Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 5, October 2023, pp. 1157-1164

DOI: https://doi.org/10.52436/1.jutif.2023.4.5.1014
p-ISSN: 2723-3863
e-ISSN: 2723-3871

IMPLEMENTATION OF TEXT PROCESSING FOR SENTIMENT ANALYSIS OF TAX PAYMENT INTEREST AFTER THE "RUBICON" PHENOMENON

Ridian Gusdiana[1], Iqbal Alfian[2], Christina Juliane[*3]

[1,2,3]Master of Information Systems, STMIK LIKMI Bandung, Indonesia
Email: [1]ridiangusdiana@uniga.ac.id, [2]musikiqbal@gmail.com, [3]christina.juliane@likmi.ac.id

(Article received: April 27, 2023; Revision: May 22, 2023; published: October 15, 2023)

Abstract

In February 2023, an incident occurred involving the child of an official from the Indonesian Directorate General of Taxes who committed violence against a member of the GP Ansor organization. The news spread widely and brought a new issue, namely suspicious reporting of the official's wealth with an amount of up to 56 billion Indonesian Rupiahs. In order to determine public sentiment towards the "RUBICON" case, which was receiving attention, sentiment analysis of tax payment interest was conducted using text mining techniques. Data processing was done using the R language and RStudio application, taking a dataset of 23,785 tweets from the public about paying taxes on Twitter. Next, text cleaning was done to remove numbers, symbols, and URLs, as well as text processing using stemming, tokenizing, stopword removal, and TF-IDF methods. The TF-IDF method shows that the words "rafael" and "case" are the top keywords. This study used a supervised model by comparing SVM, KNN, and Naive Bayes algorithms, and evaluation was done using a confusion matrix with accuracy results in descending order of 0.8922, 0.8049, and 0.7369. The conclusion of this study is that the SVM algorithm successfully classified sentiment with the highest level of accuracy and obtained the highest negative sentiment of 5,616 sentences.

Keywords: *K-Nearest Neighbor, Naïve Bayes, Sentiment Analisys,, Support Vector Machine, Tax.*

IMPLEMENTASI TEXT PROCESSING UNTUK ANALISIS SENTIMEN MINAT BAYAR PAJAK SETELAH FENOMENA *"RUBICON"*

Abstrak

Bulan Februari 2023, terjadi sebuah peristiwa yang melibatkan anak dari pejabat Ditjen Pajak Indonesia yang melakukan kekerasan terhadap anak seorang pengurus GP Ansor. Berita ini menyebar luas dan membawa masalah baru, yaitu adanya dugaan pelaporan harta kekayaan pejabat yang mencurigakan dengan jumlah mencapai 56 miliar Rupiah. Dalam rangka untuk mengetahui sentimen masyarakat terhadap kasus "*RUBICON*" yang tengah menjadi perhatian, dilakukan analisis sentimen minat bayar pajak menggunakan teknik *text mining*. Pengolahan data dilakukan menggunakan bahasa *R* dan aplikasi *RStudio* dengan mengambil dataset dari komentar masyarakat tentang bayar pajak di media sosial *Twitter* sebanyak 23.785 twit. Selanjutnya, dilakukan *text cleaning* untuk menghapus angka, simbol, dan alamat *URL*, serta *text processing* dengan metode *stemming*, *tokenizing*, *stopword removal*, dan *TF-IDF*. Metode *TF-IDF* menunjukkan bahwa kata "rafael" dan "kasus" adalah topik pada kata kunci tertinggi. Penelitian ini menggunakan model *supervised* dengan membandingkan algoritma *SVM, KNN*, dan Naïve Bayes. Evaluasi dilakukan menggunakan *confusion matrix* dengan hasil akurasi secara berurutan adalah 0.8922, 0.8049 dan 0.7369. Kesimpulan dari penelitian ini adalah algoritma *SVM* berhasil mengklasifikasikan sentimen dengan tingkat akurasi tertinggi dan mendapatkan sentimen tertinggi adalah negatif sebesar 5.616 kalimat.

Kata kunci: *K-Nearest Neighbor, Naïve Bayes, Pajak, Sentimen Analisis, Support Vector Machine.*

1. PENDAHULUAN

Peristiwa yang terjadi pada awal tahun 2023 adalah tentang kasus "*RUBICON*". Kasus yang melibatkan anak Ditjen Pajak Indonesia dan anak dari salah satu pengurus Organisasi GP Ansor. Kasus ini berkembang hingga dugaan pelaporan harta kekayaan mencurigakan. Kondisi itu membuat pertanyaan baru yaitu peminatan bayar pajak oleh masyarakat. Sumber data pada penelitian diperoleh dari penambangan komentar masyarakat dengan kata kunci "bayar pajak". Penambangan dilakukan pada periode 06 Maret sampai 22 Maret 2023 dan menghasilkan 23.785 data twit.

