Text Mining Untuk Analisis Sentimen Pelanggan Terhadap Layanan Uang Elektronik Menggunakan Algoritma Support Vector Machine

Fajar Romadoni[1]*, Yuyun Umaidah[2], Betha Nurina Sari[3]
Fakultas Ilmu Komputer[1], [2], [3]
Universitas Singaperbangsa Karawang
Karawang, Indonesia
fajar.16083@student.unsika.ac.id[1], yuyun.umaidah@staff.unsika.ac.id[2], betha.nurina@staff.unsika.ac.id [3]

***Abstract*— Electronic money is a cashless payment instrument whose money is stored in media server or chip that can be moved for the benefit of payment transactions or fund transfers. In Indonesia, there are already many electronic money products, one of which is OVO. OVO is very popular with the people of Indonesia because it offers many promos such as discounts and cashback. But over time, that much promotion is detrimental to OVO shareholders, so the portion of promo given by OVO to its customers is finally reduced. That incident caused many pros and cons opinions about OVO, one of them is on social media Twitter. Sentiment analysis can be used as a solution to process the opinions of OVO customers on Twitter. This study aims to classify the customer opinions on OVO services into positive and negative classes. This study uses the Support Vector Machine algorithm with 3852 data taken from Twitter with keyword @ovo_id using web scraping techniques. The dataset divided into two classes, 2034 positive and 1818 negative sentiment data. The classification process is carried out with four splitting data scenarios, with 60:40, 70:30, 80:20, 90:10 data ratio and with four kernel such as linear, rbf, sigomid, and polynomial. The final results show that the greatest accuracy value obtained by linear kernel with 90:10 data ratio which gets an accuracy value of 98.7%.**

Keywords*— *Classification, Electronic Money, Kernel, Sentiment Analysis, Support Vector Machine

***Abstrak*— Uang elektronik adalah alat pembayaran nontunai atau *cashless* yang nilai uangnya disimpan di dalam media server ataupun chip yang dapat dipindahkan untuk kepentingan transaksi pembayaran atau transfer dana. Di Indonesia sendiri sudah banyak produk uang elektronik, salah satunya OVO. OVO sangat digemari oleh masyarakat Indonesia karena menawarkan banyak promo seperti diskon dan *cashback*. Tetapi seiring berjalannya waktu justru pemberian promo yang sangat banyak itu merugikan para pemegang saham OVO, sehingga porsi promo yang diberikan OVO kepada para pelanggannya akhirnya dikurangi. Kejadian itu menimbulkan banyak pendapat pro dan kontra terhadap OVO salah satunya pada media sosial Twitter. Analisis sentimen dapat dijadikan solusi mengolah pendapat dari para pelanggan OVO pada Twitter. Penelitian ini bertujuan untuk mengklasifikasikan pendapat para pelanggan terhadap layanan uang elektronik OVO ke dalam kelas positif dan negatif. Penelitian ini menggunakan algoritma *Support Vector Machine* dengan *dataset* berjumlah 3852 yang diambil dari Twitter dengan kata kunci @ovo_id dengan teknik *web scraping*. *Dataset* terbagi menjadi dua kelas yaitu 2034 data sentimen positif dan 1818 data sentimen negatif. Proses klasifikasi dilakukan dengan empat skenario *splitting data* yaitu dengan rasio 60:40, 70:30, 80:20, 90:10 dan dengan empat kernel yaitu kernel linear, rbf, sigomid, dan polynomial. Hasil akhir menunjukkan bahwa nilai akurasi terbesar didapat oleh kernel linear dengan rasio data 90:10 yang mendapatkan nilai akurasi sebesar 98.7%.**

Kata Kunci*— *Analisis Sentimen, Kernel, Klasifikasi, Support Vector Machine, Uang Elektronik.

I. PENDAHULUAN

Uang elektronik adalah alat pembayaran nontunai atau *cashless* yang nilai uangnya disimpan secara elektronik di dalam media server ataupun *chip* yang dapat dipindahkan untuk kepentingan transaksi pembayaran atau transfer dana [1]. Dengan perkembangan teknologi yang semakin maju tentu saja persebaran uang elektronik semakin cepat. Dengan adanya uang elektronik tentu saja masyarakat Indonesia dimudahkan dalam banyak hal, seperti membeli pulsa, belanja, membayar pajak dan masih banyak lagi. Di Indonesia sendiri sudah banyak produk uang elektronik seperti OVO, Go-Pay, DANA, Tcash, dll. Persaingan bisnis tersebut tentu saja bukan hanya mencari keuntungan semata, tetapi mencoba menjadi yang terbaik dan memuaskan para pelanggan.

