Analisis Sentimen Pemilihan Presiden Amerika 2020 di Twitter Menggunakan Naïve Bayes dan Support Vector Machine

Gery Nugroho[1], Danang Triantoro Murdiansyah[2], Kemas M Lhaksmana[3]

[1,2,3] Universitas Telkom, Bandung
[1]gerynugroho@students.telkomuniversity.ac.id, [2]danangtriantoro@telkomuniversity.ac.id, [3]kemasmuslim@telkomuniversity.ac.id

Abstrak

Analisis sentimen adalah suatu cara untuk mengekstrak emosi dari suatu teks. Tujuan dari analisis sentiment ini adalah untuk mengetahui sentiment positif atau negatif dalam suatu tweet dari twitter mengenai pemilihan Presiden Amerika 2020. Salah satu cara untuk menentukannya adalah dengan melakukan klasifikasi teks. Dengan melakukan klasifikasi teks, kita dapat melakukan prediksi sentimen dari suatu tweet. Namun terdapat suatu masalah yaitu banyaknya atribut yang dimiliki oleh suatu teks. Oleh karena itu dilakukan seleksi fitur menggunakan metode TF-IDF (Term Frequency – Inverse Document Frequency). TF-IDF merupakan teknik pembobotan suatu kata dalam dokumen. Pada penelitian ini peneliti mencoba membandingkan 2 algoritma klasifikasi yaitu Naïve Bayes dan Support Vector Machine. Hasil evaluasi menggunakan cross-validation dengan nilai K sebesar 10 serta menggunakan *mean approach* menunjukkan bahwa model memberikan hasil akurasi terbaik sebesar 82% menggunakan kernel linear. Berdasarkan 10000 data *tweet* mengenai Donald Trump dengan akurasi terbaik 82% model berhasil memprediksi 36.88% orang memiliki pandangan netral terhadap Trump, 30.78% orang memilki pandangan positif terhadap Trump, dan 32.34% memiliki pandangan pandangan negatif terhadap Trump. Lalu 10000 data *tweet* mengenai Joe Biden, model berhasil memprediksi atau 42.39% orang memiliki pandangan netral terhadap Biden, 29.62% orang memilki pandangan positif terhadap Biden, dan 27.99% memiliki pandangan pandangan negatif terhadap Biden

Kata kunci : Pemilihan Presiden Amerika 2020,Sentiment Analysis, SVM, Naïve Bayes, TF-IDF

Abstract

Sentiment analysis is a way to extract emotion from a text. The purpose of this sentiment analysis is to determine the positive or negative sentiment in a tweet from Twitter regarding the 2020 American Presidential election. One way to determine this is to classify the text. By doing text classification, we can predict the sentiment of a tweet. However, there is a problem that is the number of attributes a text has. Therefore, feature selection is carried out using the TFIDF (Term Frequency - Inverse Document Frequency) method. TF-IDF is a technique of weighting a word in a document. In this study, the researchers tried to compare 2 classification algorithms, namely naïve Bayes and support vector machines. The evaluation result of the machine learning using cross-validation with the value of K is 10 and using mean approach shows that the model reaches the highest accuracy with 82% using linear kernel. Based on 1000 data tweet about Donald Trump, SVM model with accuracy of 82% succeeded in predicting 3688 neutral label or 36.88% of the people think neutral about Trump, 3078 positive label or 30.78% of the people think positive about trump, and 3234 negative label or 32.34% of the people think negative about Trump. Then, Based on 1000 data tweet about Joe Biden, SVM model succeeded in predicting 4239 neutral label or 42.39% of the people think neutral about Biden, 2962 positive label or 29.62% of the people think positive about Biden, and 2799 negative label or 27.99% of the people think negative about Biden.

Keywords: 2020 American Presidential Election, sentiment analysis, SVM, Naïve Bayes, TF-IDF

1. Pendahuluan

1.1. Latar Belakang

Teknologi informasi dan komunikasi saat ini berkembang pesat. Persebaran informasi dengan mudah dilakukan di internet. Akibat perkembangan ini memunculkan media-media salah satunya media sosial. Media sosial saat ini banyak digunakan sebagai media untuk mengeluarkan opini seseorang terhadap suatu hal. salah satu media yang digunakan untuk mengeluarkan opini adalah Twitter. Twitter adalah media sosial yang digunakan untuk menyampaikan opini mengenai berbagai hal baik itu pribadi maupun publik. Pada Twitter pengguna dapat memposting opininya dan postingannya dapat dilihat oleh orang lain. Postingan dari pengguna Twitter biasanya disebut tweet.

