data dengan mengubah data teks yang tidak terstruktur serta menghilangkan *noise*. *Text pre-processing* pada penelitian ini menggunakan *case folding, remove emoticon, remove punctuation, remove number, strip whitespace,* dan normalisasi kata.

b. *Sentiment Scoring*

   Sentiment scoring adalah tahapan dimana sebuah pernyataan yang disiapkan akan diberikan label, dengan dua label yaitu positif dan negatif. Pada penelitian ini dibutuhkan kumpulan kata positif dan kata negatif berbahasa inggris yang telah di tranlasi terlebih dahulu.

c. *Feature Selection*

   Feature selection ini adalah tahapan yang bertujuan untuk mengurangi dimensi dalam proses transformasi teks sehingga hasil dari *text mining* memiliki kualitas yang baik. *Feature selection* yang dilakukan pada penelitian ini adalah *stopwords removal,*

stemming, dan *tokenizing*. *Stopwords removal* adalah tahap menghilangkan kata-kata yang tidak memiliki makna pada isi kalimat. *Stemming* adalah tahap menghilangkan imbuhan kata untuk mendapatkan kata dasarnya. *Tokenizing* adalah tahap memisahkan setiap kata pada sebuah kalimat.

Support Vector Machine (SVM) upaya untuk mendapatkan *hyperplane* terbaik sebagai pemisah antara dua buah kelas. Cara yang digunakan yaitu memaksimalkan jarak antar kelas, SVM dapat menemukan fungsi *hyperlane* terbaik. SVM dapat digunakan pada dua jenis permasalahan dengan data set yang berbeda, yaitu data set yang terpisah secara linier dan yang tidak dapat terpisah secara linier dalam ruang input.

Pembobotan kata dengan TF-IDF digunakan dalam klasifikasi SVM untuk mendapatkan fungsi *kernel* yang digunakan. Pembobotan kata adalah proses guna memberikan bobot pada setiap kata dalam proses mencari informasi dari beberapa dokumen yang heterogen. Pembobotan kata pada penelitian ini menggunakan metode *Term Frequency – Inverse Document Frequency* (TF-IDF). Pembobotan menggunakan TF-IDF dihitung dengan persamaan berikut.

$$W(i,j) = \frac{n_{i,j}}{\sum_{i=1}^{p} n_{i,j}} . log_2 \frac{D}{d_i} \quad (1)$$

dengan:

$W(i,j)$: Pembobotan TF-IDF untuk *term* ke-*i* pada dokumen *j*
$n_{i,j}$: Jumlah kemunculan *term* ke-*i* dalam dokumen ke-*j*
$\sum_{i=1}^{p} n_{i,j}$: Jumlah kemunculan seluruh *term* pada dokumen ke-*j*
p : Banyaknya *term* yang terbentuk
D : Banyaknya dokumen yang digunakan
d_i : Banyaknya dokumen yang mengandung *term* ke *i*

Linearly Separable data adalah data yang dapat dipisahkan secara linier. Terdapat himpunan $X = \{x_1, x_2, \dots, x_l\}$ dengan data latih yang dinotasikan $x_i \in R$ dan label pada setiap kelas dari x_i dinotasikan $y_i \in \{-1, +1\}$ untuk $i = 1,2,3, \dots l$ dengan l merupakan banyaknya data. Asumsi kedua kelas berhasil dipisahkan dengan *hyperplane* secara sempurna, didefinisikan (Nugroho dkk, 2003):

$$(\boldsymbol{w^T}.\boldsymbol{x_i}) + b = 0 \quad (2)$$

dengan:

$\boldsymbol{w}$: Parameter bobot berupa vector berukuran $l \times 1$
$\boldsymbol{x_i}$: Data *training* yang tersedia berupa vector berukuran $l \times 1$
l : Banyaknya data *training*
b : Nilai bias atau error yang konsisten dalam menentukan nilai

