ANALISIS SENTIMEN PROGRAM ACARA DI SCTV PADA TWITTER MENGGUNAKAN METODE NAIVE BAYES DAN SUPPORT VECTOR MACHINE

Dery Anjas Ramadhan[1], Erwin Budi Setiawan S.Si., M.T[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]deryar@student.telkomuniversity.ac.id , [2]erwinbudisetiawan@telkomuniversity.ac.id ,

Abstrak

Kepuasan penonton merupakan suatu faktor yang menjadi bahan pertimbangan dalam menentukan keberhasilan program acara televisi. Dengan kepuasan penonton perusahaan tahu harus membuat jadwal tayang yang baik dan menentukan berapa durasi acara maupun berapa episode yang harus dibuat. Selain itu perusahaan dapat meminimalisir kerugian karena jumlah penonton yang tidak sesuai harapan, maka dari itu dapat dilakukan suatu *Analisis Sentimen* pada Tweet Twitter SCTV yang berupaya untuk mengetahui sentimen dari setiap Tweet. Metode klasifikasi yang digunakan adalah *Naive Bayes* dan *Support Vector Machine*. Setelah dilakukan penelitian akurasi terbaik dari Metode Naive Bayes dan Metode Support Vector Machine adalah Support Vector Machine dengan Seluruh Program Acara didapatkan hasil akurasi 88,57%.

Kata kunci : *Analisis Sentimen, Naive Bayes, Support Vector Machine*

Abstract

Audience satisfaction is a factor that is taken into consideration in determining the success of a television program. With satisfaction the audience knows the company must make a good showtimes and determine the duration of the event and how many episodes to make. In addition, the company can minimize losses because the number of viewers is not as expected, therefore a Sentiment Analysis can be done on SCTV's Twitter Tweet which seeks to find out the sentiments of each Tweet. The classification method used is Naive Bayes and Support Vector Machine. After the best accuracy research from the Naive Bayes Method and the Support Vector Machine Method is a Support Vector Machine with All Program Programs, the results are 88.57% accuracy.

Keywords: *Sentiment Analysis, Naive Bayes, Support Vector Machine*

1. Pendahuluan

Perkembangan televisi yang pesat ini telah membuat banyak televisi swasta membuat program - program unggulan upaya untuk memikat masyarakat untuk menonton program televisi mereka. Apabila program acara televisi disukai penonton maka tingkat kepopuleran chanel televisi yang dimiliki perusahaan akan meningkat dan bisa mempertahanan program acara televisi yang diunggukan perusahaan tersebut.

Pada penelitian ini akan menganalisis sentimen terhadap kepuasan program acara televisi di salah satu saluran televisi swasta yaitu SCTV (Surya Citra Televisi). SCTV memiliki kategori program acara yaitu FTV, Sinetron, Berita, dan Entertaiment. Dengan penelitian diharapkan dapat membantu menganalisis kepuasan penonton televisi agar dapat mengetahui sampai dimana tingkat sentimen positif kepuasan pada penonton, karena dengan mengetahui kepuasan program acara dapat dipertahankan dan diperbaiki lagi agar lebih menarik hati dari para penonton televisi.

Untuk mendapatkan data berita dan Tweet dari Twitter dilakukan crawling. Crawling data merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database. Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server Twitter berupa user dan Tweet beserta atribut - atributnya.[6]

Menggunakan metode Naive Bayes dan Support Vector Machine dianggap tepat karena metode ini mudah diterapkan untuk mencari *trending event*. Pengklasifikasian yang saling bebas membuat peluang tiap parameter menjadi lebih mudah. *Naïve Bayes Classifier* merupakan sebuah metode klasifikasi yang berakar pada teorema Bayes. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi/kejadian. Sebelum menjelaskan *Naïve Bayes Classifier* ini, akan dijelaskan terlebih dahulu Teorema Bayes yang menjadi dasar dari metode tersebut.[1]