Sebelumnya telah dilakukan penelitian tentang *text mining* dengan algoritma *KNN* [1]. Data didapatkan dari aplikasi Layanan Aspirasi dan Pengaduan Online Rakyat (LAPOR) dengan jumlah 103 data. Metode yang digunakan adalah algoritma *KNN* untuk membuat klasifikasi keakuratan laporan. Hasilnya yang diperoleh adalah pembobotan TF.RF menggunakan Naïve Bayes terbukti lebih baik dengan tingkat akurasi sebesar 98,67%, Precision 93,81%, dan Recall 96,67%. Berikutnya penelitian dilakukan untuk menganalisis opini masyarakat tentang kinerja layanan transportasi online dengan menggunakan text mining dan metode *SVM* [2]. Data yang digunakan diperoleh melalui penambangan *Twitter* dengan kata kunci "grab" dan "gojek". Metode yang digunakan adalah sentiment analisis untuk mendapatkan klasifikasi data berdasarkan tanggapan positif dan negatif. Hasil dari penelitian tersebut mendapatkan nilai akurasi sentiment objek Grab sebesar 74,34% dan hasil akurasi pada objek Gojek sebesar 68,84%. Penelitian terakhir membahas tentang penerapan *text mining* pada analisis sentimen pengguna *Twitter* terhadap *marketplace* di Indonesia dengan menggunakan algoritma *Support Vector Machine (SVM)* [3]. Data didapatkan dari *twitter* menggunakan kata kunci Tokopedia, Shoppe dan Bukalapak. Metode yang digunakan adalah *SVM* dan mendapatkan akurasi nilai sentiment positif sebesar 0,85 serta sentiment negatif 0,86.

Berdasarkan permasalah tersebut, maka penelitian ini menggunakan teknik *text mining* dalam mengklasifikasikan sentimen masyarakat terhadap minat bayar pajak. Tujuan penelitian untuk membandingkan algoritma *KKN*, *Naïve Bayes*, dan *SVM* untuk mengetahui akurasi terbaik.

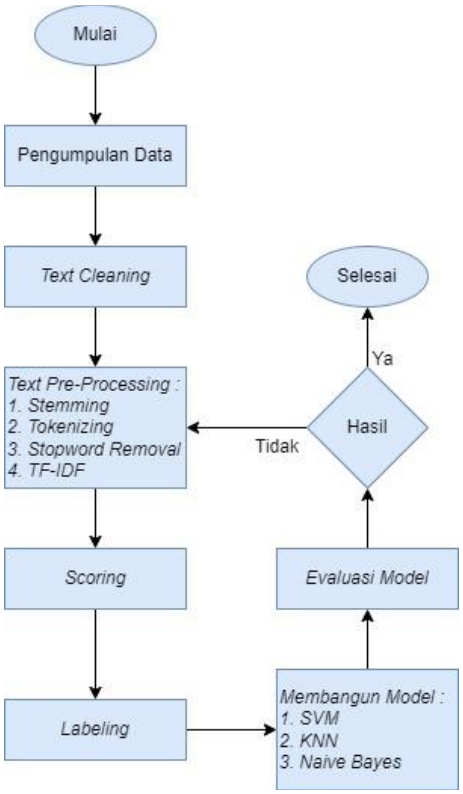
2. METODE PENELITIAN

2.1. Prosedur Penelitian

Kegiatan pertama dalam penelitian ini adalah mengumpulkan data, kemudian melakukan pre-processing. Selanjutnya, dilakukan pembuatan skor, penandaan label, berikutnya adalah pembangunan model *KNN*, *SVM*, dan *Naïve Bayes*. Terakhir yaitu evaluasi model sentiment menggunakan algoritma *SVM*, *KNN* dan *Naïve Bayes*.. Gambar 1 merupakan *Flowchart* prosedur penelitian.

2.2. Pajak

Pajak merupakan suatu kewajiban bagi individu atau organisasi untuk memberikan kontribusi kepada negara sesuai dengan Undang-Undang, tanpa imbalan langsung, dan digunakan untuk memajukan kemakmuran rakyat [4]. Beberapa klasifikasi pajak termasuk pajak penghasilan, pajak pertambahan nilai (PPN), pajak properti (PBB), pajak kendaraan bermotor, dan sebagainya. Pajak penghasilan adalah pajak yang dikenakan pada penghasilan seseorang atau badan usaha dalam satu tahun pajak. PPN adalah pajak yang dikenakan pada penjualan barang dan jasa, sedangkan PBB adalah pajak yang dikenakan pada kepemilikan tanah dan bangunan [5]. Pajak juga dapat digunakan sebagai alat untuk mengendalikan inflasi dan defisit anggaran negara. Dalam menjalankan fungsinya sebagai alat pengendalian inflasi, pajak digunakan sebagai instrumen kebijakan fiskal dengan mengatur jumlah uang yang beredar di masyarakat.