OVO adalah produk dompet digital yang memberikan banyak layanan transaksi secara daring, yaitu bisa menyimpan uang tunai menjadi uang elektronik, mengirim uang serta membayar berbagai transaksi. OVO juga aman digunakan karena sudah mendapat izin lisensi dari Bank Indonesia. Dengan aplikasi OVO masyarakat dapat melakukan transaksi tanpa harus membawa uang tunai dalam jumlah yang besar. Sampai sekarang banyak masyarakat menggunakan OVO, bahkan OVO menjadi aplikasi penyimpanan uang elektronik favorit bagi masyarakat Indonesia karena banyaknya promo yang ditawarkan oleh OVO. OVO merupakan aplikasi uang

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020

elektronik berbasis *mobile* dengan nilai transaksi terbesar pada semester pertama tahun 2019 di Indonesia, dari keseluruhan nilai transaksi sebesar Rp. 56.1T, OVO merupakan aplikasi dengan nilai transaksi terbesar yaitu dengan nilai pangsa pasar 37% atau Rp. 20.8T dari keseluruhan transaksi uang elektronik berbasis *mobile*, mengalahkan para pesaingnya yaitu Gopay yang memiliki porsi 17% atau setara Rp. 9.5T, aplikasi DANA sebesar 10%, dan LinkAja sebesar 3% [2]. OVO sangat digemari oleh masyarakat Indonesia karena menawarkan banyak promo seperti diskon dan *cashback*. Tetapi seiring berjalannya waktu justru pemberian promo yang sangat banyak itu merugikan para pemegang saham OVO, salah satunya Lippo Group sebagai pemegang saham utama OVO. Lippo Group melepas 2/3 sahamnya atau sekitar 70% dikarenakan tidak kuat lagi untuk "bakar uang" atau sebuah kegiatan menghabiskan uang yang banyak untuk sebuah proses bisnis misalnya dengan menawarkan diskon, layanan gratis atau *cashback* seperti yang dilakukan OVO kepada para pelanggannya [3]. Hal ini tentunya akan mengurangi porsi promo yang diberikan OVO kepada para pelanggannya dan dengan kejadian itu pastinya banyak tanggapan pro dan kontra terhadap OVO salah satunya pada media sosial Twitter. Maka dari itu penelitian ini akan menganalisis sentimen para pelanggan terhadap pelayanan uang elektronik OVO.

Analisis sentimen atau *opinion mining* merupakan perpaduan dari data mining dan text mining, suatu teknik untuk menganalisa pendapat, sentimen, evaluasi, sikap, penilaian, perasaan dan emosi seseorang apakah pembicara atau penulis berkenan dengan suatu topik, produk, layanan, organisasi, individu, ataupun kegiatan tertentu. Dengan analisis sentimen kita bisa mengetahui apakah isi teks itu bersifat positif atau negatif. Sentimen mengacu pada fokus topik tertentu, pernyataan suatu topik mungkin akan berbeda makna dengan pernyataan sama pada subjek berbeda, oleh karena itu, pada beberapa penelitian didahului dengan menentukan elemen dari sebuah produk yang sedang dibicarakan sebelum memulai analisis sentimen [4]. Terdapat banyak algoritma untuk melakukan analisis sentimen salah satunya algoritma *Support Vector Machine* (SVM). SVM mampu memisahkan data dengan baik dibandingkan algoritma klasifikasi lainnya karena memiliki fungsi *kernel trick* yang dapat digunakan untuk mentransformasi data ke ruang berdimensi lebih tinggi yang disebut sebagai ruang kernel [5].

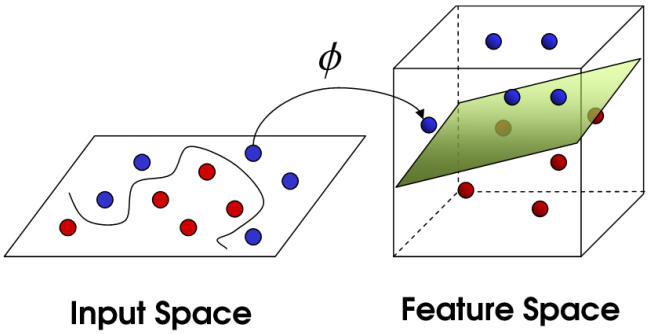
Analisis yang dilakukan bertujuan untuk mengklasifikasi sentimen para pelanggan terhadap layanan uang elektronik OVO melalui teknik *text mining* dengan data yang diperoleh dari *tweet* para pelanggan OVO pada media sosial Twitter menggunakan algoritma *Support Vector Machine* dengan berbagai macam kernel yaitu kernel linear, polynomial, rbf, dan sigmoid untuk mencari nilai akurasi terbaik. Penelitian ini diharapkan dapat membantu pihak OVO untuk mengembangkan bisnisnya serta memperbaiki apa yang dikeluhkan para pelanggannya.

II. TEORI PENUNJANG

A. Algoritma Support Vector Machine

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi dengan menggunakan metode *supervised learning* yang memprediksi kelas berdasarkan pola dari hasil proses training. Klasifikasi dilakukan dengan *hyperplane* yang memisahkan antara kelas positif dan negatif. Suatu *hyperplane* yang baik adalah yang memiliki jarak terbesar ke titik data pelatihan terdekat dari setiap kelas, karena pada umumnya semakin besar margin, semakin rendah *error* generalisasi dari pemilah. *Hyperplane* pemisah terbaik antara kedua kelas dapat dilakukan dengan mengukur *margin hyperplane* dan mencari titik maksimalnya [6].

Ketika terdapat permasalahan data yang tidak terpisah secara linear dalam ruang input SVM, SVM tidak dapat menemukan *hyperplane* pemisah yang kuat yang meminimalkan misklasifikasi dari *data points* serta menggeneralisasi dengan baik. Untuk itu, *kernel trick* dapat digunakan untuk mentransformasi data ke ruang berdimensi lebih tinggi yang disebut sebagai ruang kernel, dimana akan menjadikan data terpisah secara linear [7], konsep dasar *kernel trick* dapat dilihat pada Gambar 1.