Untuk metode SVM pertama kali diperkenalkan di tahun 1992 pada *Annual Workshop on Computational Learning Theory*. SVM dikembangkan oleh Boser bersama 2 orang temannya yakni Guyon dan Vapnik (Saifinnuha, A. Z., 2015). Prinsip kerja SVM pada awalnya sebagai metode untuk klasifikasi linier (*linear classifier*), dan dikembangkan untuk dapat menyelesaikan permasalahan klasifikasi non-linear, dengan memanfaatkan fungsi kernel untuk data atau ruang kerja dengan dimensi tinggi (Nugroho, 2007).[2]

Setelah didapatkan informasi tentang kepuasan penonton terhadap program acara televisi maka bisa digunakan untuk memprediksi seberapa bagusnya program acara televisi untuk memikat ketertarikan penonton pada saluran televisi tersebut. Sehingga bisa bersiap-siap berinovasi menghadapi perubahan dimasa mendatang.

2. Tinjauan pustaka

2.1 Analisis Sentimen

Analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau Fitur/tingkat aspek dan menentukan apakah pendapat yang di nyatakan dalam dokumen, kalimat atau Fitur entitas/aspek bersifat positif, negatif atau netral. Lebih lanjut sentiment analysis dapat menyatakan emosional sedih, gembira, atau marah Liu, B (2012).[4]

2.2 Crawling

Crawling data merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database. Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server Twitter berupa *User* dan *Tweet* beserta atribut - atributnya. Aplikasi crawling ini dibuat dengan memodifikasi *Application Programming Integration* (API) Twitter dengan menggunakan Bahasa pemograman R.

Crawling data di Twitter dapat menggunakan dua sistem pencarian, *by user* dan *by keyword*. Pencarian menggunakan *by keyword* yaitu pencarian menggunakan penggalan kata maupun hashtag dengan total Tweet yang diunduh dalam sekali proses maksimum 200 Tweet. Sedangkan pencarian dengan *by user* yaitu pencarian berdasarkan nama akun *User Twitter* dengan total Tweet yang diunduh dalam sekali proses maksimum 2000 Tweet. Ekstrasi Fitur yang didapat dari index twitter untuk data user berupa total Tweet, total follower, total following, total likes, website, source, bio profile, id, akun, nama dan lokasi. Sedangkan ekstraksi Fitur yang didapat dari index Twitter untuk data Tweet berupa url, mention, retweet, hashtag, jumlah likes dan jumlah retweet.[1]

2.3 Naive Bayes

Naive Bayes adalah tehnik yang diterapkan untuk menentuan kelas dari tiap masalah, yang sudah dibagi berdasarkan tiap-tiap masalah. perhitungan numerik berdasarkan pada pendekatan grup. Naive bayes memiliki beberapa manfaat seperti sederhana, cepat, memiliki tingkat akurasi yang tinggi (Jurafsky & Martin, 2015). Leung, menjelaskan rumus Bayes adalah, dimana Nc : nomor dokumen pada kelas α dan N: jumlah nomor pada dokumen.

$$\gamma(\alpha) = \frac{N_c}{N} \qquad (1)$$

Model angka dari Naive Bayes merekam informasi tentang frekuensi kata pada dokumen. *Maximum Likelihood Estimate* (MLE) adalah frekuensi relatif dan sesuai dengan kemungkinan nilai masing-masing parameter yang diberikan data pelatihan. Persamaan (1) menjelaskan probabilitas sebelum perkiraan. Dalam model Multinomial, mengasumsikan nilai atribut yang bebas satu sama lain yang diberikan untuk kelas tertentu $\gamma = (\alpha \mid \beta) = \gamma(\omega 1 \ldots \omega nd|\alpha)$. Dalam model multinomial, dokumen memerintahkan urutan peristiwa kata, ditarik dari kosakata V. Asumsikan bahwa panjang dokumen independent dari kelas. dengan demikian, masing-masing iβ dokumen diambil dari pembagian multinomial kata - kata dengan banyak percobaan independen sebagai panjang iβ. ini menghasilkan hal yang umum yaitu seperti kantong yang berisi banyak kata yang merepresentasikan dokumen dokumen. (Dhande dan Patnaik, 2014). Jumlah(ω,α) = Jumlah kejadian dari ω dalam dokumen training set dari α kelas, jumlah(α) = jumlah kata dalam kelas tersebut, |V| = jumlah yang di terima sebagai kosakata yang di gunakan.