Gambar 1. Prosedur Penelitian

2.3. *Text Mining*

Identifikasi pola dari suatu masalah dapat dilakukan dengan beberapa cara, salah satunya yaitu dengan metode *text mining*. *Text mining* merupakan proses eksplorasi dan analisis data teks yang tak terstruktur dengan bantuan perangkat lunak yang dapat mengidentifikasi konsep, pola, topik, kata kunci, dan atribut lainnya dalam data [6]. Pada penerapannya, *Text Mining* merupakan sebuah metode yang mengolah data dalam bentuk kata-kata untuk mendapatkan informasi. Area penerapan *Text Mining* meliputi Ekstraksi Informasi (*Information Extraction*), Pelacakan Topik (*Toping Tracking*), Perangkuman (*Summarization*), Kategorisasi (*Categorization*), Penggugusan (*Clustering*), Penautan Konsep (*Concept Lingking*), Penjawaban Pertanyaan (*Question Answering*) [7]. Proses yang dilakukan pada *text mining*, yaitu *Text Cleaning* dan *Text Processing*.

2.4. *K-Nearest Neighbor* (*KNN*)

K-Nearest Neighbor (K-NN) adalah suatu metode yang menggunakan algoritma *supervised*

dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada *K-NN* [8]. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasakan atribut dan sampel latih. Pengklasifikasian tidak menggunankan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik uji, akan ditemukan sejumlah *K* objek (titik *training*) yang paling dekat dengan titik uji. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari *K* objek. Algoritma *K-NN* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru [9].

2.5. *Support Vector Machine (SVM)*

SVM adalah sistem pembelajaran yang menggunakan fungsi linear pada ruang fitur berdimensi tinggi dan dilatih pada parameter dengan algoritma pembelajaran berdasarkan teori optimasi dan pembelajaran statistik [10]. Metode SVM mencari hyperplane terbaik untuk memisahkan dua kelas dengan maksimalkan jarak antara kelas tersebut pada ruang kelas berdimensi tinggi. Hyperplane berfungsi untuk melakukan klasifikasi pada ruang kelas dengan dimensi yang lebih tinggi [11]. *SVM* menggunakan teknik kernel untuk mengubah data ke dimensi yang lebih tinggi untuk memisahkan data secara linier. Beberapa jenis kernel yang biasa digunakan antara lain kernel linier, polinomial, dan fungsi basis radial [12].

2.6. *Naïve Bayes*

NB adalah suatu algoritma pembelajaran mesin yang mengaplikasikan prinsip teorema Bayes dengan asumsi yang sederhana [13]. Konteks teori Bayes, probabilitas yang paling mungkin dari suatu peristiwa dapat dihitung berdasarkan data yang sudah ada. [14].

2.7. *Confussion Matrix*

Confusion matrix adalah salah satu cara untuk mengevaluasi performa suatu algoritma klasifikasi. [15]. Secara umum, *confusion matrix* membandingkan hasil klasifikasi sistem dengan hasil yang seharusnya dan terdiri dari empat istilah: *True Positive (TP), True Negative (TN), False Positive (FP) dan False Negative (FN). True Negative (TN)*.

3. HASIL DAN PEMBAHASAN

3.1. Pengumpulan Data

Dataset diperoleh melalui media sosial *twitter* dari tanggal 06 Maret – 22 Maret 2023 sebanyak 23.785 *tweet* dengan dua atribut yaitu, nomor dan text yang ditunjukan pada Tabel 1.

Tabel 1. Hasil Pengumpulan Data

No	Text
1	RT @sport7trans7: Pelatih bergaji Rp282 miliar per tahun, tapi naik pesawat kelas ekenomi. Coba jadi pejabat pajak. Eh. ðŸ˜• https://t.co/z5CB34KtcD
2	RT@vctrkmng: Diingetin batas akhir lapor SPT malah keinget soal betapa hedonnya pejabat pajak. Jadi memacu diri euG tuk lebi giat agar bisa lebih hedon dari mereka ðŸ˜¡
3	RT@TetiDominicus: @PartaiSocmed Lu pade kerja banting tulang dapetnya ga sbrp krn kepotong pajak. Beli rumah kena pajak. Beli tanah mobil dsb2 kena pajak. Trus trnyata uangnya dipake buat beli sendal 12 jt ama bocahnya pejabat. Uang lu itu, dipake. Pantes klo ga bayar pajak dihukum berat, pejabatnya BU ginian.
-	-
-	-
-	-
23.785	RT @baronlukater Kayaknya udah lumrah, mobil² mewah gitu ternyata nama di dokumennya adalah orang² yg jauh dari kata mewah, kadang mereka menggunakan KTP orang lain spt supir, OB, atau ART buat membeli barang mewah, untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb

3.2. *Text Cleaning*

Selanjutnya dilakukan pembersihan teks dengan cara menghapus simbol, angka, alamat *url*, menghapus baris kosong dan menghapus data duplikat yang ditunjukan pada Tabel 2.