Dengan banyaknya pengguna media sosial, khususnya Twitter. semakin banyak juga opini yang muncul. Setiap orang berhak untuk menyampaikan opininya dengan bebas mengenai masalah apapun. Pada 2020, sedang ramai-ramainya membicarakan mengenai Pemilihan Presiden Amerika 2020. Kandidat yang maju menjadi Calon

Presiden adalah Donald Trump dan Joe Biden. Kedua kandidat ini berkata bahwa pemilihan Presiden Amerika 2020 ini merupakan pemilihan terpenting sepanjang masa [1]. Berdasarkan riset dari *The Atlantic*, banyak yang tidak suka terhadap trump karena dia narsis, pembuli, rasis, tidak professional, dan memalukan terutama kaum perempuan [2]. untuk penduduk di negara lain, pemilihan ini bisa jadi sangat penting, terpilihnya kembali Trump akan menjadi sumber ketidakstabilan politik yang lebih luas [3]. Sedangkan terpilihnya Biden tidak akan banyak merubah aturan politik luar negeri dari Amerika Serikat [3]. Dikarenakan kontroversi ini, oleh karena itu pada penelitian ini peneliti ingin mengetahui apakah Pemilihan Presiden Amerika 2020 memiliki lebih banyak sentimen positif, negatif atau netral pada media sosial twitter. Hasil sentiment tersebut dapat digunakan untuk mengetahui keadaan masyarakat dan membantu pengambilan keputusan terkait kondisi yang ada.

1.2. Topik dan Batasannya

Berdasarkan masalah yang telah dijelaskan pada latar belakang, maka rumusan masalah dapat disusun sebagai berikut :

a. Bagaimana menentukan sentimen dari *tweet* pada Twitter dengan menggunakan algoritma Naïve Bayes dan Support Vector Machine?
b. Bagaimana performa sistem yang dibangun menggunakan algoritma Naive Bayes dan Support Vector Machine?

Adapun Batasan masalah yang ada pada penelitian ini sebagai berikut :

a. Data opini yang diambil hanya menggunakan data dari Twitter.
b. Hanya menggunakan data yang berbahasa inggris.
c. Tidak menangani data yang *misspellings*.

1.3. Tujuan

Tujuan penelitian yang dilakukan adalah untuk melakukan analisis sentimen dari *tweet* pada *Twitter* dengan menggunakan algoritma Naïve Bayes dan Support Vectore Machine lalu setelah itu untuk mengetahui performa dari Naïve Bayes dan Support Vector Machine.

Organisasi Tulisan

Sistematika penulisan pada penelitian ini dilakukan dengan bab pertama membahas mengenai permasalahan yang pada penelitian ini meliputi latar belakang, rumusan permasalahan, dan tujuan. Selanjutnya pada bab kedua merupakan segala informasi yang menjadi acuan dalam melakukan penelitian dan mencari penyelesaian masalaha pada bab pertama. Pada bab ketiga merupakan rancangan sistem yang dibangun. Setelah bab ketiga, hasil dari sistem yang dievaluasi pada bab keempat. Terakhir pada bab kelima berisi kesimpulan dan saran pada penelitian yang sudah dilakukan.

2. Studi Terkait

2.1. Analisis Sentimen

Analisis sentimen atau disebut juga *opinion mining* adalah bidang studi yang menganalisa opini, sentimen, evaluasi, penilaian, sikap, dan emosi seseorang terhadap sesuatu seperti produk, jasa, organisasi, orang lain, isu, peristiwa, dan topik [4]. Pada kasus masalah penelitian ini terdapat 3 kelas sentimen yaitu positif, negatif, dan netral. Dalam pengaplikasiannya data akan memiliki polaritas positif apabila data tersebut mengandung lebih banyak kata-kata positif lalu akan diklasifikasikan sebagai kelas positif, begitu juga dengan yang polaritas negatif. Apabila data memiliki kata-kata positif yang sama dengan kata-kata negatif maka akan dikategorikan sebagai netral.

2.2. Natural Language Processing

Natural Language Processing (NLP) adalah upaya untuk mengestraksi lebih lengkap representasi makna dari teks bebas. ini dapat dikatakan secara kasar sebagai mencari tahu siapa melakukan apa kepada siapa, kapan, di mana, bagaimana dan mengapa. NLP biasanya menggunakan konsep linguistik seperti *part-of-speech* (kata benda, kata kerja, kata sifat, dll) dan struktur tata bahasa. NLP harus berurusan dengan anaphora (kata benda yang sebelumnya melakukan pronomina atau frase yang merujuk kembali berhubungan dengan) dan ambiguitas. Untuk melakukan ini, NLP menggunakan berbagai representasi pengetahuan, seperti leksikon kata, arti kata, sifat tata Bahasa, dan seperangkat aturan tata bahasa dan dari sumber lain sinonim atau singkatan [5].

2.3. Pre-Processing

Pre-processing adalah proses awal pengolahan data yang bertujuan untuk menjadikan data siap untuk diolah dan dianalisis [6]. Ada 3 tahap yang harus dilakukan untuk *pre-processing*. Tahapan tersebut diantaranya sebagai berikut :

A. *Basic Operation* dan *Cleaning* : pada tahap ini dilakukan penghapusan elemen yang tidak penting atau menggangu untuk fase analisis selanjutnya. Untuk memperoleh informasi yang signifikan suatu *tweet* tidak berisi URL, *hashtag* (seperti #happy), *mention* (seperti @realDonaldTrump), *emoticon*, dan simbol. Setelah itu dapat dilakukan penghapusan tanda baca.
B. *lemmatization*: tahap ini adalah pengubahan kata menjadi kata dasarnya seperti 'using' menjadi 'use'
C. *Stopwords* : tahap ini dilakukan pemrosesan agar kata-kata yang terlalu sering digunakan dihapus sehingga dapat memengaruhi hasil.