Terdapat kasus yang sering terjadi dikarenakan data yang tidak dapat terpisahkan secara linear (*nonlinearly separable data)*, disimulasikan bahwa tidak dapat terpisah secara sempurna pada *input space*. maka SVM dirumuskan kembali dengan teknik *soft margin*, dilakukan dengan mengubah persamaan pada *quadratic programing* (QP) dengan menambahkan variabel kendur (*slack variabel*) $\xi_i \geq 0$. Didapat *hyperplane* yang telah dimodifikasi berikut (Gunn, 1998):

$$y_i[(\boldsymbol{w^T}.\boldsymbol{x_i}) + b] \geq 1 - \xi_i \, , i = 1,2,3, \dots l, \xi_i \geq 0$$
$$min_{\boldsymbol{w}} = \tau(\boldsymbol{w}) = \frac{1}{2} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{l} \xi_i \quad (3)$$

Parameter *Cost* (C) yang digunakan untuk mengontrol *trade-off* antara pinalti *variabel slack* dengan *margin*. Parameter *Cost* (C) yang optimal ditentukan dengan *trial and error*, jika nilai C semakin tinggi maka besar pelanggaran klasifikasi. Bentuk primal sebagai berikut.

$$L_p(\mathbf{w}, b, \alpha, \xi, \beta) = \frac{1}{2}\,||\mathbf{w}||^2 + C\sum_{i=1}^{l}\xi_i \sum_{i=1}^{l}\alpha_i[y_i((\boldsymbol{w^T}.\boldsymbol{x_i}) + b) - 1 + \xi_i] \tag{4}$$

dengan α_i dan β_i merupakan *lagrange multiplier*. Optimasi *lagrange multiplier* dengan meminimalkan fungsi *lagrange* (L_p) terhadap $\boldsymbol{w}$,ξ,b disamadengankan, sehingga diperoleh sebagai berikut.

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \sum_{i=1}^{l}\alpha_i y_i\,\boldsymbol{x_i} \tag{5}$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{l}\alpha_i y_i = 0 \tag{6}$$

$$\frac{\partial L}{\partial \xi} = 0 \rightarrow C - \alpha_i - \beta_i = 0 \rightarrow C = \beta_i + \alpha_i \tag{7}$$

diperoleh persamaan dual problem (L_D) yang dimaksimalkan

$$\max L_D = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(\boldsymbol{x_i}^T\boldsymbol{x_j}) \tag{8}$$

dengan batas, $0 \le \alpha_i \le C\ dan\ \sum_{i=1}^{l}\alpha_i\, y_i = 0\,; i = 1,2,3\ldots, l$

Kasus dalam SVM dapat ditemukan data yang tidak terpisahkan secara linier, dengan menggunakan teknik *kernel* permasalahan non linier dapat terselesaikan (Cortes dan Vapnik, 1995). Penggunaan teknik kernel dengan memetakan dari ruang dimensi rendah menuju ruang berdimensi tinggi (*feature space*). Dalam SVM non linier data $\boldsymbol{x_i}$ dipetakan oleh fungsi *kernel* $\phi(\boldsymbol{x_i})$ ke ruang vector yang memiliki dimensi lebih tinggi. Proses selanjutnya bergantung pada *dot product* dari data yang telah ditransformasikan pada ruang berdimensi tinggi yaitu $\phi(\boldsymbol{x_i}).\phi(\boldsymbol{x_j})$. Dikarenakan transformasi ϕ sulit dipahami, maka *dot product* digantikan fungsi *kernel* yang dinotasikan sebagai berikut:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_i}).\phi(\boldsymbol{x_j}) \tag{9}$$

Salah satu fungsi *kernel* yang digunakan dalam SVM adalah *kernel Radial Basis Function* (RBF) sebagai berikut.

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp(-\gamma\left|\left|\boldsymbol{x_i} - \boldsymbol{x_j}\right|\right|^2) \tag{10}$$

dengan x_i, x_j merupakan pasangan data latih, dan ϕ merupakan fungsi pemetaan dari *inner space* kedalam *feature space*. Parameter γ (*gamma*) merupakan suatu parameter positif dan jika nilai *gamma* makin besar maka kurva juga makin sempit. Proses klasifikasi pada sebuah objek data *x* diformulasikan sebagai berikut (Nugroho dkk, 2003).

$$f(\Phi(x)) = sign\left(\sum_{i=1}^{p}\alpha_i\, y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b\right) \tag{11}$$

$$f(\Phi(x))\begin{cases} 1, jika \sum_{i=1}^{p} \alpha_i\, y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b \geq 0 \\ -1, jika \sum_{i=1}^{p} \alpha_i\, y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b < 0 \end{cases}$$

dengan:

$f(\Phi(x))$: Hasil klasifikasi dari data $\boldsymbol{x}$
y_i : Kelas data
α_i : Koefisien *lagrange*
$K(\boldsymbol{x_i}, \boldsymbol{x_j})$:Fungsi *kernel* data latih dan data uji
b : Bias
p : Banyaknya *support vector*

Maximum entropy salah satu teknik klasifikasi yang menggunakan nilai *entropy* untuk mengklasifikasikan suatu data. *Entropy* digunakan untuk mengukur tingkat keberagaman dari beberapa sampel data (Sabily dkk, 2019). Metode tersebut digunakan untuk mendapatkan probabilitas distribusi yang memiliki nilai *entropy* tinggi, dengan probabilitas yang didapat merupakan prediksi dari suatu kalimat yang terdapat informasi dari kalimat tersebut. Fungsi umum *entropy* adalah sebagai berikut (Alroy, 2019).

$$Entropy\ (X) = -\sum_{i=1}^{n} pi * log_2(pi) \tag{12}$$

dengan:

X : Jumlah seluruh *sample* data
n : Jumlah kelas klasifikasi
pi : Probabilitas *term* ke-i

Proses klasifikasi teks metode *maximum entropy* mecari probabilitas himpunan *y* terhadap term *x* yang dinyatakan dengan $f(x, y)$ sebagai berikut:

$$f(x,y)\begin{cases} 1, jika\ y' = y\ dan\ x = benar \\ 0, \qquad\qquad lainnya \end{cases} \tag{13}$$

Probabilitas bersyarat (*conditional probability*) atau suatu keadaan *y* terhadap *x* dapat dilihat dengan persamaan berikut (Putra dkk, 2019):

$$p(y|x) = \frac{\prod_i \alpha_i f(x,y)}{Z(x)} \tag{14}$$

$\prod_i \alpha_i f(x, y)$ merupakan perhitungan probabilitas *term* ke-i didalam suatu dokumen terhadap kelas y, sedangkan nilai $Z(x)$ merupakan nilai normalisasi dari setiap kata dengan menggunakan persamaan berikut (Putra dkk, 2019):

$$Z(x) = \sum_y \prod_i \alpha_i f(x,y) \tag{15}$$

dengan:

$P(y|x)$: Probabilitas bersyarat keadaan *y* terhadap *x*
$Z(x)$: Nilai normalisasi

Menilai atau memvalidasi keakuratan sebuah model dalam klasifikasi *maximum entropy* dengan pembangunan model terdiri dari pembentukan data latih dan data uji yang telah diambil secara acak menggunakan *K-Fold Cross Validation*. *K-Fold Cross Validation* mempartisi himpunan data D secara acak menjadi K *fold* (sub himpunan) yang saling bebas yaitu $f_1, f_2, \ldots, f_k$. *K-fold cross validation* dataset yang digunakan dibagi menjadi K buah segmen dengan ukuran proporsi yang sama kemudian dilakukan pelatihan dan validasi

sebanyak K kali. Kemudian pada setiap pelatihan dan validasi pada tiap perulangannya, diambil satu segmen sebagai data tes, selanjutnya K-1 segmen lainnya diambil dan digunakan sebagai data latih. Dari validasi yang dilakukan akan didapatkan tingkat akurasi yang digunakan sebagai keakuratan model klasifikasi.

Proses evaluasi model klasifikasi dilakukan untuk mengukur akurasi prediksi dari model klasifikasi yang telah dilakukan. Evaluasi model klasifikasi dibutuhkan metode *confusion matrix*, *confusion matrix* merupakan matrix yang berisikan informasi mengenai klasifikasi yang akan diprediksi oleh sistem, dan akan dibandingkan dengan kelas yang asli (Salim dan Mayary, 2020).

