Sentiment Analisis Review Film Di IMDB Menggunakan Algoritma SVM

Sentiment Analysis of Film Review at IMDB using SVM algorithm

Ikhsan Subagyo*[1], Lukman Dwi Yulianto[2], Wahyu Permadi[3], Arian Wahyu Dewantara[4], Anggit Dwi Hartanto[5]

[1,2,3,4] S1 Informatika Fakultas Ilmu Komputer Universitas Amikom Yogyakarta,

[5] S2 Teknik Informatika Fakultas Ilmu Komputer Universitas Amikom Yogyakarta

E-mail: *[1]**ikhsan.subagyo@students.amikom.ac.id**, [2]lukman.3728@students.amikom.ac.id, [3]wahyu.permadi@students.amikom.ac.id, [4]arian.dewantara@students.amikom.ac.id, [5]anggit@amikom.ac.id

Abstrak

Sentimen Analisis yaitu suatu penelitian dalam bidang keilmuan Machine Learning yang membahas tentang opini dalam bentuk teks. IMDb adalah website yang telah lama digunakan untuk menyajikan informasi dan berbagi opini antar penikmat film yang ada di seluruh dunia, tanggapan mereka menjadi tolak ukur kesuksesan dari sebuah film. Penelitian yang dilakukan ini bertujuan untuk mengklasifikasi opini positif dan negatif menggunakan SVM dan SVM dengan SGD. Hasil dari klasifikasi memiliki akurasi dengan max features 10000 87,620% untuk SVM dan 87,404% untuk SVM – SGD untuk max features 30000 87,380% untuk SVM dan 86,948% untuk SVM – SGD dan untuk max features 50000 87,268% untuk SVM dan 86,780% untuk SVM – SGD.

***Kata Kunci**—SVM, SGD, IMDb, Klasifikasi, opini, akurasi*

Abstract

Sentiment analysis is a research in the science field of Machine Learning that discusses opinion in text. IMDb is a website that has long been used to present information and share opinions between filmmakers around the world, their response to the success of a film. The research is aimed at classifying positive and negative opinions using SVM and SVM with SGD. Result of classification has accuracy with Max features 10000 87.620% for SVM and 87.404% for SVM – SGD for Max features 30000 87.380% for SVM and 86.948% for SVM – SGD and for Max features 50000 87.268% for SVM and 86.780% for SVM – SGD.

***Keywords**— SVM, SGD, IMDb, classification, opinions, accuracy*

1. Pendahuluan

Sentimen analisis adalah sebuah pemrosesan bahasa dengan menggunakan sebuah pendekatan untuk mendefinisikan bahasa tersebut mengarah ke arah positif atau negatif. Sentimen analisis merupakan aspek yang sangat popular dan memberikan keuntungan yaitu prediksi penjualan dan pengambilan keputusan kepada para investor [1].

Pada saat ini, opini dan komentar masyarakat banyak membahas tentang yang berangkaian dengan aspek ekonomi, perilaku sosial, fenomena alam, perdagangan, pendidikan,

hiburan, dan lain- lain [3]. Berkaitan pada hiburan, film adalah sebuah hiburan yang banyak di minati oleh banyak golongan dari anak-anak, dewasa, dan orang tua. Dari sebuah film pasti banyak sekali masyarakat yang mengomentari sebuah film itu bagus atau tidak. Salah satu tempat untuk melihat komentar atau mengomentari film yaitu IMDB. IMDB merupakan sebuah situs web yang berguna untuk melihat rincian dari film seperti aktor/aktris yang main, sinopsis film, rating, serta komentar/review film yang dapat melihat film tersebut menghasilkan respon positif atau negatif dari para masyrakat yang sudah menonton [2].

Ada banyak contoh algoritma klasifikasi sentimen salah satunya yaitu SVM. SVM merupakan metode yang berfungsi dari prinsip *Structural Risk Minimization (SRM)* yang bertujuan untuk manemukan *hyperplane* terbaik yang membedakan dua buah *input space* [12]. Strategi SRM pada SVM memberikan *error* generalisasi yang lebih kecil daripada yang didapatkan dari pendekatan *Empirical Risk Management* (ERM) pada *neural network* atau metode lainnya [12]. Namun, data yang tidak terstruktur sebagaimana data tekstual akan menjadikan banyaknya atribut saat proses pada klasifikasi [12]. Faktor tersebut mengakibatkan proses klasifikasi menjadi lebih berat, membuat hasil akurasinya menurun hingga 20% [12]. Salah satu proses yang dapat menurunkan atau mengurangi atribut dalam data tekstual yaitu dengan cara filtering. Fungsi dari filtering untuk memisahkan dan menghapus kata yang tidak penting atau bisa dibilang yang tidak berguna pada klasifikasi [12].