2.4. Naïve Bayes

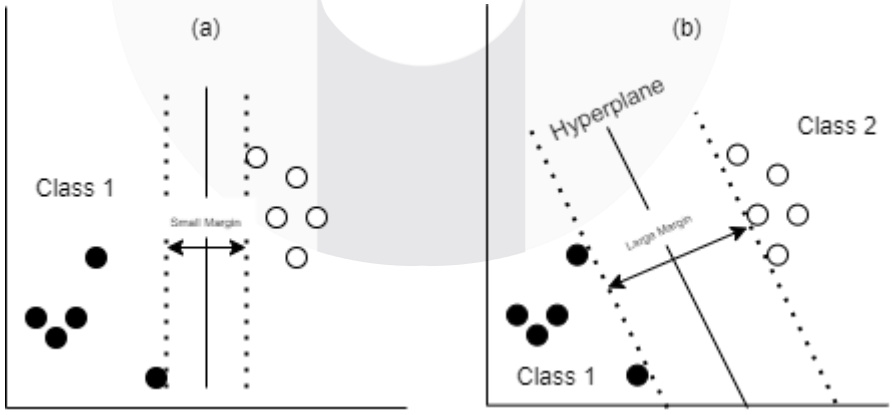
Dalam teori probabilitas, teorema Bayes menghubungkan probabilitas kondisional dan marginal dari dua kejadian acak. Ini sering digunakan untuk menghitung probabilitas posterier yang diberikan pengamatan [7]. Klasifikasi Naïve Bayes adalah klasifikasi yang didasari oleh teorema Bayes. Klasifikasi Naïve Bayes mengkombinasikan efisiensi (kinerja waktu yang optimal) dengan akurasi yang wajar [8].

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Keterangan :
A : hipotesis data merupakan suatu kelas tertentu
B : data yang masih belum diketahui kelasnya
P(A|B) : probabilitas hipotesis A dengan syarat kondisi B
P(A) : probabilitas hipotesis A
P(B) : probabilitas B
P(B|A) : probabilitas B berdasarkan kondisi hipotesis A. [9]

2.5. Support Vector Machine

Support vector machine (SVM) adalah algoritma untuk klasifikasi, regresi, dan density information [10]. Karakteristik yang luar biasa dari SVM adalah kemampuannya untuk belajar tidak bergantung pada dimensionalitas fitur ruang. SVM mengukur kompleksitas Hipotesis berdasarkan margin yang memisahkan bidang dan bukan jumlah fitur lalu dicari perpisahan margin yang paling optimum untuk memisahkan antara kelas yang satu dengan yang lainnya. SVM memiliki kelebihan yaitu tidak bergantung pada jumlah fitur sehingga dia dapat mengatasi data yang memilki banyak fitur [11].



Gambar 1 Ilustrasi Hyperplane pada SVM [10]

Pada gambar 1(a) memilki fungsi atau margin yang memisahkan *class* 1 dan *class* 2, tetapi tidak optimum dalam memisahkan. Sementara pada gambar 1(b) margin yang memisahkan antara *class* 1 dan *class* 2 lebih baik dalam memisahkan.

Pada klasifikasi teks, SVM melakukan klasifikasi kata menjadi kategori berdasarkan *feature set* [12]. Data yang dimasukkan akan ditransformasikan mengunakan fungsi kernel secara matematis untuk membuat pemisahan data secara linear dari berbagai kategori. Terdapat beberapa jenis kernel pada SVM. Berikut jenis kernelnya [13].

Tabel 1. Kernel Support Vectore Machine

Jenis Kernel	Fungsi
Linear	$K\left(x_i, x_j\right) = sum\ (x_i\ , x_j)$
Radial Basis Function	$K\ (x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
Polynomial	$K\left(x_i, x_j\right) = (x_i, x_j + 1)^d$

2.6. Studi Terkait

Penelitian terkait analisis sentimen sudah banyak dilakukan. Untuk mendukung penelitian ini, peneliti menggunakan penelitian yang sudah ada untuk dijadikan sebagai acuan dalam penelitian ini. Berikut merupakan hasil dari penelitian sebelumnya yang dijadikan referensi atau acuan oleh penulis.

Pada tahun 2014, [14] melakukan penelitian tentang analisis sentimen data twitter. Penelitian ini dilakukan untuk mengetahui sentimen dari tweet pada twitter. Terdapat 2 sentimen yang digunakan yaitu positif dan negatif. Metode yang digunakan pada penelitian tersebut ada 4 yaitu Naïve Bayes, Maximum Entropy, Support Vector Machine (SVM), dan Semantic Analysis. Setelah dilakukan pengujian didapatkan akurasi sebesar 88.2% pada Naïve Bayes, 83.8% pada Maximum Entropy, 85.5% pada SVM, dan 89.9% untuk Semantic Analysis.