$$\gamma(\omega|\alpha) = \frac{jumlah(\omega, a) + 1}{jumlah(\alpha) + |V|} \qquad (2)$$

Masalah dengan estimasi MLE adalah bahwa hal itu adalah nol untuk kombinasi term class yang tidak terjadi dalam data pelatihan. Untuk menghilangkan masalah probabilitas nol, menggunakan *add - one* atau *Laplace Smoothing*. Tambahkan satu *smoothing* dapat diartikan sebagai uniform prior (setiap istilah terjadi sekali untuk setiap kelas) yang kemudian diperbarui sebagai bukti dari data pelatihan yang masuk. Kemudian, probabilitas dokumen yang diberikan kelasnya secara sederhana distribusi multinomial direpresentasikan pada persamaan (2). Akhirnya

mengklasifikasikan dokumen baru menggunakan probabilitas posteriori. α_{NB} adalah probabilitas posterior, α_j adalah salah satu kelas dari α kelas dan βi adalah dokumen.

$$\alpha_{NB} = \arg\max \quad \alpha_j \epsilon \alpha \prod_i \gamma(\beta_i|\alpha_i) \qquad (3)$$

Uji akurasi dilakukan dengan tujuan mengetahui tingkat ketepatan hasil prediksi klasifikasi suatu kelas terhadap kelas yang sebenarnya. Akurasi diukur menggunakan persamaan (4).[1]

$$\text{Akurasi}(\textit{Kelas}) = \frac{|\text{anggota } (\textit{kelas}) \text{ sebenarnya} \cap \text{anggota (kelas) hasil prediksi}|}{|\text{seluruh anggota hasil prediksi}|} \qquad (4)$$

2.4 Support Vector Machine

Untuk memproses data teks maka digunakan kernel linear. Fungsi kernel linear mengubah data teks menjadi matriks kernel. Setiap elemen matriks kernel $K(x_i , x_j)$ digunakan untuk menggantikan dot-product xi.x j dalam persamaan dualitas *Lagrange*:

$$\begin{aligned} \max L_D(\alpha) &= \frac{1}{2}\left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)\left(\sum_{i=1}^{n} \alpha_i y_i x_i\right) - \sum_{i=1}^{n} \alpha_i y_i \left(\left(\sum_{i=1}^{n} \alpha_i y_i x_i\right) x_i + b\right) + \sum_{i=1}^{n} \alpha_i \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j (x_i . x_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j (x_i . x_j) - b\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j (x_i . x_j) \end{aligned} \qquad (5)$$

Untuk melakukan pengukuran kinerja klasifikasi digunakan matriks konfusi. Berdasarkan isi matriks konfusi, dapat diketahui nilai presisi, recall dan akurasi dari hasil klasifikasi yang diperoleh. Setiap sel f_{ij} dalam matriks menyatakan jumlah record (data) dari kelas i yang hasil prediksinya masuk ke kelas j, f_{11} data benar, f_{01} data positif f_{10} data negatif. Untuk menghitung akurasi, presisi dan recall digunakan formula:[2]