Algoritma SVM menghasilkan nilai akurasi tinggi seperti menguji sentimen terhadap wacana politik pada media sosial online, analisis sebuah sentimen komentar mahasiswa pada sistem pembelajaran di perguruan tinggi, analisis pada tweets di Twitter yang mengeluarkan opini tentang produk mobil otomatis dan produk Apple, dan rata-rata tingkat akurasi yang didapatkan adalah 50 %-90 % [4] [5] [8].

Menurut informasi yang telah dijelaskan, akan dilakukan penelitian tentang sentiment analisis review film di IMDB menggunakan algoritma SVM. Data yang diperoleh akan diproses dengan memerlukan *text mining*, kemudian akan dilanjutkan mengklasifikasikan komentar IMDB menjadi dua kelas, yaitu positif dan negatif. Klasifikasi yang dilakukan menerapkan algoritma *Support Vector Machine (SVM)*. Klasifikasi berfungsi memberikan kemudahan kepada pengguna untuk mengetahui opini positif atau negatif. Nilai akurasi yang dihasilkan dengan menggunakan algoritma akan memberikan pengaruh pada hasil klasifikasi.

2. Metode Penelitian

Penelitian ini menempuh beberapa alur yang bertujuan agar proses penelitian ini menjadi runtut dan tertata. Metode pada penelitian ini adalah melakukan eksperimen dan evaluasi menggunakan *tools scikit-learn library* yaitu pustaka perangkat lunak dalam bahasa python yang berisikan alur dan algoritma untuk kebutuhan *machine learning*. Alur tersebut ditunjukan pada Gambar 1:

Gambar 1. Alur Proses Klasifikasi

2.1. Pengambilan Dataset

Dataset yang digunakan adalah 50.000 data dari kumpulan review film didalam website IMDb yang telah dikumpulkan oleh Andrew Mass yang telah dibagi menjadi 2 bagian yaitu: 1) 25.000 review yang digunakan untuk training dan 2) 25.000 review yang digunakan untuk test. Setiap 25.000 dibagi 12.500 untuk review positif dan review negative.

2.2.Pre-processing Data

Tahapan preprocessing adalah tahapan untuk membersihkan, menata dan menstruktur data mentah yang tidak terstruktur dan memiliki banyak *noise* yang berupa tanda baca atau kalimat yang tidak berarti [16]. Preprocessing data memiliki 4 tahap: 1) *Case Folding* untuk membuat semau *text* menjadi *lowercase* dan menghilangkan tanda baca dan tag-tag HTML dikarenakan dataset ini

merupakan hasil *crawling* dari website IMDb 2) *Cleaning* yaitu membersihkan text dari tanda baca atau HTML tag jika hasil *crawling* dari website. 3) Filtering dengan menggunakan stopwords list dengan menghilangkan kata sambung sesuai dengan bag-of-words. 4) Stemming mencari kata dasar dengan mengilangkan kata imbuhan seperti me-, di, -kan dan mengelompokannya dari text data yang telah digunakan di penelitian ini proses stemming ini menggunakan metode SnowBall.

2.3. Transformasi Data

Transformasi data dipenelitian ini memberikan nilai numerik dan menghitung setiap *term* yang berada dalam text review ini. Hal ini perlu dilakukan dikarenakan *term* dapat berbentuk kata atau frase dan agar dokumen atau text dapat diketahui konteksnya oleh sistem maka harus diberi indikator berupa pembobotan berupa nilai biner yang dinamakan *term weight*. Penelitian ini menggunakan metode TF-IDF dengan formula:

$$TF\text{-}IDF_{t,d} = TF_{t,d} \times IDF_t \quad (1)$$

$$IDF_t = log\frac{N}{DF} \quad (2)$$

$TF_{t,d}$ adalah jumlah muncul sebuah term *t* didalam dokumen *d*, N merupakan jumlah semua dokumen DF_t merupakan jumlah dokumen yang memiliki term *t.*