Gambar 1. Konsep Kernel SVM

Pada dasarnya pembelajaran SVM untuk menentukan *support vector* hanya bergantung pada *dot product* dari data pada ruang fitur, yaitu $\phi_i \phi_j$. Pada umumnya transformasi ϕ tidak diketahui dan sangat sulit dipahami. Oleh karena itu, perhitungan *dot product* dapat diganti dengan fungsi kernel $Kx_i x_j$ yang mendefinisikan secara implisit fungsi transformasi ϕ, itulah yang dimaksud *kernel trick*. *Kernel trick* diformulasikan sebagai berikut [8]:

$$K(x_i, x_j) = \phi_j(x_i).\phi_j(x_j) \quad (1)$$

Pada umumnya terdapat empat jenis fungsi kernel yang dapat digunakan, yaitu:

1. Kernel Linier
$$K(x, x_k) = x_k^T x \quad (2)$$
2. Kernel Polynomial
$$K(x, x_k) = x_k^T x^d + 1 \quad (3)$$
3. Kernel Radial Basis Function (RBF)
$$K(x, x_k) = exp\{-||x - x_k||_2^2 / \sigma^2\} \quad (4)$$
4. Kernel Sigmoid
$$K(x, x_k) = tanh\,[k\, x_k^T x + \theta] \quad (5)$$

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020

B. *Knowledge Discovery in Databases (KDD)*

Knowledge Discovery in Databases merupakan metode untuk mendapatkan pengetahuan dari basis data yang ada. Didalam basis data tersebut terdapat tabel-tabel yang saling berelasi. Hasil pengetahuan dalam proses tersebut dapat digunakan sebagai *knowledge base* untuk kepereluan pengambilan keputusan. KDD dan data mining sering digunakan untuk penggalian informasi tersembunyi dari basis data yang besar. Proses KDD dibagi menjadi 5 tahapan, diantaranya *yaitu data selection, preprocessing, transformation, data mining*, dan *evaluation* [9].

C. *Confusion Matrix*

Confusion matrix merupakan suatu alat yang berfungsi untuk menganalisis seberapa baik hasil klasifikasi dalam mengenali *tuple* dari kelas yang berbeda. *Confusion matrix* juga bisa digunakan untuk mencari nilai akurasi. Akurasi merupakan rasio prediksi yang benar dengan data keseluruhan. Contoh confusion matrix dapat dilihat pada Tabel 1 [10].

TABLE I TABEL CONFUSION MATRIX

Data		Aktual	
		True (Positif)	False (Negatif)
Prediksi	True (Positif)	TP	FP
	False (Negatif)	FN	TN

Penjelasan dari Tabel 1 yaitu sebagai berikut :

1. *True Positive* (TP) yaitu data kelas positif yang diklasifikasikan sebagai kelas positif.
2. *False Positive* (FP) yaitu data kelas negatif yang diklasifikasikan sebagai kelas positif.
3. *True Negative* (TN) yaitu data kelas negatif yang di klasifikasikan sebagai kelas negatif.
4. *False Negative* (FN) yaitu data kelas positif yang diklasifikasikan sebagai kelas negatif.

Untuk menghitung nilai akurasi bisa menggunakan persamaan seperti berikut :

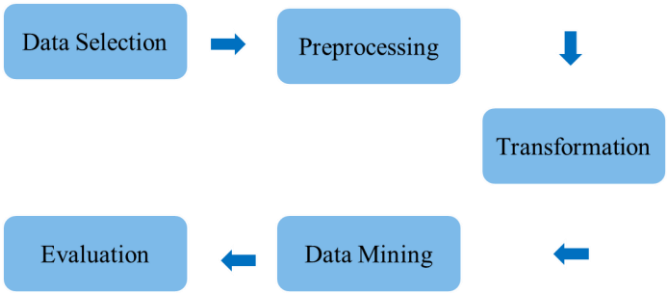
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (6)$$

D. *Web Scraping*

Web scraping atau sering disebut sebagai ekstrasi *web* yaitu proses pengambilan sebuah informasi dokumen dari sebuah *web* dan menganalisis dokumen tersebut untuk mengambil data dari halaman tersebut untuk digunakan bagi kepentingan yang lain [11]. *Scraping* merubah data tidak terstruktur menjadi terstruktur dan disimpan dalam sebuah *database*. Proses *scraping* sangat membantu dalam pengambilan sebuah data yang besar salah satunya mengambil data *tweet* dari Twitter dalam jumlah yang banyak.

III. METODOLOGI

Metodologi penelitian yang digunakan yaitu metode *Knowledge Discovery in Databases* (KDD) yang memiliki 5 tahap yaitu *selection data , preprocessing, transformation, data mining, dan evaluation* seperti pada Gambar 2.