Tabel 1. *Confusion Matrix*

		Aktual	
	Class	**Positive**	**Negative**
Prediksi	**Positive**	*True Positive (TP)*	*False Negative (FN)*
	Negative	*False Positive (FP)*	*True Negative (TN)*

keterangan:

True Positive (TP) : Mendefinisikan data positif dan di prediksi positif
False Positive (FP) : Mendefinisikan data positif dan di prediksi negatif
False Negative (FN) : Mendefinisikan data negatif dan di prediksi positif
True Negative (TN) : Mendefinisikan data negatif dan di prediksi negatif

Tabel 2. Ukuran Evaluasi Model Klasifikasi

Ukuran	**Rumus**
Accuracy	$\frac{TP + TN}{(TP + FP + TN + FN)}$
Kappa Statistics	$P_0 = Accuracy$ $P_c = \left[\left(\frac{TP + FP}{Total}\right)\left(\frac{TP + FN}{Total}\right)\right] + \left[\left(\frac{FN + TN}{Total}\right)\left(\frac{FP + TN}{Total}\right)\right]$ $Kappa\ Statistic = \frac{P_0 - P_c}{(1 - P_c)}$

Visualisasi diperlukan dalam tahapan analisis sentimen. Visualisasi dengan *wordcloud* merupakan hasil yang didapat dari metode *text mining* yang menampilkan gambaran visual dari kumpulan kata popular yang berkaitan dengan kata kunci internet. *Wordcloud* terbentuk berdasarkan banyaknya kemunculan kata, jika kata yang paling muncul dalam teks akan menjadi memiliki ukuran yang paling besar, begitu juga dengan yang paling sedikit muncul dalam teks.

3. METODE PENELITIAN

Jenis data yang digunakan dalam penelitian ini merupakan data sekunder. Sumber data yang digunakan adalah data set ulasan terhadap pengguna aplikasi Ajaib *Google Play*. Pengumpulan data menggunakan teknik *web scrapping*, diperoleh 2000 ulasan untuk diteliti yang diambil sejak bulan November 2021 sampai dengan Februari 2022 untuk mengetahui kinerja aplikasi Ajaib.

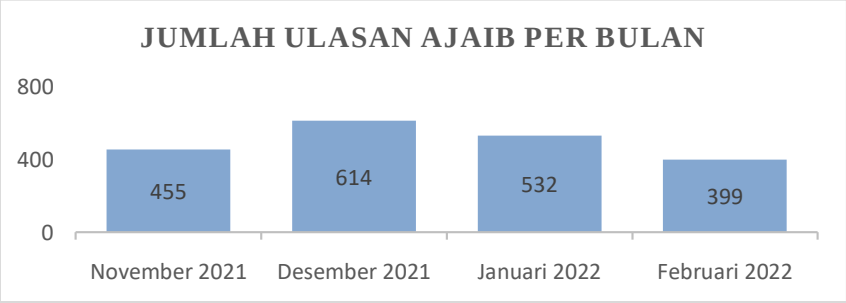
Metode yang digunakan dalam penelitian ini adalah *Support Vector Machine* (SVM) dan *Maximum Entropy*. Penelitian dilakukan dengan bantuan *software Data Miner* untuk *web scrapping*, *Rstudio* dan *Google Colaboratory* untuk analisis data. Adapun tahapan analisis yang dilakukan:

1. *Scrapping* data ulasan aplikasi Ajaib pada *Google Play*.

2. Analisis Deskriptif.
3. Data *Pre-processing* (*case folding*, *cleaning*, dan normalisasi kata).
4. *Sentiment Scoring*.
5. *Feature Selection (stopwords, stemming*, dan *tokenizing*).
6. Pembobotan TF-IDF.
7. Pembagian data latih dan data uji untuk klasifikasi.
8. Membangun model klasifikasi *Support Vector Machine* (SVM) dengan fungsi kernel RBF.
9. Menghitung akurasi dan kappa menggunakan *confusion matrix* untuk evaluasi model klasifikasi SVM.
10. Membangun model klasifikasi dan melakukan pengujian model klasifikasi *Maximum Entropy*.
11. Mengevaluasi performan model klasifikasi *Maximum Entropy.*
12. Interpretasi hasil ke dalam bentuk visual menggunakan *wordcloud.*