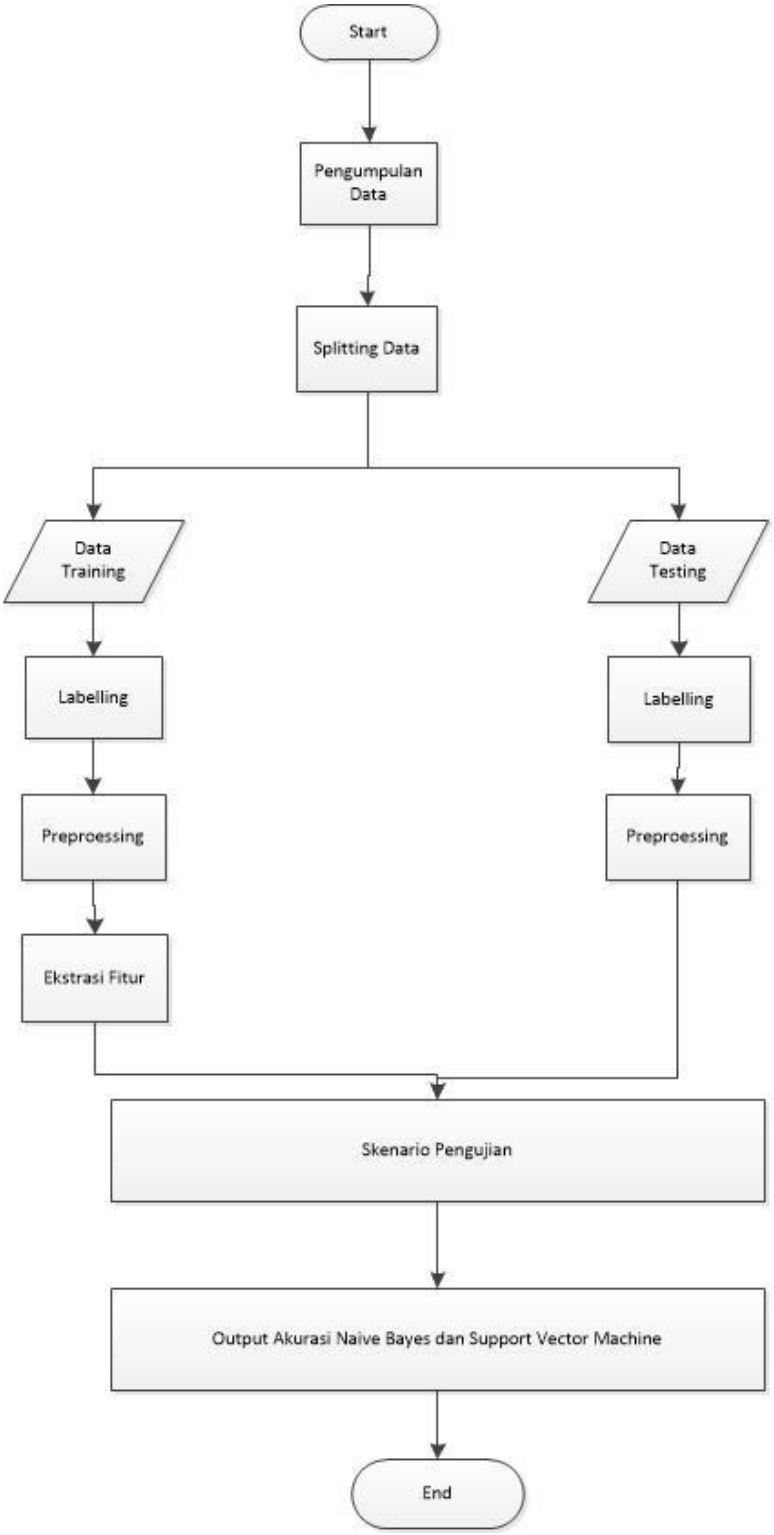
$$\begin{aligned} \textit{Presisi} &= \frac{f_{11}}{f_{11} + f_{01}}, \\ \textit{Recall} &= \frac{f_{11}}{f_{11} + f_{10}}, \\ \text{Akurasi} &= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}. \end{aligned} \qquad (6)$$

2.5 N-gram

N-Gram adalah proses pemecahan kalimat menjadi kata - kata berdasarkan n kata yang dipilih. Berdasarkan [9] *N - Gram* dibagi menjadi tiga jenis yaitu *Unigram* untuk pemecahan kalimat menjadi token yang berisi satu kata, *Bigram* untuk pemecahan kalimat menjadi token yang berisi dua kata dan *Trigram* untuk pemecahan kalimat menjadi token yang berisi tiga kata.