Tabel 2. Hasil *Text Cleaning*

No	Sebelum	Sesudah
1	RT @sport7trans7: Pelatih bergaji Rp282 miliar per tahun, tapi naik pesawat kelas ekenomi. Coba jadi pejabat pajak. Eh. ðŸ˜• https://t.co/z5CB34KtcD	pelatih bergaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi pejabat pajak eh
2	RT@vctrkmng: Diingetin batas akhir lapor SPT malah keinget soal betapa hedonnya pejabat pajak. Jadi memacu diri euG tuk lebi giat agar bisa lebih hedon dari mereka ðŸ˜¡	diingetin bata akhir lapor spt malah keinget soal betapa hedonnya pejabat pajak jadi memacu diri eug tuk lebi giat agar bisa lebih hedon dari mereka
3	RT@TetiDominicus: @PartaiSocmed Lu pade kerja banting tulang dapetnya ga sbrp krn kepotong pajak. Beli rumah kena pajak. Beli tanah mobil dsb2 kena pajak. Trus trnyata uangnya dipake buat beli sendal 12 jt ama bocahnya pejabat. Uang lu itu, dipake. Pantes klo ga bayar pajak dihukum berat, pejabatnya BU ginian.	lu pade kerja bant tulang dapetnya ga sbrp krn kepotong pajak beli rumah kena pajak beli tanah mobil dsb kena pajak trus trnyata uangnya dipak buat beli sendal jt ama bocahnya pejabat uang lu itu dipak pant klo ga bayar pajak dihukum berat pejabatnya bu ginian
-	-	-
-	-	-
-	-	-
10.873	RT @baronlukater Kayaknya udah lumrah, mobil² mewah gitu ternyata nama di dokumennya adalah orang² yg jauh dari kata mewah, kadang mereka menggunakan KTP orang lain spt supir, OB, atau	kayaknya udah lumrah mobil mewah gitu ternyata nama di dokumennya adalah orang yg jauh dari kata mewah kadang mereka menggunakan ktp orang lain spt supir ob atau abuat memb

No	Sebelum	Sesudah
	ART buat membeli barang mewah, untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb	barang mewah untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb

3.3. *Text Preprocessing*

Tahap pertama dalam *Text Preprocessing* adalah *Stemming*, yaitu mengganti kata-kata yang berimbuhan menjadi kata dasar. *Stemming* dilakukan dengan menggunakan bahasa indonesia yang bersumber pada Kamus Besar Bahasa Indonesia. Hasil *Stemming* ditampilkan pada Tabel 3.

Tabel 3 *Text Stemming*

No	Sebelum	Sesudah
1	pelatih bergaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi pejabat pajak eh	latih gaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi jabat pajak eh
2	diingetin bata akhir lapor spt malah keinget soal betapa hedonnya pejabat pajak jadi memacu diri eug tuk lebi giat agar bisa lebih hedon dari mereka	diingetin bata akhir lapor spt malah keinget soal betapa hedonnya jabat pajak jadi pacu diri eug tuk lebi giat agar bisa lebih hedon dari mereka
3	lu pade kerja bant tulang dapetnya ga sbrp krn kepotong pajak beli rumah kena pajak beli tanah mobil dsb kena pajak trus trnyata uangnya dipak buat beli sendal jt ama bocahnya pejabat uang lu itu dipak pant klo ga bayar pajak dihukum berat pejabatnya bu ginian	lu pade kerja bant tulang dapetnya ga sbrp krn potong pajak beli rumah kena pajak beli tanah mobil dsb kena pajak trus trnyata uang pak buat beli sendal jt ama bocah jabat uang lu itu pak pant klo ga bayar pajak hukum berat jabat bu ginian
-	-	-
-	-	-
-	-	-
10.873	kayaknya udah lumrah mobil mewah gitu ternyata nama di dokumennya adalah orang yg jauh dari kata mewah kadang mereka menggunakan ktp orang lain spt supir ob atau abuat memb barang mewah untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb	kayak udah lumrah mobil mewah gitu nyata nama di dokumen adalah orang yg jauh dari kata mewah kadang mereka guna ktp orang lain spt supir ob atau abuat b barang mewah untuk hindar pajak progresif atau asal usul uang untuk beli mobil tsb

Proses kedua yaitu *Tokenization* yang betujuan untuk memecah kalimat hasil *Stemming* menjadi kata-kata. Tabel 4 menunjukan data hasil tokenisasi.

Tabel 4. *Text Tokenization*

No	Sebelum	Sesudah
1	pelatih bergaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi pejabat pajak eh	latih gaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi jabat pajak eh

Proses ketiga yaitu melakukan metode *stopword removal* untuk menghilangkan kata-kata tidak bermakna. Kata-kata dapat dipilih berdasarkan kemunculan pada hasil tokenisasi. Terpilih sebanyak 12,912 kata telah dihilangkan dan ditampilkan pada Tabel 5.

Tabel 5 *Stopword Removal*

No	Sebelum	Sesudah
1	latih	latih
	gaji	gaji
	rp	rp
	miliar	miliar
	per	
	tahun	
	tapi	
	naik	
	pesawat	pesawat
	kela	kela
	ekenomi	ekenomi
	coba	coba
	jadi	
	jabat	jabat
	pajak	pajak
	eh	eh

Proses keempat adalah melakukan perhitungan *Document Frequenciy (DF), Term Frequency (TF), Inverse Document Frequency (IDF)* dan *Term Frequency/Inverse Document Frequency (TF/IDF).* Tujuan dilakukannya perhitungan tersebut agar dapat melihat frekuensi kata, Tabel 6 hasil proses Frekuensi Kata dan Gambar 2 adalah *wordclound* frekuensi kata.