4. HASIL DAN PEMBAHASAN

Pengumpulan data ulasan pada Aplikasi *Ajaib* di *Google play* menggunakan metode *scrapping data* dengan menggunakan aplikasi *Data Miner.* Data yang diambil berisi tanggal memberi ulasan, nama akun *google*, serta ulasan yang tersimpan dalam bentuk *Comma Separated Valuse* (CSV). Berikut banyaknya data ulasan pada setiap bulan sejak November 2021 hingga Februari 2022



Gambar 1. Diagram Jumlah Ulasan Ajaib per-Bulan

Setiap ulasan yang telah terkumpul memiliki struktur penulisan yang beragam, tahap *pre-processing* ini dilakukan untuk membentuk data yang siap diproses dengan menghilangkan karakter atau symbol yang tidak bernilai. Penelitian ini menggunakan *case folding, remove emoticon, remove punctuation, remove numbers, remove strip whitespace*, dan normalisasi kata.

Data telah siap diproses kemudian dilakukan pelabelan kelas sentimen. Pelabelan kelas sentimen dengan *sentiment scoring* dengan bantuan kamus kata positif dan kamus kata negatif dengan cara menghitung skor sentimen. Data ulasan yang memiliki skor akhir hitung ≥ 0 masuk ke dalam kelas sentimen positif, jika data ulasan memiliki skor akhir hitung < 0 masuk kategori kelas sentiment negatif. Pelabelan menggunakan teknik *sentiment scoring* lebih dominan kelas positif dibandingkan kelas negatif, dengan hasil ulasan masuk ke dalam kelas positif sebanyak 1528 ulasan, dan ulasan yang masuk ke dalam kelas negatif sebanyak 460 ulasan. Pelabelan menggunakan teknik *sentiment scoring* masih menimbulkan kesalahan sebanyak 3,15% dari 2000 ulasan. Kesalahan disebabkan karena skor akhir dari setiap ulasan yang dilakukan oleh program komputasi berbeda dengan penilaian yang diputuskan oleh manusia, hal ini sering terjadi karena sentimen merupakan hal yang subjektif. Adanya

kesalahan tersebut, dilakukan perbaikan maka didapat jumlah ulasan positif menjadi 1528 ulasan dan ulasan negatif sebanyak 466 ulasan.

Setelah melakukan tahap pelabelan kelas sentimen pada data, kemudian dilakukan proses *feature selection*. Proses *feature selection* terdapat beberapa tahapan yaitu *stopwords Removal*, *Stemming* dan *Tokenizing*. Dalam proses *stopwords removal*, jumlah kata pada seluruh dokumen sebanyak 1023 term. Proses *stemming* digunakan untuk mengubah kata-kata dalam dokumen menjadi kata dasar, dan *tokenizing* digunakan untuk memisahkan kata per kata dalam dokumen menjadi kata yang tidak saling berpengaruh.

Data yang akan digunakan untuk membentuk model klasifikasi dibagi terlebih dahulu menjadi data latih dan data uji, dengan perbandingkan data latih dan data uji yang digunakan yaitu 80%:20% secara berturut-turut. Dapat diartikan bahwa dengan data jumlah keseluruhan terdapat 2000 ulasan, sebanyak 1600 ulasan menjadi data latih dan 400 ulasan menjadi data uji.

Analisis sentimen teks membutuhkan tahapan pembobotan *term* untuk mengubah kata-kata menjadi bilangan yang dapat dihitung. TF-IDF dilakukan dengan mencari nilai bobot dari seluruh *term* berdasarkan banyaknya kemunculan atau frekuensi munculnya *term* tersebut pada dokumen ulasan yang dinamakan dengan *Term Frequency* (TF). Contoh perhitungan *Term Frequency-Inverse Document Frequency* (TF-IDF).

$$W_{ajaib,dok-1} = \frac{Jumlah\ term \text{ ajaib } pada\ dok\ 1}{Jumlah\ seluruh\ term\ pada\ dok\ 1} \times \log_2 \frac{Jumlah\ seluruh\ dok}{Jumlah\ dok\ mengandung\ term\ "ajaib"}$$

$$W_{ajaib,dok-1} = \frac{1}{12} \log_2 \frac{2000}{571} = 0{,}1754$$

Tabel 3. Hasil pembobotan TF-IDF pada ulasan ke-1

Term	Ulasan ke-1	*Term*	Ulasan ke-1
ajaib	0,175	loss	0,698
tambah	0,914	persentase	0,782
gain	0,830	saham	0,204
hai	0,747	tolong	0,335
hasil	0,453	transaksi	0,456
keren	0,368	:	:
kolom	0,637	investasi	0