3. Sistem yang Dibangun

Secara umum sistem yang akan dibuat dalam tugas akhir ini adalah sebagai berikut.



3.1 Pengumpulan Data

Pengambilan data dari Twitter dilakukan menggunakan aplikasi *Automatic Crawling Twitter* melalui API Twitter dengan mengambil seluruh cuitan yang mengandung kata "ftv, sinetron, entertaimentsctv dan liputan6sctv" pada Twitter. Data diambil pada awal Mei sampai akhir Juni 2019 sebanyak 4.198 Tweet.

Tabel 1. Contoh cuitan beserta labelnya

No	Cuitan	Label
1	ambulans berisi batu uang diamankan kerusuhan	Positif

2	ungkapan syukur marcella zalianty injakkan kaki tanah suci	Positif
3	audimarissa malam nonton diam diam suka pkl penasaran ulah naomi sctv	Netral
4	robby satria gitaris band geisha tertangkap narkoba ganas kecam publik figur tampil televisi sctv now yaa	Negatif
5	kemuning kerjaannya ngeledekin kak melati om ben deh	Netral
6	momen abah zaenal sekeluarga kaget kemiripan bu arini almarhumah bu dahlia	Netral

3.2 Preprocessing

Data yang diperoleh dari tahap sebelumnya dilakukan pemrosesan untuk mengubahnya menjadi data yang siap diolah. Tahap *preprocessing* ini memiliki beberapa tahapan di antaranya, yaitu.

a. *Case folding* : mengubah seluruh huruf menjadi huruf kecil
b. *Cleansing* : membersihkan dokumen dari karakter – karakter special yang tidak terbaca sistem
c. *Stemming* : menghilangkan imbuhan pada kata
d. *Stopwords* : menghilangkan kata – kata yang tidak penting seperti kata sambung dan kata ganti orang

3.3 Ekstrasi Fitur

Data training kemudian masuk ke dalam tahap ekstraksi Fitur, pada tahap ini akan digunakan Fitur *N - gram.* Dalam Fitur *N-gram* yang gunakan ada Unigram, Bigram, Trigram, Unigram+Bigram, Bigram+Trigram dan Unigram+Bigram+Trigram.

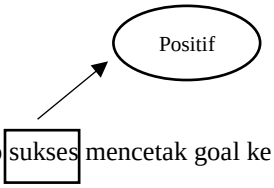
3.4 Skenario Pengujian

Proses skenario pengujian ini dilakukan dengan menggunakan Fitur *N - gram* 6 skenario, menggunakan komposisi 90%Training data dan 10%Testing data, manual labelling dan otomatis labelling.

3.5 Analisis Sentimen

Analisis sentimen dilakukan dengan 2 cara yaitu pertama dengan cara otomatis labelling mencocok kamus data yang telah disediakan berisi kata positif dan negatif dengan kalimat yang akan di analisis sentimen[10] kedua dengan cara manual labelling pada kalimat. Alur proses analisis sentimen untuk mendapatkan sentimen positif, negatif dan netral. Urutan proses pada sistem adalah, memasukkan Dokumen Tweet, Preprocessing, Analisis Sentimen, lalu keluar hasil sentimen positif, negatif dan netral. Berikut contoh otomatis labelling dan manual labelling.

a. Otomatis Labelling



Kata ***sukses*** ada pada kamus data positif, maka kalimat tersebut menjadi kalimat positif.

b. Manual Labelling

Kiper arsenal dapat menepis tendangan benzema (Positif)

4. Hasil dan Analisis

Pada bab ini skenario yang telah dibuat akan diuji menggunakan data latih yang didapatkan dari tahap splitting data.

4.1 Hasil Pengujian

Berikut hasil akurasi yang didapatkan dari hasil percobaan dan data yang digunakan sebagai berikut

Tabel 1.