Kemudian pada tahun 2018, [15] melakukan penelitian tentang analisis sentimen artikel berita Menteri Koordinator Bidang Kelautan. Penelitian ini dilakukan untuk mengetahui klasifikasi konten dari suatu artikel berita apakah terdapat kesimpulan yang positif atau negatif terkait berita tersebut. Metode yang digunakan ada 2 yaitu SVM dan Naïve Bayes. Pada penelitian ini digunakan *Particle Swarm Optimization* sebagai feature extraction. Setelah dilakukan pengujian, didapatkan hasil akurasi sebersar 87.50% untuk SVM, 90.50% untuk SVM + PSO, 89.50% untuk Naïve Bayes, dan 92.00% untuk Naïve Bayes + PSO.

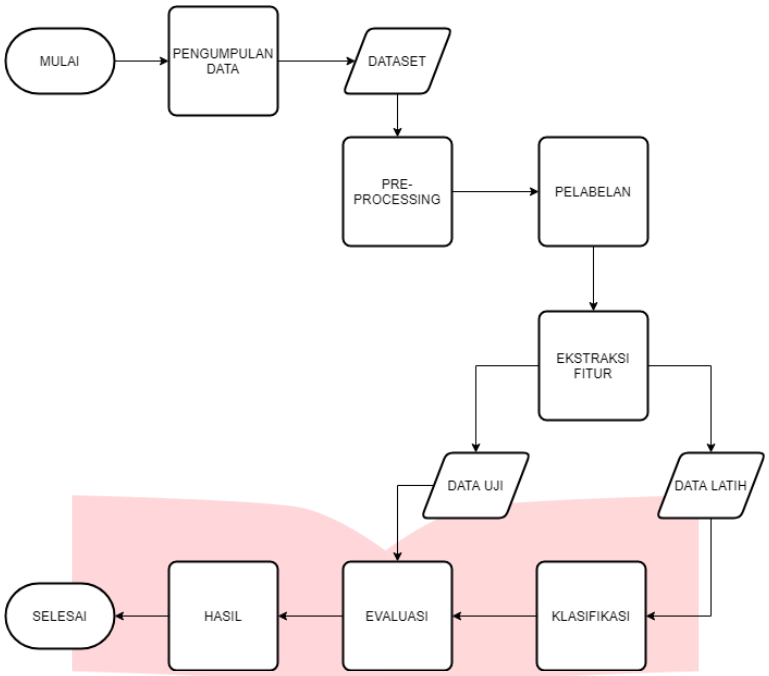
Selanjutnya, [16] melakukan penelitian tentang analisis sentimen penjualan barang pada *online store*. Penelitian ini dilakukan agar hasil dari analisis sentimen dapat digunakan untuk evaluasi penjualan. Metode yang digunakan ada 2 yaitu SVM dan Naïve Bayes. Pada penelitian ini digunakan TF-IDF dengan *max features* yang berbeda sebagai *feature extraction*. Setelah dilakukan pengujian, didapatkan hasil akurasi rata-rata paling besar 90.65% untuk Naïve Bayes dengan 100% *max feature* dan 93.65% untuk SVM dengan 25% *max feature*.

Penelitian selanjutnya, [17] melakukan penelitian tentang analisis sentimen e-sport untuk kurikulum pendidikan. penelitian ini dilakukan untuk mengukur pendapat atau memisahkan antara sentimen positif dan sentimen negatif terhadap pendidikan e-sport. Metode yang digunakan ada 2 yaitu SVM dan Naïve Bayes. Pada penelitian ini digunakan *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk evaluasi. Setelah dilakukan pengujian, didapatkan hasil akurasi 50.74% untuk Naïve Bayes, 70.32% untuk Naïve Bayes + SMOTE, 70.50% untuk SVM, dan 66.92% untuk SVM + SMOTE.

Yang terakhir, [18] melakukan analisis sentimen terhadap *hatespeech* pada data twitter. Penelitian ini dilakukan dengan menganalisis tagar #HateSpeech untuk mengetahui apakah semua tweet yang diberikan tagar tersebut memiliki sentimen negatif. Setelah dilakukan pengujian, didapatkan hasil akurasi 66.6% untuk SVM dan 63.7% untuk Naïve Bayes.

3. Sistem yang Dibangun

3.1. Deksripsi Umum Sistem



Gambar 2 Deskripsi Umum Sistem

3.2. Dataset

Dataset yang digunakan pada penelitian ini adalah data tweet pada twitter dari user mengenai pemilihan Presiden Amerika 2020. Data didapatkan melalui website *Kaggle.com* dalam bentuk *comma separated value* (CSV) sebanyak 1.72 juta data. Data tersebut didapatkan melalui scraping menggunakan hashtag #joebiden dan #donaldtrump.

Tabel 2. Contoh Sampel Data

No	tweet
1	Go watch the entire story. What a nightmare it must have been for all those passengers. The #CDC really messed up here, thanks to #Trump. It's not China's fault; it's the lack of leadership which started the spread of #COVID19 in the US. #TrumpKnewAndDidNothing https://t.co/FrFYOPyDMJ
2	@atrupar Accidentally? They are probably broke and these people were ok with payment after service unlike the hotel!! 😶#Trump #TrumpCampaign

3.3. Pre-Processing

Pre-processing merupakan tahap mengubah data mentah menjadi data yang dapat dipahami oleh komputer sehingga dapat diklasifikasikan. Tahap ini terdiri dari cleaning, negation, pengubahan kata salah eja atau kata gaul, stemming, dan penghapusan stopword.