Klasifikasi SVM pada penelitian ini menggunakan fungsi *kernel Radial Basis Function* (RBF). *Kernel* RBF menggunakan dua parameter yaitu nilai *Cost* dan *gamma*. Pengoptimalan parameter C dilakukan dengan cara *trial and error.* Pada penelitian ini nilai *Cost* (C) yang digunakan adalah 0,1; 1; 10 dan nilai *gamma* yang digunakan adalah 0,002.

Dot product pada SVM dilakukan dengan memasukkan *kernel* RBF pada data, dengan seluruh data dihitung dengan cara yang sama sehingga menghasilkan nilai *kernel*. Contoh perhitungan fungsi kernel RBF ulasan ke-1, dan dilakukan dengan cara yg sama pada ulasan berikutnya.

$$K(x_i, x_i) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) = \exp(-\gamma \sum_{j=1}^{n} (x_i - x_j)^2)$$

$$K(x_1, x_1) = \exp(-0{,}002((0{,}175 - 0{,}175)^2 + \cdots + (0{,}456 - 0{,}456)^2 + \cdots + (0 - 0)^2)$$

$$K(x_1, x_1) = 1$$

SVM terdapat proses mencari nilai $\boldsymbol{\alpha}$ dan $\boldsymbol{b}$ yang optimal dengan menggunakan metode *Quadratic Programming* (QP) P*roblem* pada setiap parameter *Cost* (C) 0,1; 1; 10 dengan bantuan *software Rstudio. Dot product* data latih dan data uji diperoleh dengan cara yang sama, yaitu dengan memasukkan data uji ke data latih pada fungsi *kernel* RBF.

Tabel 4. Nilai $\boldsymbol{\alpha}$ *dan* $\boldsymbol{b}$ pada tiap Nilai C

C	0,1	1	10
$\boldsymbol{\alpha}$	$\begin{bmatrix} 0 \\ 0 \\ 0{,}1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}_{2000\times1}$	$\begin{bmatrix} 0 \\ 0{,}13 \\ 0{,}65 \\ 0 \\ 0 \\ \vdots \end{bmatrix}_{2000\times1}$	$\begin{bmatrix} 0{,}54 \\ 1{,}27 \\ 5{,}41 \\ 0 \\ 0 \\ \vdots \end{bmatrix}_{2000\times1}$
$\boldsymbol{b}$	-1,000435	-0,9171324	0,07713039

Nilai $\boldsymbol{\alpha}$ dan $\boldsymbol{b}$ selanjutnya digunakan pada persamaan SVM untuk memprediksi klasifikasi data uji dan perhitugan *hyperplane*. Persamaan hyperplane dengan $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ merupakan fungsi kernel dari *dot product* data training yang terpilih sebagai *support vector*. Diperoleh *hyperplane* dari salah satu *Cost* yaitu 0,1 dengan *support vector* ke-848 dalam *kernel* RBF sebagai berikut.

$$Hyperplane = (\boldsymbol{w^T}.\boldsymbol{x_i}) + b = \left(\sum_{i=1}^{p} \alpha_i\, y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) - 1{,}000435\right)$$

$$Hyperplane = \left(\left(\alpha_1 y_1 K(x_1 x_1)\right) + \cdots + \left(\alpha_{848} y_{848} K(x_{848} x_{848})\right)\right) - 1{,}000435$$

persamaan SVM *kernel* RBF dengan menggunakan contoh *Cost* 0,1 adalah sebagai berikut.

$$f(\Phi(x)) = sign\left(\sum_{i=1}^{p} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x_j}) + b\right) = f(\Phi(x)) = sign\left(\sum_{i=1}^{p} \alpha_i y_i K(\boldsymbol{x_{training}}, x_{testing}) + b\right)$$

$$f(\Phi(x)) = sign\left(\sum_{i=1}^{p} \alpha_i y_i K(\boldsymbol{x_{training}}, \boldsymbol{x_{testing}}) - 1{,}000435\right)$$

perhitungan dilakukan untuk semua data uji pada setiap *Cost* yang digunakan.