No	Jenis Acara	Negatif		Positif		Netral	
		Otomatis	Manual	Otomatis	Manual	Otomatis	Manual
1	Berita	432	221	268	238	355	596
2	Entertaiment	411	58	268	251	355	746
3	Sinetron	432	194	268	271	355	590
4	FTV	411	47	272	220	372	788
5	Seluruh Program Acara	1685	520	1277	980	1236	2720

Tabel 2.

Jenis Program Acara	Skenario	Akurasi (%)			
		NB		SVM	
		Otomatis	Manual	Otomatis	Manual
Seluruh Program Acara	Unigram	66,09	41,70	88,57	55,37
	Bigram	65,40	41,53	79,52	53,44
	Trigram	64,08	44,21	77,14	52,96
	Unigram+Bigram	67,60	41,42	87,62	56,05
	Bigram+Trigram	67,13	37,40	78,33	52,44
	Unigram+Bigram+Trigram	68,12	42,63	88,10	56,57
Berita	Unigram	62,35	37,06	73,08	64,15
	Bigram	52,42	34,91	62,50	59,43
	Trigram	48,01	28,95	48,08	41,51
	Unigram+Bigram	63,49	40,46	77,88	63,21
	Bigram+Trigram	50,51	34,53	64,42	60,38
	Unigram+Bigram+Trigram	60,47	40,39	79,81	64,15
Entertaiment	Unigram	67,83	27.35	86,73	79.05
	Bigram	65,36	29.16	71,43	78.10
	Trigram	56,04	26.77	52,02	79.05
	Unigram+Bigram	67,01	31.65	88,78	76.19
	Bigram+Trigram	65,86	28.30	71,43	78.10
	Unigram+Bigram+Trigram	66,50	28.13	89,80	80.95
Sinetron	Unigram	64,79	50.90	69,74	71.43

	Bigram	59,75	44.80	64,47	69.05
	Trigram	52,44	46.99	57,89	67.86
	Unigram+Bigram	62,21	53.99	73,68	72.62
	Bigram+Trigram	62,11	50.24	59,21	67.86
	Unigram+Bigram+Trigram	58,98	55.56	73,68	65.48
FTV	Unigram	56,33	27,04	87,74	70,75
	Bigram	50,46	25,88	77,36	68,87
	Trigram	46,87	23,82	47,17	66,98
	Unigram+Bigram	55,70	23,83	86,79	67,92
	Bigram+Trigram	54,27	19,70	77,36	68,87
	Unigram+Bigram+Trigram	58,15	21,05	87,74	68,89

4.2 Analisis Hasil Pengujian

Berdasarkan hasil dari pengujian yang telah dilakukan di dapatkan hasil akurasi sebagai berikut

1. Otomatis Labelling
   a. Berita didapatkan hasil akurasi Naive Bayes 63,49% dengan menggunakan Unigram+Bigram dan Support Vector Machine 79,81% dengan menggunakan Unigram+Bigram+Trigram.
   b. Entertaiment didapatkan hasil akurasi Naive Bayes 67,83% dengan menggunakan Unigram dan Support Vector Machine 89,80% dengan menggunakan Unigram+Bigram+Trigram.
   c. Sinetron didapatkan hasil akurasi Naive Bayes 64,79% dengan menggunakan Unigram dan Support Vector Machine 73,68% dengan menggunakan Unigram+Bigram.
   d. FTV didapatkan hasil akurasi Naive Bayes 58,15% dengan menggunakan Unigram+Bigram+Trigram dan Support Vectro Machine 87,74% dengan menggunakan Unigram.
   e. Seluruh Program Acara didapatkan hasil akurasi Naive Bayes 68,12% dengan menggunakan Unigram+Bigram+Trigram dan Support Vector Machine 88,57 dengan menggunakan Unigram.
2. Manual Labelling
   a. Berita didapatkan hasil akurasi Naive Bayes 40,46% dengan menggunakan Unigram+Bigram dan Support Vector Machine 64,15% dengan menggunakan Unigram+Bigram+Trigram.
   b. Entertaiment didapatkan hasil akurasi Naive Bayes 31.65% dengan menggunakan Unigram+Bigram dan Support Vector Machine 80.95% dengan menggunakan Unigram+Bigram+Trigram.
   c. Sinetron didapatkan hasil akurasi Naive Bayes 55,56% dengan menggunakan Unigram+Bigram+Trigram dan Support Vector Machine 71.43% dengan menggunakan Unigram.
   d. FTV didapatkan hasil akurasi Naive Bayes 27,04% dengan menggunakan Unigram dan Support Vectro Machine 70,75% dengan menggunakan Unigram.
   e. Seluruh Program Acara Program Acara didapatkan hasil akurasi Naive Bayes 42,63% dengan menggunakan Unigram+Bigram+Trigram dan Support Vector Machine 56,57% dengan menggunakan Unigram+Bigram+Trigram.