Gambar 2. Tahapan Knowledge Discovery in Databases

A. *Selection Data*

Pada tahapan awal ini akan dilakukan pencarian dan pengambilan data *tweet* dari Twitter yang relevan dengan penelitian. Data yang diambil menggunakan teknik *web scraping* dengan bantuan *package* Twitterscraper dengan *tools* Anaconda Prompt. Data yang diambil dari Twitter yaitu *tweet* dari bulan januari 2019 sampai bulan januari 2020 dengan kata kunci @ovo_id. Untuk proses *scraping* pada Twitterscraper dapat dilihat pada Gambar 3.

```
(base) C:\Users\ASUS>twitterscraper "@ovo_id" -bd 2019-01-01 -ed 2020-01-01 --lang=id --output=ovo.csv -
INFO: {'User-Agent': 'Mozilla/5.0 (Windows; U; Windows NT 6.1; rv:2.2) Gecko/20110201'}
INFO: queries: ['@ovo_id since:2019-01-01 until:2019-01-19', '@ovo_id since:2019-01-19 until:2019-02-06'
19-02-24 until:2019-03-15', '@ovo_id since:2019-03-15 until:2019-04-02', '@ovo_id since:2019-04-02 until
_id since:2019-05-08 until:2019-05-27', '@ovo_id since:2019-05-27 until:2019-06-14', '@ovo_id since:2019
7-20', '@ovo_id since:2019-07-20 until:2019-08-08', '@ovo_id since:2019-08-08 until:2019-08-26', '@ovo_i
until:2019-10-01', '@ovo_id since:2019-10-01 until:2019-10-20', '@ovo_id since:2019-10-20 until:2019-11-
:2019-11-25 until:2019-12-13', '@ovo_id since:2019-12-13 until:2020-01-01']
INFO: {'User-Agent': 'Mozilla/5.0 (Windows; U; Windows NT 6.1; rv:2.2) Gecko/20110201'}
INFO: {'User-Agent': 'Mozilla/5.0 (compatible, MSIE 11, Windows NT 6.3; Trident/7.0; rv:11.0) like Gecko
INFO: {'User-Agent': 'Mozilla/5.0 (Windows; U; Windows NT 6.1; x64; fr; rv:1.9.2.13) Gecko/20101203 Fire
INFO: {'User-Agent': 'Mozilla/5.0 (Windows; U; Windows NT 6.1; rv:2.2) Gecko/20110201'}
INFO: {'User-Agent': 'Mozilla/5.0 (compatible, MSIE 11, Windows NT 6.3; Trident/7.0; rv:11.0) like Gecko
INFO: {'User-Agent': 'Mozilla/5.0 (compatible, MSIE 11, Windows NT 6.3; Trident/7.0; rv:11.0) like Gecko
INFO: Querying @ovo_id since:2019-01-19 until:2019-02-06
INFO: {'User-Agent': 'Mozilla/5.0 (Windows NT 5.2; RW; rv:7.0a1) Gecko/20091211 SeaMonkey/9.23a1pre'}
INFO: {'User-Agent': 'Opera/9.80 (X11; Linux i686; Ubuntu/14.10) Presto/2.12.388 Version/12.16'}
INFO: Querying @ovo_id since:2019-01-01 until:2019-01-19
INFO: Scraping tweets from https://twitter.com/search?f=tweets&vertical=default&q=@ovo_id%20since%3A2019
INFO: Using proxy 45.234.200.18:53281
INFO: {'User-Agent': 'Mozilla/5.0 (Windows NT 5.2; RW; rv:7.0a1) Gecko/20091211 SeaMonkey/9.23a1pre'}
INFO: Scraping tweets from https://twitter.com/search?f=tweets&vertical=default&q=@ovo_id%20since%3A2019
```

Gambar 3. Proses scraping menggunakan Twitterscraper

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020

Penjelasan sintaks dari Gambar 3 yaitu :

a. Twitterscraper merupakan *package* dari python yang berfungsi untuk *scraping* data dari Twitter
b. "@ovo_id" merupakan kata kunci untuk pencarian *tweet* dari Twitter
c. -bd (*begin date*) merupakan tanggal awal dari *tweet* yang akan dicari
d. -ed (*end date*) merupakan akhir atau batas pencarian
e. --lang merupakan bahasa dari *tweet* yang akan diambil
f. --output merupakan nama file hasil *scraping*
g. --csv merupakan ektensi dari file hasil *scraping*.

Output dari hasil *scraping* merupakan file *tweet* dari Twitter berbentuk CSV dengan 21 *attribut*, data tersebut dapat dilihat pada Gambar 4.


RStudio Source Editor

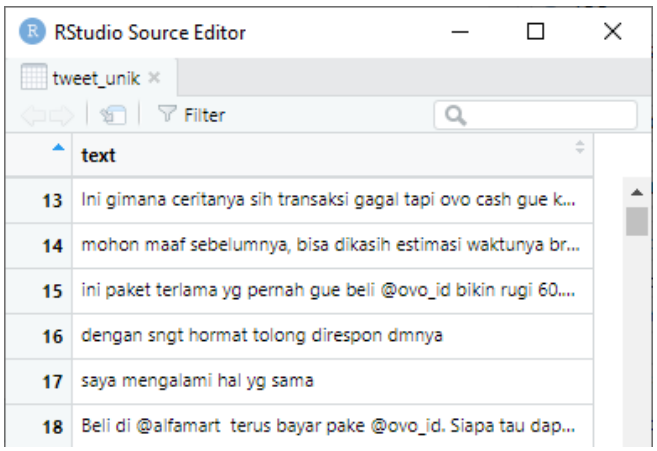
ovo

Filter

	screen_name	text
1	ovo_id	Hi Kak, mohon maaf terkait kendala Kakak. Tidak perlu khaw...
2	IMISSYOO__	@ovo_id kenapa nih min?
3	claudirafa	Kak udah complain via email? Aku semalem complain via e...
4	ovo_id	Hi Kak steven, bisa diinformasikan secara detail kendala mel...
5	stevenwijayaa	@ovo_id pagi, mau tanya dong min. Saya udh top up dgn n...
6	ovo_id	Hai Kak Pratiwi, dimohon untuk menunggu proses tindaklan...
7	litathings	Min sy udh lampirin bukti tf ovo ke dm ya, tp sampai skrg b...
8	tiwitiwukk	Susah di DM.. Sampai pagi ini saldonya belum juga masuk..
9	ovo_id	Hi Kak Asdas, kami informasikan jika saat ini untuk proses u...