Tingkat kinerja model klasifikasi dapat diukur dengan melihat dari nilai *Accuracy* dan nilai *Kappa* pada setiap parameter *Cost* yang diuji. Berikut Nilai *Accuracy* dan nilai *Kappa* yang dihasilkan pada setiap parameter *Cost.*

Tabel 5. Nilai *Accuracy* dan *Kappa* pada *Kernel* RBF

Evaluasi Model	*Cost*		
	0,1	**1**	**10**
Accuracy	0,7625	0,845	0,8575
Kappa	0	0,535	0,5807

Diperoleh nilai *Accuracy* dan *Kappa* terbaik adalah *Cost* 10 karena menghasilkan nilai terbesar untuk kedua pengukuran. Kemudian dilakukan evaluasi kinerja klasifikasi dengan tabel *confusion matrix* dengan nilai *Cost* terbaik yaitu 10 sebagai berikut.

Tabel 6. *Confusion Matrix* SVM *Cost* 10

Prediksi	Aktual	
	Negatif	**Positif**
Negatif	58	37
Positif	20	285

Perhitungan *accuracy* dan *kappa* pada *confusion matrix* C=10

$$Accuracy = \frac{TP+TN}{(TP+FP+TN+FN)} = \frac{285+58}{(285+58+20+37)} = 85{,}75\%$$

$$Kappa\ Statistic = \frac{P_0-P_c}{(1-P_c)} = \frac{0{,}5575-0{,}660125}{(1-0{,}660125)} = 58{,}07\%$$

Pembangunan model terdiri dari pembentukan data latih dan data uji yang diambil secara acak dengan menggunakan *k-fold cross validation* dengan 5 kali iterasi dengan *trial and error* terlebih dahulu untuk mendapatkan nilai akurasi dan *kappa* prediksi terbaik sehingga masing-masing *fold* akan terisi 400 data dengan total data ulasan sebanyak 2000 ulasan. Hasil rata-rata akurasi dan *kappa* pengujian dari 5 iterasi untuk ulasan pengguna aplikasi Ajaib sebagai berikut.

Tabel 7. Hasil Kinerja *Maximum Entropy* dengan 5 *Fold.*

	Accuracy	***Kappa***
Rata-rata	0,83	0,5050

Rata-rata hasil akurasi dari kinerja klasifikasi *Maximum Entropy* sebesar 83% menandakan dari 2000 ulasan yang ada, terdapat 1660 ulasan terklasifikasi tepat pada kelas sentimennya. Hasil kappa statistik dari kinerja klasifikasi *Maximum Entropy* sebesar 50,5% yang menandakan bahwa kinerja klasifikasi *Maximum Entropy* baik.

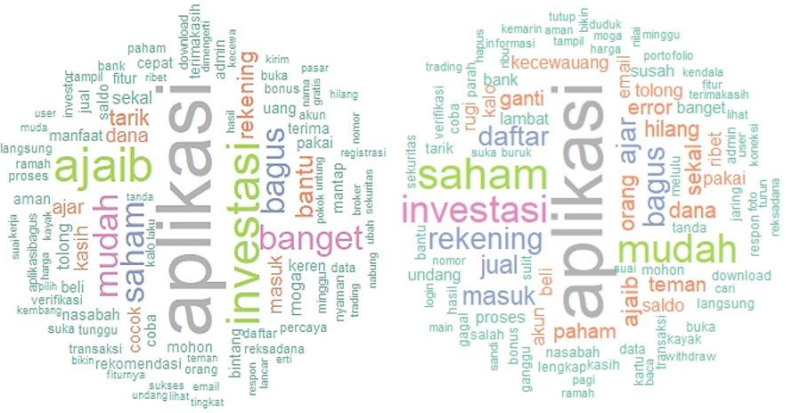
Perbandingan hasil kinerja klasifikasi menggunakan *Support Vector Machine* dan *Maximum Entropy* dapat dilihat dari nilai akurasi akhir yang didapatkan dari kedua algoritma tersebut. Nilai akurasi akhir yang paling tinggi merupakan algoritma yang paling baik dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Ajaib. Tingkat akurasi algoritma *Support Vector Machine* dan *Maximum Entropy* sebagai berikut.