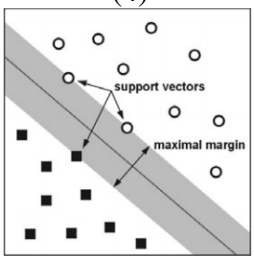
2.4. Klasifikasi SVM(Support Vector Machine)

SVM merupakan teknik supervised machine learning yang dikembangkan oleh Vapnik tahun 1995 dan dikembangkan lebih lanjut oleh Joachims tahun 1998. Berbeda dengan metode terdahulu SVM memiliki konsep dan teori yang terstruktur dan baik sehingga metode ini adalah metode dengan akurasi terbaik dalam bidang klasifikasi text yang digunakan dalam berbagai macam kasus untuk sentiment klasifikasi[7][11]. SVM bekerja dengan membagi data training menjadi 2 kelas dengan memperkirakan garis hyperplane dan mencari jarak maksimal dari *hyperplane* ke data training terdekat agar didapatkan generalisasi untuk proses klasifikasi dengan data test[7][12]. Lihat Gambar 2. Berdasarkan Gambar 2, garis hyperplane berada pada tengah dari nilai maksimal margin dari data terdekat. Persamaan untuk mendapatkan hyperline:

$$\overrightarrow{\omega}.\vec{x} - b = 0 \quad (3)$$

Formula menghitung maksimal margin adalah:

$$\overrightarrow{\omega}.\vec{x} - b = 1, \text{ dan } \overrightarrow{\omega}.\vec{x} - b = -1 \quad (4)$$



Gambar 2 Support Vector Machine

2.5. Klasifikasi SVM (Support Vector Machine) dengan Stochastic Gradient Descent

Stochastic Gradient Descent metode yang afektif dan cocok untuk melatih SVM dalam melakukan berbagai macam klasifikasi maupun kegiatan machine learning lainnya sesuai dengan masalah yang ada [13] [14]. Metode ini bekerja dengan memilih data training yang telah ber-label yang sering digunakan untuk melakukan superviced machine learning secara acak dan melakukan perubahan model bobot melalui Gradient Descent secara instan dan berkelanjutan oleh karena itu SVM dan SGD dapat dipadu agar menangani masalah klasifikasi untuk data yang besar [13] [14].

2.6 Evaluasi

Evaluasi dengan melihat kinerja dari SVM yaitu menggunakan beberapa parameter yaitu Accuracy, precision dan recall dengan menggunakan Confusion Matrix untuk mengetahui seberapa banyak data yang berhasil di klasifikasi dan yang gagal diklasifikasi menggunakan metode SVM dan SVM dengan SGD ini. Untuk bentuk dasar Confusion Matix dapat dilihat pada Tabel 1.

Tabel. 1 Confusion Matrix

	Predicted	
	Positive Documents	Negative Documents
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative(FN)	True Positive (TP)

Accuracy merupakan perbandingan antara semua hasil kasus dengan nilai identifikasi benar[15]. Berikut formula untuk menemukan *Accuracy*.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Recall merupakan kasus positif yang terindentifikasi benar adanya sesuai dengan proporsi yang ada[15]. Berikut formula *Recall*.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Precision adalah proporsi dari hasil positif yang terindentifikasi benar adanya[15]. Berikut formula *Precision*

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

2.7 Eksperimen

Eksperimen dilakukan setelah melatih sistem dengan dataset training yang telah disiapkan agar sistem melakukan pemilahan sesuai dengan review positif atau review negatif.

2.8 Evaluasi Hasil

Setelah eksperimen dilakukan maka dilakukan tahap mengevaluasi sehingga diketahui nilai *Accurasi, Precision* dan *recall* dari metode SVM.

3. Hasil Dan Pembahasan

Hasil dan pembahasan akan dijabarkan dengan beberapa tahap yaitu:

3.1 Persiapan

Persiapan yang dilakukan adalah mempersiapkan alat dan dataset yang akan menjadi objek utama penelitian ini. Penelitian ini dikerjakan menggunakan processor 13-2310M 2.1Ghz dengan RAM 4GB dan menggunakan sistem operasi Windows 7 64-bit yang telah terinstall library python 3.7.3 Anaconda Environment dan menggunakan library scikit-learn untuk membantu melakukan proses klasifikasi ini. Dataset review IMDb yang memiliki 50.000 review dapat digambarkan sesuai dengan tabel berikut:

Tabel 2 Data Review IMDb

Dataset Review Film IMDb			
Data Latihan		Data Test	
Positif	negatif	Positif	negatif
12.500	12.500	12.500	12.500

3.2 Preprocessing Data

Dalam tahap ini data tersebut akan diproses agar dimensi text menjadi lebih sederhana. Contoh text yang akan memasuki preprocessing: **"***When I first see it, I am too young around 15, to understand all the things the movie what to show us. The film gave me a very different feeling. Hope is a dangerous thing, is the root of mental distress.".** Hasil dapat dilihat pada tabel 3 berikut.