Gambar 4. Dataset hasil scraping

Langkah selanjutnya setelah proses *scraping* yaitu melakukan penghapusan kolom atau *attribut* yang tidak terpakai, untuk proses sentimen sendiri hanya memerlukan *attribut* text saja yang berisikan *tweet* dari para pelanggan, maka dari itu *attribut* yang tidak terpakai akan dihapus seperti pada Gambar 5.


RStudio Source Editor

tweet_unik

Filter

	text
13	Ini gimana ceritanya sih transaksi gagal tapi ovo cash gue k...
14	mohon maaf sebelumnya, bisa dikasih estimasi waktunya br...
15	ini paket terlama yg pernah gue beli @ovo_id bikin rugi 60....
16	dengan sngt hormat tolong direspon dmnya
17	saya mengalami hal yg sama
18	Beli di @alfamart terus bayar pake @ovo_id. Siapa tau dap...

Gambar 5. Dataset setelah penghapusan attribut

B. Preprocessing

Setelah data *tweet* tersebut selesai diseleksi, tahapan selanjutnya yaitu tahapan *preprocessing*. Tahapan *preprocessing* bertujuan untuk membersihkan data yang kotor. Pada tahapan ini URL, *mention*, *hashtag*, angka, simbol, dan tanda baca akan dibersihkan. Selanjutnya mengubah kata-kata yang disingkat dan mengubah kata-kata gaul (*slangword*) menjadi kata yang sebenarnya sesuai kaidah bahasa Indonesia. Kemudian menghapus kata-kata yang tidak penting (*stopword*) seperti "yang", "di", dll. Setelah data dibersihkan selanjutnya masuk ke tahapan terakhir *preprocessing* yaitu *tweet* tersebut akan diberikan label dengan menggunakan teknik skoring. Contoh tahapan preprocessing dapat dilihat pada Tabel 2.

TABLE II PREPROCESSING

Sebelum preprocessing	Sesudah preprocessing
@ovo_id Kemarin topup di indomart blm masuk min.	kemarin topup indomart belum masuk
#CashbackOVO mantap sekali pake ovo selalu dpt cashback	mantap pake ovo selalu dapat cashback

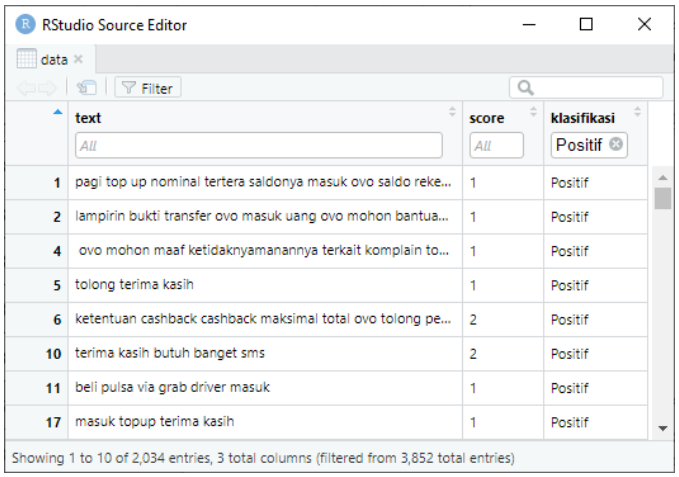
Setelah dataset bersih selanjutnya melakukan tahapan teknik skoring. Skoring merupakan tahap untuk memberikan label pada data *unsupervised*. Sebelum mengklasifikasikan *dataset* menggunakan algoritma SVM, data tersebut harus mempunyai label terlebih dahulu. Setiap kata yang mengandung kata positif akan mendapat skor +1, sementara kata negatif mendapat skor –1 berdasarkan kamus kata positif dan negatif [12]. Setelah itu untuk menentukan kelas *tweet* yaitu dengan cara menghitung jumlah kata positif dikurang jumlah kata negatif. Hasil kalimat dari perhitungan skor tersebut yaitu jika >0 akan dilabeli kelas positif, sementara skor <0 akan dilabeli negatif [13]. Jadi ketika di dalam sebuah *tweet* terdapat kata positif sebanyak 2 kata, maka skor akhirnya adalah 2 yang termasuk ke dalam kelas positif, dan jika di dalam sebuah *tweet* terdapat 2 kata negatif dan 1 kata positif maka skornya adalah -1 dan masuk ke dalam kelas negatif. Contoh perhitungan dengan teknik skoring dapat dilihat pada tabel 3.