Tabel 6 Frekuensi Kata

no	word	n
1	pajak	10956
2	pajak	7178
3	bayar	4014
4	pejabat	2964
5	bayar	2771
6	rafael	2484
7	alun	2168

Gambar 2. Frekuensi Kata

Proses kelima yaitu pembobotan kata-kata menggunakan metode *TF-IDF* yang akan disajikan pada Tabel 7. Lalu, mencari frekuensi keterkaitan yang ditunjukan pada Tabel 8. selanjutkan hasil *wordcloud* ditampilkan pada Gambar 3.

Tabel 7 *TF-IDF*

no	word	n	tf	idf	Tf_idf
1	pajak	10956	0.05068	0	0
2	pajak	7178	0.0602	0	0
3	bayar	4014	0.018568	0	0
4	pejabat	2964	0.013711	0	0
5	bayar	2771	0.02324	0	0
6	rafael	2484	0.01149	0.133531	0.001534
7	alun	2168	0.010029	0.133531	0.001339
8	rafael	1981	0.059954	0.133531	0.008006
9	alun	1891	0.05723	0.133531	0.007642
10	kasus	1820	0.008419	0.133531	0.001124

Tabel 8 Keterkaitan Kata

X	word	n
1	bayar pajak	3033
2	rafael alun	1706
3	pajak kendara	745
4	jabat pajak	728
5	alun trisambodo	614
6	gawai pajak	382
7	sri mulyani	297
8	pamer harta	259
9	hai kak	255
10	ditjen pajak	252

Gambar 3. Keterkaitan Kata

3.4. *Text Processing*

Selanjutnya dilakukan *text processing* dengan langkah pertama yaitu pembuatan skor sentimen. Skor -4 sampai -1 adalah sentimen negatif, skor 0 sampai 4 adalah sentimen positif. Tabel 9 adalah hasil perhitungan skor sentimen pada dataset kalimat.

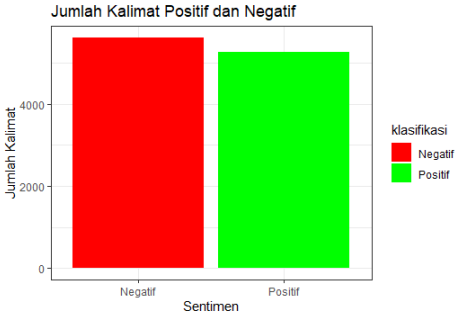
Tabel 9 *Text Scoring*

No	score	text
1	0	pelatih bergaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi pejabat pajak eh
2	-1	diingetin bata akhir lapor spt malah keinget soal betapa hedonnya pejabat pajak jadi memacu diri eug tuk lebi giat agar bisa lebih hedon dari mereka
3	-4	lu pade kerja bant tulang dapetnya ga sbrp krn kepotong pajak beli rumah kena pajak beli tanah mobil dsb kena pajak trus trnyata uangnya dipak buat beli sendal jt ama bocahnya pejabat uang lu itu dipak pant klo ga bayar pajak dihukum berat pejabatnya bu ginian
-	-	-
-	-	-
-	-	-
10.873	-2	kayaknya udah lumrah mobil mewah gitu ternyata nama di dokumennya adalah orang yg jauh dari kata mewah kadang mereka menggunakan ktp orang lain spt supir ob atau abuat memb barang mewah untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb

Selanjutnya membuat label berdasarkan hasil skor yang telah dibuat yang ditunjukan oleh Tabel 10 dan ditampilkan *barplot* pada Gambar 4.

Tabel 10. Pelabelan

no	klasifikasi	score	text
1	Positif	0	pelatih bergaji rp miliar per tahun tapi naik pesawat kela ekenomi coba jadi pejabat pajak eh
2	Negatif	-1	diingetin bata akhir lapor spt malah keinget soal betapa hedonnya pejabat pajak jadi memacu diri eug tuk lebi giat agar bisa lebih hedon dari mereka
3	Negatif	-4	lu pade kerja bant tulang dapetnya ga sbrp krn kepotong pajak beli rumah kena pajak beli tanah mobil dsb kena pajak trus trnyata uangnya dipak buat beli sendal jt ama bocahnya pejabat uang lu itu dipak pant klo ga bayar pajak dihukum berat pejabatnya bu ginian
-	-	-	-
-	-	-	-
-	-	-	-
10.873	Negatif	-2	kayaknya udah lumrah mobil mewah gitu ternyata nama di dokumennya adalah orang yg jauh dari kata mewah kadang mereka menggunakan ktp orang lain spt supir ob atau abuat memb barang mewah untuk menghindari pajak progresif atau asal usul uang untuk beli mobil tsb