Tabel 3 Hasil Preprocessing

Tahap	Output
Cleaning dan Case-Folding	when i first see it i am too young around to understand all the things the movie what to show us the film gave me a very different feeling hope is a dangerous thing is the root of mental distress
Stopword removal	first see young around understand things movie show us film gave different feeling hope dangerous thing root mental distress
Stemming	first see young around understand thing movi show us film gave differ feel hope danger thing root mental distress

Dapat dilihat bahwa melakukan preprocessing data dapat memangkas data yang berawal besar menjadi lebih kecil dan lebih bermakna. Karakter pada data yang belum diprocessing terhitung sebanyak 207 karakter kamudian setelah melewati proses Cleaning dan Case-Folding karakter data menjadi 196, data tersebut memasuki tahap stopwords removal dan berhasil memangkas karakter menjadi 124 karakter, tahap terakhir adalah stemming data tersebut dan berhasil memangkas karakter menjadi 113. Dalam hal ini preprocessing data lebih tepatnya normalisasi data yang berawal 207 karakter menjadi hanya 113 karakter yang dapat diproses untuk klasifikasi.

3.3 Klasifikasi

Tahap ini melakukan eksperimen yang dengan melakukan klasifikasi text menggunakan dalam dataset menggunakan algoritma SVM yang pertama adalah Linier SVM dan SVM yang telah dilatih atau diperbarui menggunakan metode pembelajaran stochastic gradient descent atau dalam penelitian ini kami pesingkat dengan SVM SGD. Untuk SVM linier menggunakan parameter C=0.01 dan tolerasi 0.001 selain itu menggunakan parameter default dari *scikit-learn library* dan untuk SVM SGD menggunakan parameter alpha = 0.001.

Tabel 4 Perbandingan dengan max features 10000

25000 data	SVM Linier	SVM SGD
Accuracy	0.8762000000000000	0.8740400000000000
Precision	0.8636047320807237	0.8511453248216297
Recall	0.8935200000000000	0.9066400000000000
F1-Score	0.8783077104549208	0.8780166569823746

Tabel 5 Perbandingan dengan max features 30000

25000 data	SVM Linier	SVM SGD
Accuracy	0.8738000000000000	0.8694800000000000
Precision	0.8616766003560647	0.8434084318536694
Recall	0.8905600000000000	0.9074400000000000
F1-Score	0.8758802470592864	0.8742533430960730

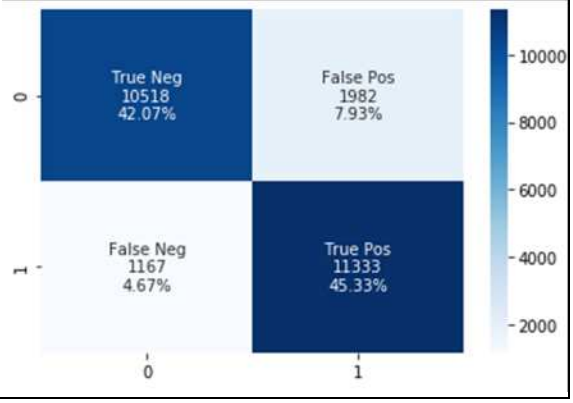
Tabel 6 Perbandingan dengan max features 50000

25000 data	SVM Linier	SVM SGD
Accuracy	0.8726800000000000	0.8678000000000000
Precision	0.8616567036720751	0.8399763366116986
Recall	0.8879200000000000	0.9087200000000000
F1-Score	0.8745912296599819	0.8729969642239557

Dari hasil yang ditampilkan dari tabel 4, tabel 5 dan tabel 6 bahwa nilai SVM sedikit lebih besar melebihi SVM SGD, dalam *Recall* SVM SGD selalu unggul dalam semua uji coba dengan menggunakan *max features* 10000, 30000 dan 50000.