TABLE III CONTOH PERHITUNGAN SKORING

Tweet	Positif	Negatif	Skor	Sentimen
promo OVO cashback	promo, cashback	-	2-0 = 2	Positif
salah kirim isi ulang hangus, tapi cashback masuk	cashback	salah, hangus	1-2 = -1	Negatif

Dari total 3852 data yang sudah dibersihkan dan sudah melalui tahap skoring, sebanyak 2034 data berhasil diklasifikasikan ke dalam kelas positif, dan 1818 data diklasifikasikan ke dalam kelas negatif. Data tersebut dapat dilihat pada Gambar 6.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020



Gambar 6. Dataset setelah tahap skoring

Selanjutnya data-data tersebut akan ditransformasi menjadi data berbentuk vektor agar bisa diolah oleh algoritma SVM.

C. *Transformation*

Transformation merupakan tahap perubahan data, karena algoritma SVM hanya bisa memproses data numerik atau berbentuk vektor, maka data *tweet* yang sudah diberi label diubah menjadi data numerik dengan cara pembobotan kata. Tahapan *term weighting* atau pembobotan kata dilakukan dengan menghitung frekuensi kemunculan kata (*term frequency*) di dalam sebuah *tweet* dan kata di dalam *tweet* tersebut menjadi *attribut*. *Term Frequency* (TF) merupakan salah satu skema pembobotan paling populer saat ini, 83% dari sistem rekomendasi berbasis teks di perpustakaan menggunakan TF [14]. Untuk contoh perhitungan TF, Misalnya *dataset* yang akan diolah memiliki 2 *tweet* yaitu "c*ashback* ovo baik baik" dan "pelayanan ovo mengecewakan", dari data tersebut maka *term frequency* yang diperoleh adalah sebagai sebagaimana pada tabel 4.

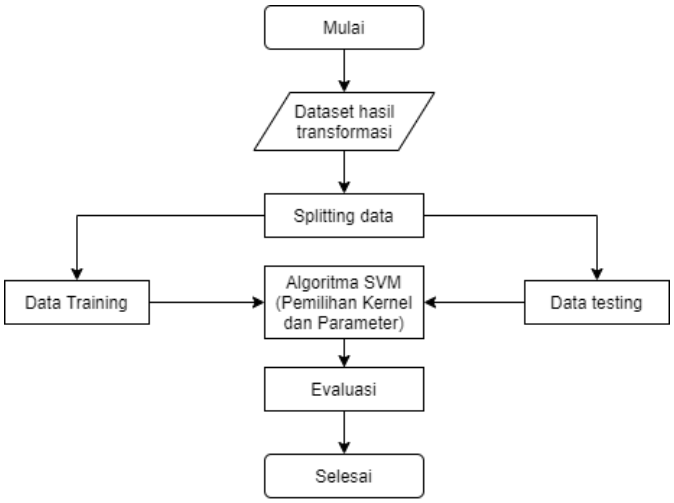
TABLE IV HASIL TERM FREQUENCY

Nomor	*cashback*	ovo	baik	pelayanan	mengecewakan
1	1	1	2	0	0
2	0	1	0	1	1

Tabel 4 merupakan hasil dari pembobotan kata dengan *term frequency*, kata di dalam *tweet* menjadi *attribute* dan di dalam *attribute* tersebut merupakan frekuensi munculnya kata di dalam sebuah *tweet*.

D. *Data Mining*

Tahapan *data mining* merupakan tahapan penerapan algoritma untuk klasifikasi. Pada tahapan ini sebelum mulai melakukan penerapan algoritma, kita harus membagi terlebih dahulu dataset menjadi *data training* dan *data testing*.



Gambar 7. Alur proses data mining

Gambar 7 menjelaskan alur proses *data mining*. *Dataset* yang telah di transformasi selanjutnya dibagi menjadi *data training* dan *data testing*. *Data training* bertujuan untuk melatih *dataset* agar algoritma mengenali mana data yang termasuk ke dalam kelas positif dan mana yang termasuk ke dalam kelas negatif. Sesudah dilatih, selanjutnya akan dilakukan tes pada *data testing* terhadap model yang telah didapat dari *data training* tersebut. Pada penelitian ini untuk mencari akurasi terbaik maka *splitting data* dibagi menjadi 4 skenario yaitu membagi dataset secara acak menjadi data training dan data testing dengan rasio sebagai berikut :

Pertama : 60% *data training* dan 40% *data testing*
Kedua : 70% *data training* dan 30% *data testing*
Ketiga : 80% *data training* dan 20% *data testing*
Keempat : 90% *data training* dan 10% *data testing*

Untuk hasil *splitting data* bisa dilihat pada Tabel 5.

TABLE V. HASIL SPLITTING DATA

Skenario	*Training*	*Testing*
60 – 40	2311	1541
70 – 30	2697	1155
80 – 20	3081	771
90 – 10	3467	385

Setelah *splitting data* dilakukan maka selanjutnya adalah melakukan *training* terhadap algoritma SVM dengan menggunakan *data training* hasil dari *splitting data*, setelah itu hasil *training* tersebut bisa langsung diterapkan untuk klasifikasi sentimen menggunakan *data testing*. Untuk mencari akurasi terbaik maka dilakukan perbandingan kernel dari SVM antara lain yaitu kernel linear, rbf, sigmoid, dan polynomial.

E. *Evaluation*

Tahapan ini merupakan tahap akhir dari KDD, untuk memudahkan perhitungan akurasi hasil *testing* dari setiap *dataset* pada tahap sebelumnya, maka digunakanlah *confusion matrix*. Akurasi menunjukkan banyaknya data yang benar

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020

diklasifikasikan sesuai label. Hasil salah satu confusion matrix dapat dilihat pada Gambar 8.