Tabel 8. Perbandingan Hasil Kinerja Klasifikasi

	Support Vector Machine	***Maximum Entropy***
Accuracy	85,75%	83%
Kappa	58,07%	50,5%

Support Vector Machine memiliki akurasi terbaik sebesar 85,75% dan memiliki nilai *kappa* sebesar 58,07%. Pada algoritma *Maximum Entropy* didapat rata-rata akurasi sebesar 83% dan memiliki nilai *kappa* sebesar 50,5% Hasil perbandingan bahwa algoritma *Support Vector Machine* mempunyai tingkat akurasi dan nilai *kappa* yang lebih besar dibandingkan dengan algoritma *Maximum Entropy* dalam mengklasifikasikan sentimen ulasan pengguna aplikasi Ajaib.

WordCloud atau awan kata merupakan kumpulan kata yang sering muncul dari teks yang dianalisis. Visualisasi dengan *wordcloud* menggunakan *software* Rstudio.



Gambar 2. *Wordcloud* Sentimen Positif dan Negatif

5. KESIMPULAN

Berdasarkan hasil dan pembahasan yang dilakukan terhadap sentimen ulasan pengguna aplikasi Ajaib sebagai beikut.

1. Pengguna aplikasi Ajaib cenderung menuliskan ulasan pada aplikasi Ajaib dengan sentiment positif. Didukung dengan hasil 1528 ulasan termasuk kelas sentimen positif dan 466 ulasan termasuk kelas sentimen negatif.
2. Klasifikasi sentiment menggunakan *Support Vector Machine* menghasilkan nilai *Accuracy* dan *Kappa* dengan parameter *Cost* pada *kernel Radial Basis Function* (RBF) 0,1; 1; 10 sebesar 85,75% dan 58,07%. Klasifikasi sentimen menggunakan *Maximum Entropy* menghasilkan nilai *Accuracy* dan *Kappa* sebesar 83% dan 50,5%.
3. Klasifikasi sentimen pada ulasan aplikasi Ajaib menggunakan *Support Vector Machine* (SVM) dan *Maximum Entropy*, terbukti bahwa SVM merupakan metode terbaik dalam analisis sentimen ulasan aplikasi Ajaib.

DAFTAR PUSTAKA

Alroy, A. B. (2019). *Klasifikasi Hoaks Menggunakan Metode Maximum Entropy Dengan Seleksi Fitur Information Gain* (Doctoral dissertation, Universitas Brawijaya).

Cortes, C., & Vapnik, V. 1995. *Support-vector networks*. Machine learning, 20(3), 273-297.

Gunn, S. R. 1998. *Support Vector Machines for Classification and Regression*. Southampton: University of Southampton

Liu, B. 2010. *Sentiment analysis and subjectivity*. Handbook of natural language processing, 2(2010), 627-666.

Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). Support vector machine. *Proceeding Indones. Sci. Meeiting Cent. Japan*.

Prasetyo, E., 2012. *Data Mining Konsep dan Aplikasi menggunakan Matlab.* Yogyakakara: Andi

Putra, M. F., Herdiani, A., & Puspandari, D. (2019). Analisis Pengaruh Normalisasi, Tf-idf, Pemilihan Feature-set Terhadap Klasifikasi Sentimen Menggunakan Maximum Entropy (studi Kasus: Grab Dan Gojek). *eProceedings of Engineering*, *6*(2).

Sabily, A. F., Adikara, P. P., & Fauzi, M. A. (2019). Analisis Sentimen Pemilihan Presiden 2019 pada Twitter menggunakan Metode Maximum Entropy. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN*, *2548*, 964X.

Salim, S. S., & Mayary, J. (2020). Analisis Sentimen pengguna Twitter terhadap dompet elektronik dengan metode lexicon based dan k–nearest neighbor. *Jurnal Ilmiah Informatika Komputer*, *25*(1), 1-17.

Tandelilin, E. 2010. *Portofolio dan Investasi: Teori dan aplikasi*. Kanisius.
Berita Satu. 2021. Jumlah Investor Pasar Modal Tembus 6,59 Juta. https://www.beritasatu.com/ekonomi/845369/jumlah-investor-pasar-modal-tembus-659-juta.