Gambar 4. Jumlah Kalimat Positif dan Negatif

4. PERBANDINGAN ALGORITMA

4.1. *Support Vector Machine (SVM)*

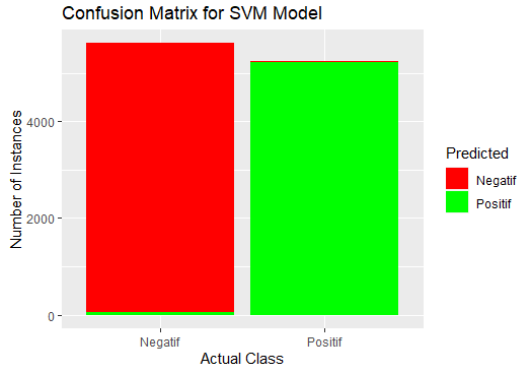
Pertama yang dilakukan membuat dataset menjadi dua bagian yaitu data latih dan uji. Hasil

pemisahan tersebut adalah data train 250:10873 dan data uji adalah 1:10873.

Kedua memilih *kernel* yang sesuai yaitu "*Linear*". Ketiga melakukan pelatihan data dengan *kernel* yang telah dipilih. Keempat evaluasi menggunakan metode *confussion matrix*. Kegiatan tersebut akan ditampilkan pada Tabel 11 dan Gambar 5 untuk evaluasi.

Tabel 11. Bangun Model *SVM*

variable	operator	function
container	=	create_container(dtMatrix, cf$klasifikasi, trainSize = 250:10873, testSize = 1:10873, virgin = FALSE)
model	=	train_model(container, "SVM", kernel = "linear", cost = 1)
result	=	classify_model(container, model)
result$SVM_LABEL	=	factor(result$SVM_LABEL)
confussionMatrix	=	confusionMatrix(cf$klasifikasi[1:10873], result[,"SVM_LABEL"])



Gambar 5. Hasil evaluasi *Confussion Matrix SVM*

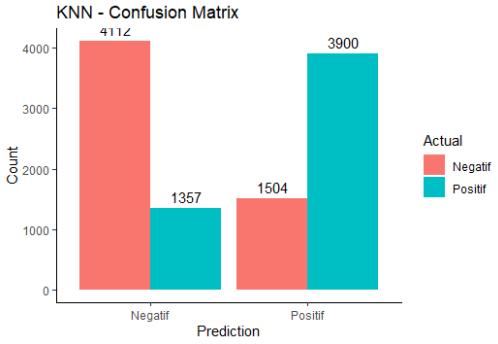
4.2. *K-Nearest Neighbor (KNN)*

Pertama yang dilakukan memilih jumlah tetangga terdekat. Kedua menghitung jarak untuk data baru dan kepada semua titik yang ada pada dataset. Ketiga mengurutkan langkah kedua lalu ambil titik data dengan jarak terkecil dalam jumlah *K* titik. Keempat menghitung rata-rata nilai pada variabel pada tiap titik. Kelimat memasukan rata-rata itu sebagai hasil output.

Proses tersebut membutuhkan waktu yang lama jika jumlah data besar. Tabel 12 adalah kode dalam bahasa *R* untuk membangun algoritma *KNN* dan Gambar 6 adalah hasil evaluasi.

Tabel 12. Bangun Model *KNN*

variable	**operator**	**function**
i	=	1
k.optm	=	1 for (i in 1:3) {knn.mod <- knn(train = dtm.train, test = dtm.test, cl = cf.train$klasifikasi, k=i) k.optm[i] <- 100 * sum(cf.test$klasifikasi == knn.mod)/NROW(cf.test)k=i cat(k,'=',k.optm[i],")}
model	=	kknn(klasifikasi~., data_train1, data_test1, k=20, kernel = "optimal")
perform	=	confusionMatrix(model$fitted.values, data_test1$klasifikasi)
xval	=	train.kknn(klasifikasi~., data_train1, kmax = 20, kernel = c("optimal", "rectangular", "inv", "gaussian", "triangular"), scale = T)



Gambar 6. Hasil evaluasi *Confussion Matrix KNN*

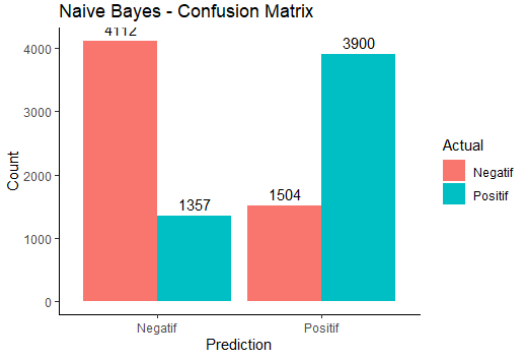
4.3. *Naïve Bayes*

Pertama yang dilakukan membuat dataset menjadi dua bagian, yaitu data latih dan uji. Hasil pemisahan tersebut adalah data train 5000:10873 dan data uji adalah 1:10873. Kedua adalah menghitung probabilitas kelas untuk setiap kelas yang mungkin dalam data pelatihan. Ketiga Urutkan jarak pada langkah 2 dari kecil ke besar, lalu ambil titik data dengan jarak terkecil sejumlah K titik. Keempat Menghitung probabilitas setiap fitur dalam setiap kelas dalam data pelatihan. Kelima Menggunakan probabilitas yang dihitung paling mungkin untuk setiap data uji dalam data pengujian.