Tabel 3. Perbandingan Data Pre-Processing

No	tweet	
1	Go watch the entire story. What a nightmare it must have been for all those passengers. The #CDC really messed up here, thanks to #Trump. It's not China's fault; it's the lack of leadership which started the spread of #COVID19 in the US. #TrumpKnewAndDidNothing https://t.co/FrFYOPyDMJ	go watch entire story nightmare must passenger cdc really mess thanks trump china fault lack leadership start spread covid19 us trumpknewanddidnothing
2	@atrupar Accidentally? They are probably broke and these people were ok with payment after service unlike the hotel!! 😶#Trump #TrumpCampaign	accidentally probably broke people ok payment service unlike hotel trump trumpcampaign

3.4. Labeling Data

Proses pelabelan data dilakukan agar algoritma Naïve Bayes dan Support Vector Machine dapat mempelajari dataset. Kelas yang dibentuk terdiri dari 3 yaitu positif, negatif, dan netral. Metode pelabelan dilakukan dengan menggunakan perbandingan kata baik dan buruk untuk menentukan kelasnya. Apabila kata baik lebih banyak daripada kata buruk maka akan dilabeli kelas positif dan sebaliknya akan dilabeli kelas negatif. Apabila kata positif sama dengan kata negatif maka akan dilabeli dengan kelas netral.

3.5. Data Split

Data split dilakukan untuk memisahkan data menjadi train-set dan test-set. Train-set akan digunakan untuk membangun model klasifikasi, sedangkan test-set akan digunakan untuk mengukur performa model yang dibangun.

3.6. Ekstraksi Fitur

Ekstraksi fitur dilakukan untuk mengubah data yang sudah di pre-processing menjadi data yang memiliki fitur-fitur. Pada penelitian ini, peneliti menggunakan metode TF-IDF untuk ekstraksi fitur. TF-IDF Adalah gabungan dari term frequency dan inverse document frequency (IDF). TF adalah pengulangan terminologi dalam dokumen. IDF adalah perhitungan terminologi yang sering didistribusikan dalam dokumen [17]. Untuk menghitung IDF dapat digunakan rumus sebagai berikut.

$$IDF_j = \log(D/df_i)$$

Dimana D merupakan jumlah semua dokumen sedangkan df_i adalah jumlah dokumen yang mengandung TF. Rumus TF-IDF sebagai berikut.

$$TFIDF = TF \times IDF$$

3.7. Evaluasi Performa

untuk mengevaluasi performa suatu model, digunakan *confusion matrix*. *Confusion matrix* berisi informasi asli dan yang diprediksi [20]. Dari matriks tersebut dapat diketahui akurasi, presisi, dan recall [16]. Contoh confusion matrix dapat dilihat pada tabel berikut.

Tabel 4. Confusion Matrix

Kelas		Predicted Class		
		Positif	Negatif	Neutral
Actual Class	Positif	**True Positive (TP)**	**False Negative (FN)**	**False Neutral (FNet)**
	Negatif	**False Positive (FP)**	**True Negative (TN)**	**False Neutral (FNet)**
	Neutral	**False Positive (FP)**	**False Negative (FNet)**	**True Neutral (TNet)**

Perhitungan akurasi dilakukan untuk mencari tahu jumlah data yang benar yang diklasifikasikan oleh sistem [14]. Untuk menghitung akurasi dari sebuah model dapat dilakukan dengan persamaan :

$$Accuracy = \frac{TP + TN + TNet}{TP + TN + FP + FN + FNet}$$

Perhitungan *precision* untuk mengetahui perbandingan jumlah kelas positif yang diprediksi benar dibandingkan dengan jumlah semua kelas positif [20]. presisi dari sebuah model dapat dihitung dengan persamaan :

$$Precision = \frac{TP}{TP + FP}$$

Perhitungan *recall* dilakukan untuk mengetahui perbandingan jumlah kelas positif yang diprediksi benar dibandingkan dengan kelas yang benar positif [20]. menghitung recall dari sebuah model dapat dilakukan dengan persamaan :

$$Recall = \frac{TP}{TP + FN + Fnet}$$

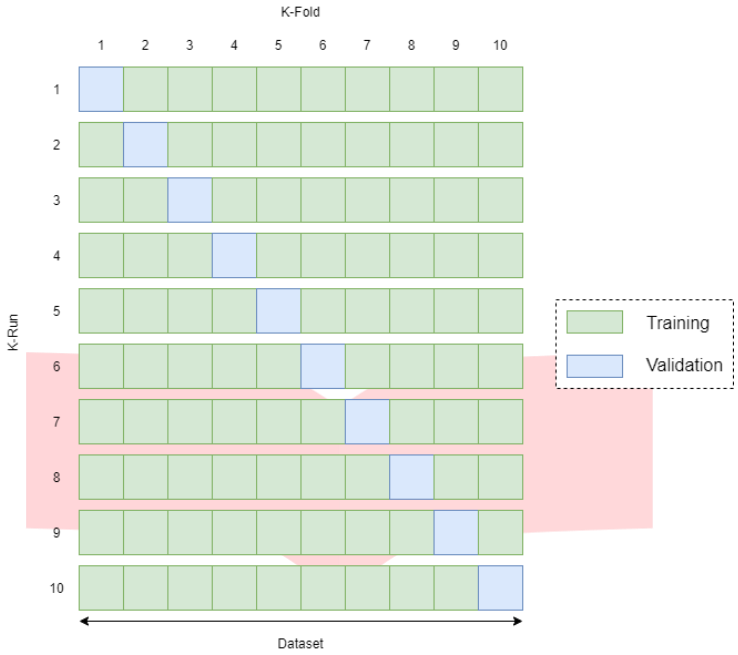
Perhitungan F-*Measure* diperoleh dari *precision* dan *recall*. Menghitung F-*Measure* dari sebuah model dapat dilakukan dengan persamaan :