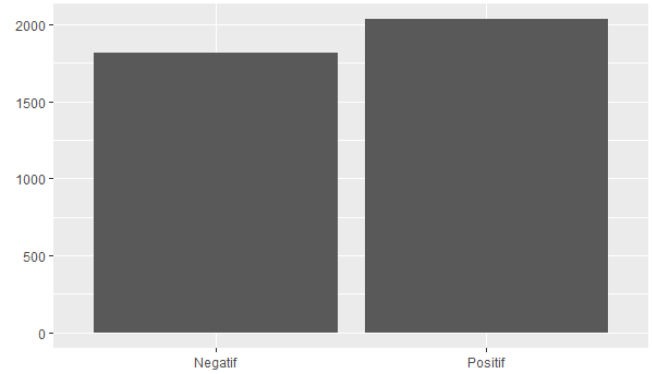
```
Confusion Matrix and Statistics

          Reference
Prediction Negatif Positif
   Negatif     182       5
   Positif       0     198

               Accuracy : 0.987
                 95% CI : (0.97, 0.9958)
    No Information Rate : 0.5273
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.974

 Mcnemar's Test P-Value : 0.07364

            Sensitivity : 0.9754
            Specificity : 1.0000
```

Gambar 8. Confusion matrix

Pada Gambar 8 dapat disimpulkan bahwa sebanyak 182 data negatif berhasil diklasifikasikan kedalam kelas negatif (true negative), sementara 198 data positif berhasil diklasifikasikan kedalam kelas positif (true positive), dan sebanyak 5 data yang salah diklasifikasikan.

IV. HASIL DAN PEMBAHASAN

Dari proses KDD yang telah dilakukan, data yang didapat dari proses *scraping* pada Twitter mulai dari bulan januari 2019 hingga bulan januari 2020 yaitu sebanyak 3852 yang terbagi menjadi 2034 kelas positif, dan 1818 kelas negatif seperti yang terdapat pada Gambar 9.



Gambar 9. Visualisasi data hasil labelling

Setelah *dataset* memiliki label, langkah selanjutnya data-data tersebut akan ditransformasi menjadi data numerik agar bisa diolah oleh algoritma SVM. Proses transformasi menggunakan teknik *term frequency* yang hasilnya dapat dilihat pada Gambar 10.

```
<<DocumentTermMatrix (documents: 3852, terms: 5723)>>
Non-/sparse entries: 32611/22012385
Sparsity           : 100%
Maximal term length: 23
Weighting          : term frequency (tf)
Sample             :
      Terms
Docs   cashback kasih masuk mohon ovo pakai saldo terima top transfer
  1106        0     0     0     0   2     1     0      0   0        0
  1158        0     1     0     0   0     0     0      1   0        4
  1393        0     1     0     2   0     0     0      1   0        0
  1813        0     2     0     0   1     2     0      2   0        1
  1955        0     1     0     1   0     0     0      1   0        0
  2532        0     1     0     0   0     0     0      1   0        0
  295         2     0     0     0   2     2     0      0   0        0
  3495        0     0     0     1   0     0     0      0   0        0
  3644        0     0     0     0   1     0     0      0   0        2
  502         0     0     1     0   2     1     1      0   0        0
```

Gambar 10. Hasil transformasi data

Pada Gambar 10 bisa dilihat bahwa dari sebanyak 3852 *tweet* yang sudah dibersihkan, terdapat sebanyak 5732 terms (kata) yang berbeda-beda. Kata tersebut nantinya akan menjadi *attribute* pada *dataset*. Setelah proses transformasi selesai dilakukan, tahapan selanjutnya merupakan tahap *data mining*.

```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  polynomial
       cost:  1

Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  sigmoid
       cost:  1
```

Gambar 11. Kernel SVM

Gambar 11 merupakan kernel-kernel SVM dengan parameter C = 1 yang akan digunakan untuk proses klasifikasi. Parameter C diatur menjadi 1 karena pada penelitian [15] akurasi terbaik yaitu 100% didapatkan ketika parameter C = 1. Setelah melakukan proses *testing*, maka akan dicari nilai akurasi terbaik dengan menggunakan *confusion matrix* pada tahapan *evaluation*. Untuk hasil nilai akurasi yang didapat dari confusion matrix dapat dilihat pada tabel 6.

TABLE VI NILAI AKURASI BERBAGAI KERNEL SVM

Rasio *Splitting Data*	Kernel			
	Linear	RBF	Sigmoid	Polynomial
60 : 40	0.9637	0.9176	0.8735	0.8884
70 : 30	0.9671	0.9333	0.8736	0.8952
80 : 20	0.9741	0.9351	0.8677	0.8833
90 : 10	**0.987**	0.9325	0.839	0.8909

Berdasarkan tabel 6 dapat dilihat bahwa nilai akurasi terendah terdapat pada kernel sigmoid dengan rasio data 90 : 10 yang mendapatkan nilai akurasi sebesar 0.839. Sementara nilai akurasi terbesar terdapat pada kernel linear dengan rasio data 90 : 10 yang mendapatkan nilai akurasi **0.987**.

Dengan hasil akurasi yang sangat besar yaitu mencapai **98.7%** pada penelitian ini membuktikan bahwa algoritma SVM dan teknik skoring untuk *labelling* memiliki performa yang sangat baik untuk mengklasifikasi data sentimen positif atau negatif.