Kegiatan tersebut ditampilkan pada Tabel 13 dan Gambar 7 untuk evaluasi.

Tabel 13. Bangun Model *Naive Bayes*

variable	**operator**	**function**
df.train	=	df[5000:10873,]
df.test	=	df[1:10873,]
convert_countNB	<-	function(x) {y <- ifelse(x > 0, 1,0) y <- factor(y, levels=c(0,1), labels=c("No", "Yes"))y}
classifier	<-	naiveBayes(trainNB, df.train$klasifikasi, laplace = 1)
pred	<-	predict(classifier, testNB)
NB_table	=	table("Prediction"= pred, "Actual" = df.test$klasifikasi)



Gambar 7. Hasil evaluasi *Confussion Matrix Naive Bayes*

Perbandingan ketiga algoritma tersebut mendapatkan masing-masing nilai akurasi yang berbeda, lalu ditampilkan pada Tabel 14.

Tabel 14. Hasil Evaluasi Perbandingan Algoritma

Perbandingan	SVM	KNN	Naive Bayes
Negative	5564	4212	4112
Positive	5524	3700	3900
Accuracy	0.8922	0.8049	0.7369
Balanced Accuracy	0.8921	0.8054	0.7370

5. DISKUSI

Data yang dihasilkan dari media sosial *twitter* berjumlah 23.785 *tweet*. Dataset berbentuk kalimat opini masyarakat. Lalu, diolah menggunakan teknik *text cleaning* sehingga jumlah data berubah menjadi 10.873 dikarena terdapat duplikasi data berjumlah 12,912 *tweet*.

Proses langkah *text processing* dengan teknik tokenisasi menghasilkan 235.581 kata yang bersumber pada kalimat hasil proses *text cleaning*. Selanjutnya pada teknik stopword removal terdapat perubahan jumlah kata menjadi 157.905 dikarena menghapus kata-kata yang tidak relevan sebanyak 77,676 kata.

Frekuensi kata yang sering muncul adalah ”pajak” dengan jumlah 10.079 kata dan frekuensi keterkaitan antar kata menghasilkan kata ”bayar pajak” menjadi yang tertinggi dengan jumlah 3.033 kata.

Pembuatan *score* pada kalimat mendapatkan 4 yang tertinggi untuk Klasifikasi Positif dan -4 terendah untuk Klasifikasi Negatif. Klasifikasi Negatif mendapatkan jumlah 5616 kalimat dan klasifikasi Positif berjumlah 5257 kalimat.

Perbandingan algoritma tertinggi didapatkan pada *SVM* dengan akurasi 0.8922, kemudian *KNN* mendapatkan akurasi 0.8094 dan *Naive Bayes* menjadi yang terendah dengan akurasi 0.7369.

6. ANALISIS HASIL

Fenomena yang terjadi pada nilai akurasi adalah *text mining* tidak selalu dapat mendeteksi sarkasme dengan akurasi yang tinggi. Hal ini dikarenakan sarkasme sendiri dapat ditampilkan dalam berbagai bentuk dan dapat menjadi sangat bergantung pada konteks. Selain itu, sarkasme seringkali disampaikan dengan menggunakan konvensi linguistik tertentu, seperti menggunakan kata-kata yang berlawanan dengan makna yang sebenarnya, atau dengan menambahkan intonasi tertentu dalam kalimat yang disampaikan.

Saran pada permasalahan tersebut, perlu dilakukan penelitian lebih lanjut terkait sarkasme untuk mendapatkan nilai akurasi klasifikasi yang lebih baik dalam *text mining*. Terdapat 2.861 kalimat yang berisikan sarkastik ditampilkan pada Tabel 15.

Tabel 15. Kalimat Sarkastik

no	klasifikasi	score	text
1	Positif	1	yuk bayar pajak yuk biar pejabat tdk melarat yuk
2	Positif	0	sangat hebat utang menumpuk rakyat bayar pajak uangnya dimakan pejabat pajak dan gak bayar pajak banyak deh lainnya
-	-	-	-
-	-	-	-
-	-	-	-
2.861	Positif	0	tenang ini hanya perpindahan uang dari masyarakat ke pejabat daerah lalu ke kang pajak mereka jg bagian dari masyarakat ingat dari masyarakat untuk masyarakat tetap semangat semua untuk sale memberi

7. KESIMPULAN

Penelitian ini menggunakan teknik *text mining* untuk membuat klasifikasi sentimen analisis minat bayar pajak. Hasilnya menunjukkan bahwa terdapat lebih banyak kalimat dengan sentimen negatif (5.616 kalimat) daripada positif (5.257 kalimat). Perbandingan algoritma berhasil dilakukan dan *SVM* memiliki tingkat akurasi tertinggi (0.8922), diikuti oleh *KNN* (0.8094) serta *Naive Bayes* (0.7369).