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.8. Validasi

Cross-validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja suatu model dan untuk mencegah terjadinya overfitting pada model. Pada CV data dipisahkan menjadi dua subset yaitu data training dan data testing. Pada penelitian ini digunakan CV K-fold untuk validasi. Pada penelitian ini

digunakan 10 fold. Dimana data dibagi menjadi 10 fold berukuran sama, sehingga terdapat sebanyak 10 kali pengujian. Untuk masing-masing pengujian, CV akan menggunakan 9 fold (80%) untuk *training* dan 1 fold (20%) untuk *testing*. Berikut ilustrasinya.



Gambar 3 K-fold

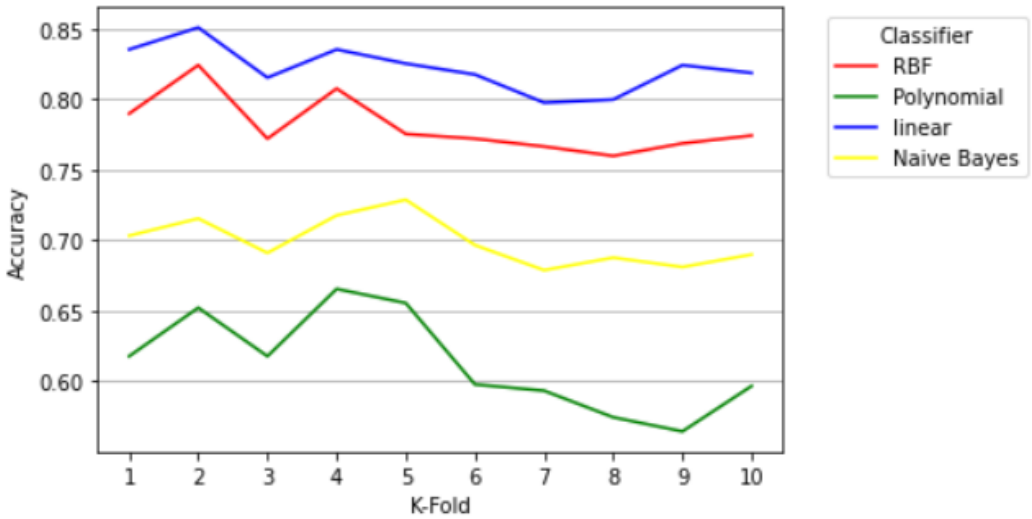
4. Evaluasi

4.1 Hasil Pengujian

Pada penelitian ini dilakukan beberapa skenario pengujian. Skenario pertama, Algoritma SVM dengan kernel *rbf*, *linear*, dan *polynomial* dibandingkan dengan Naïve Bayes. Skenario kedua, split *data train* dan *data test* dengan rasio 90:10, 80:20, dan 70:30.

Setelah dilakukan pengujian, didapatkan hasil sebagai berikut.