Dari penelitian yang sudah dilakukan dapat dievaluasi

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020

bahwa tingkat akurasi ditentukan oleh proses *preprocessing*, semakin bersih data yang akan diolah semakin bagus juga algoritma SVM dalam mengklasifikasi data. Selain itu hal berpengaruh lainnya untuk menghasilkan akurasi yang bagus adalah jumlah *data training* dan *data testing*, dan pemilihan kernel.

V. KESIMPULAN

Setelah dataset diberikan label sentimen positif atau negatif, untuk mengklasifikasikan dataset tersebut menggunakan SVM maka dataset tersebut harus ditransformasikan terlebih dahulu menjadi data numerik karena SVM hanya bisa memproses data numerik saja. Proses transformasi menggunakan teknik TF *(term frequency*) yaitu menghitung frekuensi kemunculan kata pada tweet. Setelah selesai di transformasi menjadi data numerik selanjutnya dataset di split menjadi data training dan data testing yaitu dengan pembagian rasio data 60:40, 70:30, 80:20, dan 90:10. Setelah splitting data maka proses klasifikasi dengan algoritma SVM sudah dapat dilakukan dengan menggunakan berbagai kernel yang tersedia pada algoritma SVM.

Dari sebanyak 16 kali percobaan klasifikasi, dengan 4 *data testing* dan 4 kernel yang berbeda-beda, nilai akurasi terbaik didapatkan oleh kernel linear dengan rasio data 90% (3467 data) untuk *data training* dan 10% (385 data) untuk *data testing* dengan nilai akurasi yang didapat yaitu sebesar 0.987 atau 98.7%.

REFERENCES

[1] Suharni, "Uang Elektronik (E-Money) ditinjau dari Perspektif Hukum dan Perubahan Sosial," *Jurnal Spektrum Hukum*, pp. 15-43, 2018.

[2] "Kata Data," 25 September 2019. [Online]. Available: https://katadata.co.id/berita/2019/09/25/ovo-jadi-dompet-digital-terbesar-di-indonesia-berkat-ekosistem-grab.

[3] "CNN Indonesia," 29 November 2019. [Online]. Available: https://www.cnnindonesia.com/ekonomi/20191129080739-92-452526/bendera-putih-lippo-bakar-uang-untuk-ovo.

[4] E. M. Sipayung, H. Maharani and I. Zefanya, "Perancangan Sistem Analisis Sentimen Komentar Pelanggan Menggunakan Metode Naive Bayes Classifier," *Jurnal Sistem Informasi (JSI), Vol. 8, NO.1,* pp. 958-965, 2016.

[5] I. A. Muis and M. Affandes, "Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet," *Journal of Science Technology and Industry*, vol. 12, no. 2, pp. 189-197, 2015.

[6] Neneng, K. Adi and R. R. Isnanto, "Support Vector Machine Untuk Klasifikasi Citra Jenis Daging Berdasarkan Tekstur Menggunakan Ekstraksi Ciri Gray Level Co-Occurance Matrices (GLCM)," *Jurnal Sistem Informasi Bisnis*, pp. 1-10, 2016.

[7] M. Awad and R. Khanna, Efficient Learning Machines : Theories, Concepts, Applications for Engineers and System Designers, New York City: Apress, 2015.

[8] R. Munawarah, O. Soesanto and M. R. Faisal, "Penerapan Metode Support Vector Machine Pada Diagnosa Hepatitis," *Kumpulan JurnaL Ilmu Komputer (KLIK)*, vol. 04, no. 01, pp. 103-113, 2016.

[9] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Jurnal Edik Informatik*, pp. 213-219, 2017.

[10] M. F. Fibrianda and A. Bhawiyuga, "Analisis Perbandingan Akurasi Deteksi Serangan Pada Jaringan Komputer Dengan Metode Naïve Bayes Dan Support Vector Machine (SVM)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, pp. 3112-3123, 2018.

[11] F. R. Wibowo, D. S. Rusdianto and A. Arwan, "Pengembangan Sistem Pengumpulan Promo E-Commerce Berbasis Website Dengan Menerapkan Teknik Web Scraping Dalam Proses Pengambilan Data Promo," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, pp. 2887-2893, 2019.

[12] D. Haryalesmana, "Github," 24 March 2017. [Online]. Available: https://github.com/masdevid/ID-OpinionWords.

[13] R. Mahendrajaya, G. A. Buntoro and M. B. Setyawan, "Analisis Sentimen Pengguna Gopay Menggunakan Metode Lexicon Based dan Support Vector Machine," *KOMPUTEK : Jurnal Teknik Universitas Muhammadiyah Ponorogo*, pp. 52-63, 2019.

[14] M. W. Sardjono, M. Cahyanti, M. Mujahidin and R. Arianty, "Pendeteksi Kesamaan Kata Untuk Judul Penulisan Berbahasa Indonesia Menggunakan Algoritma Stemming Nazief-Ardiani," *SEBATIK*, pp. 138-146, 2018.

[15] R. Diani, U. N. Wisesty and A. Aditsania, "Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker," *Ind. Journal On Computing Volume 2*, pp. 109-117, 2017.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v9i2.903, Copyright ©2020
Submitted : 23 Juni 2020, Revised : 17 Juli 2020, Accepted : 21 Juli 2020, Published : 22 Juli 2020