Jumlah sentimen negatif yang tinggi terhadap pajak menunjukkan bahwa adanya indikasi ketidakpuasan atau ketidakpercayaan masyarakat terhadap sistem perpajakan dan pemerintah yang mengelolanya. Bedasarkan hasil analisis sentimen yang telah dilakukan, maka saran yang diajukan adalah pemerintah perlu ketegasan terhadap kasus dugaan harta tidak wajar pada Ditjen Pajak Indonesia untuk menunjukkan komitmen dalam pemberantasan korupsi dan memperbaiki citra pemerintah.

DAFTAR PUSTAKA

[1] A. Deolika, K. Kusrini, and E. T. Luthfi, “Analisis Pembobotan Kata Pada Klasifikasi Text Mining,” *J. Teknol. Inf.*, vol. 3, no. 2, p. 179, 2019, doi: 10.36294/jurti.v3i2.1077.

[2] H. Irsyad and M. R. Pribadi, “Implementasi Text Mining Dalam Pengelompokan Data Tweet Pertanian Indonesia Dengan K-Means,” *KURAWAL J. Teknol. Inf. dan Ind.*, vol. 3, no. 2, pp. 164–172, 2020, [Online]. Available: https://t.co/FXtzMcbdHp

[3] D. A. Agustina, S. Subanti, and E. Zukhronah, “Implementasi Text Mining Pada Analisis Sentimen Pengguna Twitter Terhadap Marketplace di Indonesia Menggunakan Algoritma Support Vector Machine,” *Indones. J. Appl. Stat.*, vol. 3, no. 2, p. 109, 2021, doi:

10.13057/ijas.v3i2.44337.

[4] T. Arimurti *et al.*, "Pengaruh Leverage , Return on Asset (Roa) Dan Intensitas Modal Terhadap Penghindaran Pajak Dengan," *J. KRISNA Kumpul. Ris. Akunt.*, vol. 13, no. 2, 2022.

[5] S. Sanjaya and R. P. Sipahutar, "Pengaruh Current Ratio, Debt to Asset Ratio dan Total Asset Turnover terhadap Return on Asset pada Perusahaan Otomotif dan Komponennya yang Terdaftar di Bursa Efek Indonesia," *J. Ris. Akunt. dan Bisnis*, vol. 19, no. 2, pp. 136–150, 2019, doi: 10.30596/jrab.v19i2.4599.

[6] I. M. D. Ardiada, M. Sudarma, and D. Giriantari, "Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour," *Maj. Ilm. Teknol. Elektro*, vol. 18, no. 1, p. 55, May 2019, doi: 10.24843/mite.2019.v18i01.p08.

[7] A. S. Sri Widaningsih, "Klasifikasi Jurnal Ilmu Komputer Berdasarkan Pembagian," *Semin. Nas. Teknol. Inf. dan Komun. 2018 (SENTIKA 2018)*, vol. 2018, no. Sentika, pp. 320–328, 2018.

[8] F. Sodik and I. Kharisudin, "Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter," vol. 4, pp. 628–634, 2021.

[9] N. Tri Romadloni, I. Santoso, S. Budilaksono, and M. Ilmu Komputer STMIK Nusa Mandiri Jakarta, "Perbandingan Metode Naive Bayes, Knn Dan Decision Tree Terhadap Analisis Sentimen Transportasi Krl Commuter Line," *J. IKRA-ITH Inform.*, vol. 3, no. 2, 2019.

[10] I. Olive, D. Putra, K. R. Prilianti, P. Lucky, and T. Irawan, "Implementasi Text Mining untuk Analisis Layanan Transportasi Online dengan Analisis Faktor," *J. SimanteC*, vol. 8, no. 2, pp. 1–9, 2020.

[11] I. Oktanisa and A. A. Supianto, "Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank a Comparison of Classification Techniques in Data Mining for," *Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, pp. 567–576, 2018, doi: 10.25126/jtiik20185958.

[12] M. A. Ayu, E. Irawan, and T. Mantoro, "Text mining approaches for analyzing an Indonesian tafseer and translation of the Holy Quran," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 25, no. 3, 2022, doi: 10.11591/ijeecs.v25.i3.pp1469-1480.

[13] R. A. Wildan, R. A. Rajagede, and R. Rahmadi, "Analisis Sentimen Politik Berdasarkan Big Data dari Media Sosial Youtube : Sebuah Tinjauan Literatur," *Automata*, vol. 2, 2021.

[14] M. H. Asnawi, I. Firmansyah, R. Novian, and R. S. Pontoh, "Perbandingan Algoritma Naïve Bayes , K-NN , dan SVM dalam Pengklasifikasian Sentimen Media Sosial," 2021.

[15] R. Pratiwi, M. N. Hayati, and S. Prangga, "Perbandingan Klasifikasi Algoritma C5.0 Dengan Classification and Regression Tree (Studi Kasus : Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 2, pp. 273–284, 2020, doi: 10.30598/barekengvol14iss2pp273-284.