4.1.1 Skenario 1



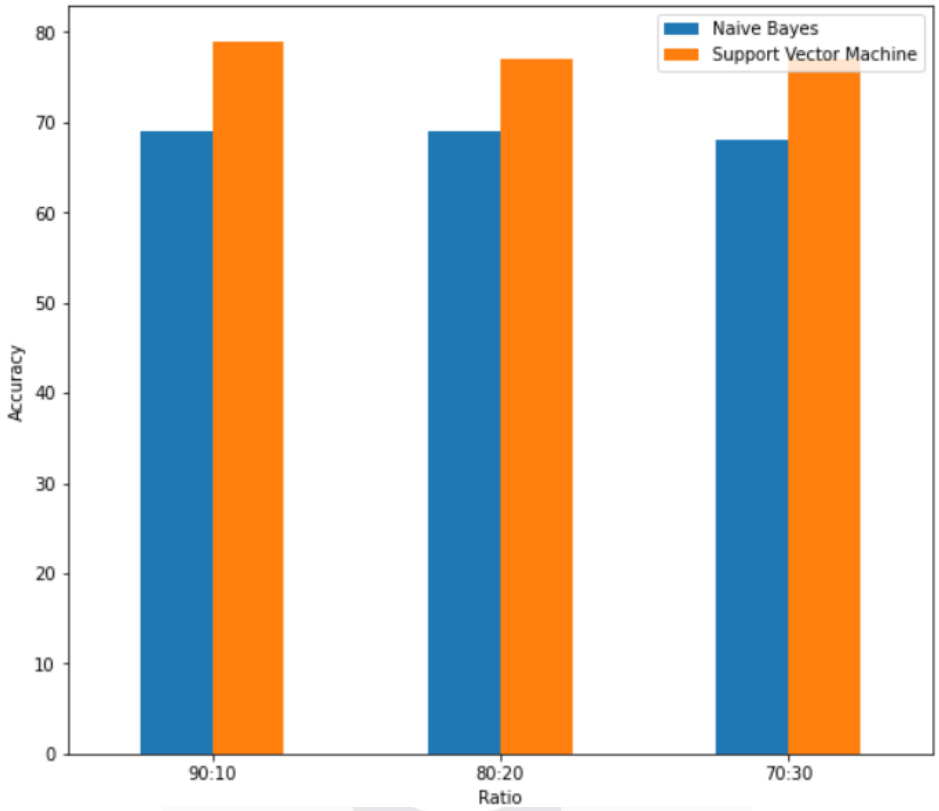
Gambar 4 Hasil Pengujian Skenario 1

Tabel 6. Hasil Evaluasi Skenario 1

Klasifier	Akurasi (Rata-rata)
Support Vector Machine dengan kernel RBF	78.1%
Support Vector Machine dengan kernel polynomial	61.3%
Support VectoreMachine dengan kernel linear	82.2 %
Naïve Bayes	69.9%

Pada skenario pertama, didapatkan algoritma SVM dengan kernel linear mendapatkan hasil rata-rata paling bagus daripada algoritma lainnya. Namun bukan berarti SVM lebih bagus daripada Naïve Bayes dalam mengklasifikasi. Dapat dilihat Naïve Bayes dapat mengungguli SVM yang menggunakan kernel polynomial.

4.1.2 Skenario 2



Gambar 5 Hasil Evaluasi Skenario 2

Berdasarkan hasil evaluasi pada skenario 2 dapat dilihat bahwa performa dari SVM lebih baik dari Naïve Bayes dari segi akurasi dengan penggunaan parameter standar.

4.2 Analisis Hasil Pengujian

Berdasarkan penelitian yang dilakukan dalam hal akurasi SVM lebih baik dari Naïve Bayes. Namun, ada beberapa hal Naïve Bayes lebih baik daripada SVM. Klasifikasi SVM menghasilkan akurasi terbaik dengan rata-rata sebesar 82.2% sedangkan algoritma Naïve Bayes menghasilkan akurasi sebesar 69%.

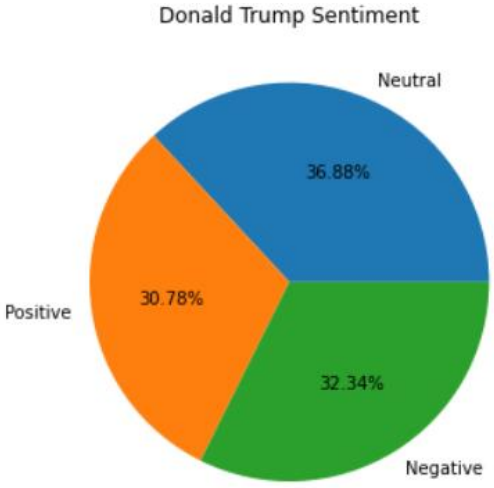
Berdasarkan penelitian yang dilakukan, dalam hal akurasi SVM lebih baik dari Naïve Bayes. Didapatkan bahwa perbedaan kernel dapat membuat perbedaan hasil akurasi pada skenario 1. Pada skenario 2 didapatkan bahwa perubahan rasio *data train* dan *data test* dapat meningkatkan hasil akurasi.

Tabel 9. Hasil Precision, Recall, dan F1-Score SVM dengan Kernel Linear

Class	Precision	Recall	F1-Score
Positive	77.00%	88.00%	82.00%
Negative	79.00%	82.00%	80.00%
Neutral	91.00%	78.00%	84.00%
Mean	82.00%	83.00%	82.00%

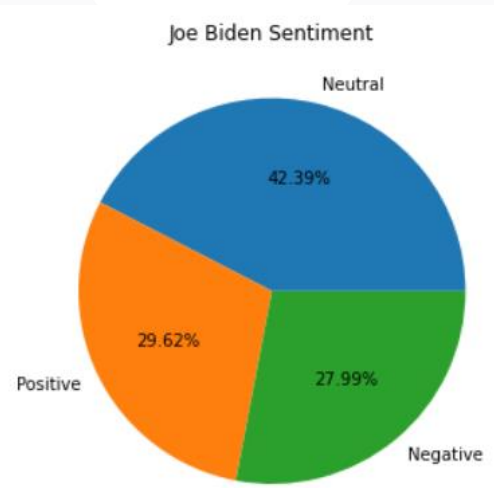
Berdasarkan tabel diatas, dapat diketahui bahwa model dapat memprediksi kelas netral lebih baik daripada kelas lainnya dengan nilai presisi sebesar 91%. dengan rata-rata Recall sebesar 83% dapat dikatakan bahwa 83% data yang prediksi adalah data yang sebenarnya.