5. Kesimpulan

Berdasarkan hasil penelitian untuk Analisis Sentimen Program Acara di SCTV pada Twitter menggunakan Metode Naive Bayes dan Support Vector Machine yang telah dilakukan mendapatkan Labelling terbaik didapatkan dengan cara Labelling Otomatis sedangkan akurasi terbaik dari Metode Naive Bayes dan Support Vector Machine adalah Metode Support Vector Machine dengan Seluruh Program Acara didapatkan hasil akurasi 88,57% berikut perkategori yang diperoleh Berita didapatkan hasil akurasi 79,81%,

Entertaiment didapatkan hasil akurasi 89,80%, Sinetron didapatkan hasil akurasi 73,68% dan FTV didapatkan hasil akurasi 87,74%.

Untuk penelitian selanjutnya harus menambahkan Fitur tambahan dalam pengolahan data seperti Tweed - Based dan TF - IDF untuk memvariasikan pengolahan datanya agar mendapatkan perbandingan akurasi yang lebih baik lagi. Membuat Korpus dan Stopword yang lebih baik lagi agar dalam melakukan pembuatan data tidak terjadi kehilangan kata - kata penting saat preprocessing.

Daftar Pustaka

1. Akhriza, T. (Februari 2017). Aplikasi Web untuk Analisis Sentimen pada Opini Produk dengan Metode Naive Bayes Classifier. *Conference Paper*.
2. Darma, I. B., Perdana, R. S., & Indriati. (Januari 2018). Penerapan Sentimen Analisis Acara Televisi Pada Twitter Menggunakan Support Vector Machine dan Algoritma Genetika sebagai Metode Seleksi Fitur.
3. Murfi, H. (n.d.). Metode Kernel. *Ebook*.
4. Rozi, I. F., Hamdana, E. N., & Alfahmi, M. B. (2018). PENGEMBANGAN APLIKASI ANALISIS SENTIMEN TWITTER MENGGUNAKAN METODE NAÏVE BAYES CLASSIFIER (Studi Kasus SAMSAT Kota Malang).
5. Taufik, & Pamungkas, S. (2018). ANALISIS SENTIMEN TERHADAP TOKOH PUBLIK MENGGUNAKAN ALGORITMA SUPPORT VECTOR MACHINE (SVM).
6. Castillo, C., Mendoza, M., & Poblete, B. (2011). Information Credibility on Twitter, 675–684.
7. Eka Sembodo, J., Budi Setiawan, E., & Abdurahman Baizal, Z. (2016). Data Crawling Otomatis pada Twitter. *Indosc 2016*, (September 2016), 11–16. https://doi.org/10.21108/INDOSC.2016.111
8. Zul, M. I. (2016). *Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi*.
9. Daniel Jurafsky, James H. Martin. 2000. *Speech and Languange Processing*. New Jersey: Prentice Hall PTR Upper Saddle River.
10. Devid. (2019, Mei 12). *GitHub*. Retrieved from https://github.com/masdevid/ID-OpinionWords