4.3 Hasil Sentimen



Gambar 6 Sentimen Donald Trump

Berdasarkan gambar 6, model SVM dengan akurasi terbaik yaitu 82% berhasil memprediksi 10000 data *tweet* dengan *hashtag #donaldtrump* dengan komposisi 3688 label netral atau 36.88% orang memiliki pandangan netral terhadap Trump, 3078 label positif atau 30.78% orang memiliki pandangan positif terhadap Trump, 3234 label negatif atau 32.34% orang memiliki pandangan negatif terhadap Trump



Gambar 7 sentimen Joe Biden

Berdasarkan gambar 6, model SVM dengan akurasi terbaik yaitu 82% berhasil memprediksi 10000 data *tweet* dengan *hashtag #joebiden* dengan komposisi 4239 label netral atau 42.39% orang memiliki pandangan netral terhadap Biden, 2962 label positif atau 29.62% orang memiliki pandangan positif terhadap Biden, 2799 label negatif atau 27.99.34% orang memiliki pandangan negatif terhadap Biden

5. Kesimpulan

Pada penelitian ini peneliti mencoba membandingkan algoritma SVM dan Naïve Bayes dalam kasus Pemilihan Presiden Amerika 2020 dan untuk menunjukkan algoritma yang tepat untuk klasifikasi sentimen ini. Hasilnya didapatkan bahwa SVM lebih baik daripada Naïve Bayes dengan menggunakan kernel linear dan perbedaan rata-rata 12%, dimana akurasi dari SVM sebesar 82% dan Naïve Bayes sebesar 69% . dengan begitu, proses analisis sentiment pada Pemilihan Presiden Amerika 2020 lebih baik dalam menggunakan Support Vectore Machine.

5.1 Saran

Saran untuk pengembangan lebih lanjut dari penelitian ini sebagai berikut :

1. Mencoba menggunakan metode klasifikasi lainnya yang belum digunakan pada penelitian ini
2. Menggunakan ekstraksi fitur lainnya pada proses klasifikasi sentimen.

Referensi

[1] Klein, E. (2020, October 19). *Vox*. Retrieved from It's the most important election in our lifetime, and it always will be: https://www.vox.com/21504729/most-important-election-bush-gore-kerry-trump

[2] Longwell, Sarah (2020, October 19). *The Atlantic*. Retrieved from Why People Who Hate Trump Stick With Him : https://www.theatlantic.com/ideas/archive/2020/10/why-people-who-hate-trump-stick-him/616758/

[3] Palaiologos, Yannis (2020, November 3). *Ekathimerini*. Retrieved from The importance of the 2020 US presidential elections : https://www.ekathimerini.com/opinion/258752/the-importance-of-the-2020-us-presidential-elections/

[4] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[5] Kao, A., & Poteet, S. R. (Eds.). (2007). Natural language processing and text mining. Springer Science & Business Media.

[6] Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., & Manicardi, S. (2016, September). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb.

[7] Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December). Naive bayes classification of uncertain data. In 2009 Ninth IEEE International Conference on Data Mining (pp. 944-949). IEEE.

[8] Gamallo, P., Garcia, M., & Fernández-Lanza, S. (2013, September). TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. In Workshop on Sentiment Analysis at SEPLN (TASS2013) (pp. 126-132).

[9] Kurniawan, I., & Susanto, A. (2019). Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019. Jurnal Eksplora Informatika, 9(1), 1-10.

[10] Burbidge, R., & Buxton, B. (2001). An introduction to support vector machines for data mining. Keynote papers, young OR12, 3-15.

[11] Patil, G., Galande, V., Kekan, V., & Dange, K. (2014). Sentiment analysis using support vector machine. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), 2607-2612.

[12] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[13] Nadkarni, P., Ohno-Machado, L., & Chapman, W. W. 2011. Natural Language Processing: an Introduction. Journal of the American Medical Informatics Association, 18(5), 544-551

[14] Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh International Conference on Contemporary Computing (IC3) (pp. 437-442). IEEE.

[15] Wardhani, N. K., Rezkiani, S. K., Setiawan, H. E. N. D. R. A., Gata, G. R. A. C. E., Tohari, S. I. S. W. A. N. T. O., Gata, W. I. N. D. U., & Wahyudi, M. O. C. H. A. M. A. D. (2018). Sentiment analysis article news coordinator minister of maritime affairs using algorithm naive bayes and support vector machine with particle swarm optimization. J. Theor. Appl. Inf. Technol, 96(24), 8365-8378.

[16] Lutfi, A. A., Permanasari, A. E., & Fauziati, S. (2018). Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine. Journal of Information Systems Engineering and Business Intelligence, 4(1), 57-64.

[17] Ardianto, R., Rivanie, T., Alkhalifi, Y., Nugraha, F. S., & Gata, W. (2020). SENTIMENT ANALYSIS ON E-SPORTS FOR EDUCATION CURRICULUM USING NAIVE BAYES AND SUPPORT VECTOR MACHINE. Jurnal Ilmu Komputer dan Informasi, 13(2), 109-122.

[18] Buntoro, G. A. (2016). Analisis Sentimen Hatespeech pada Twitter dengan Metode Naive Bayes Classifier dan Support Vector Machine. Jurnal Dinamika Informatika, 5(2).

[19] Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In 2014 international conference on computer, communications, and control technology (I4CT) (pp. 333-337). IEEE.

[20] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.