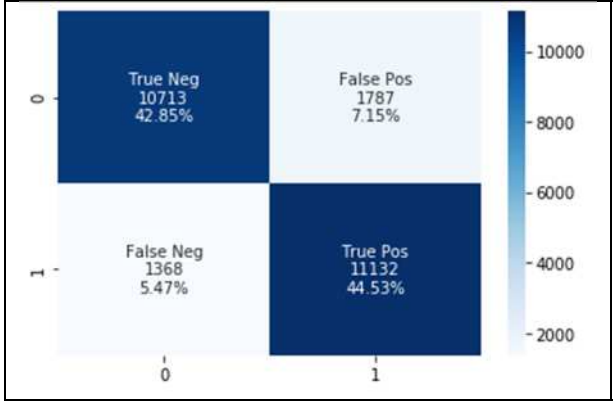
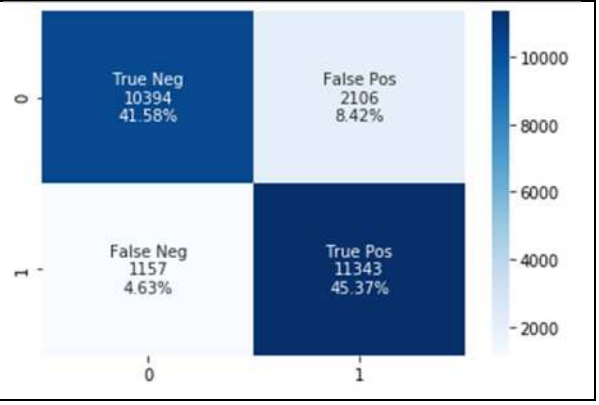
Hasil dari *Accuracy, Precision, Recall* dan *F1-Score* didapat menggunakan *Confusion Matrix*. Setiap metode yang digunakan memiliki *Confusion matrix* sendiri dengan *value* yang berbeda-beda sesuai dengan kinerja metode yang digunakan. Eksperimen ini menggunakan *library Seaborn* untuk membuat diagram *Confusion Matrix* menjadi lebih mudah. Berikut tabel perbandingan *Confusion Matrix* antara SVM dengan SVM SGD.

Tabel. 7 Confesion Matrix

SVM	SVM SGD
	

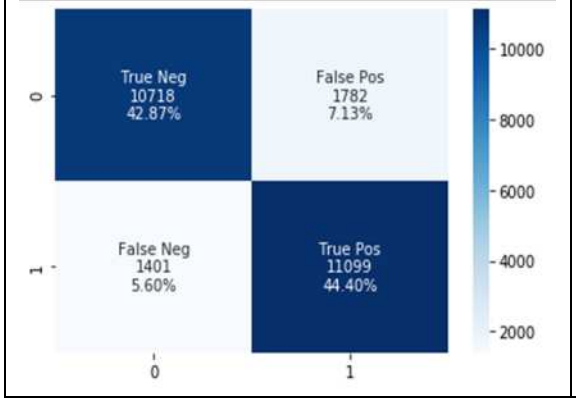
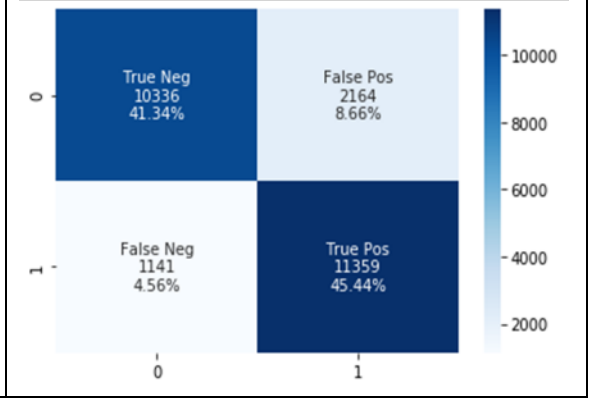
Tabel 7 merupakan perbandingan *Confusion Matrix* antara SVM dengan SVM SGD dengan max features 10000. Dapat dilihat bahwa nilai dari *False Positive* dan *False Negative* dibawah 10% dan untuk *True Positive* dan *True Negative* nilai hampir mendekati 50%, lalu untuk perbandingan dengan *max features* 30000 tertera didalam Tabel 6 dengan nilai *False Positive* dan *False Negative* masih dibawah 10 % namun terjadi peningkatan dan penurunan persentase pada nilai *True Positive* dan *True Negative* namun masih lebih dari 40%.

Tabel 8 Confusion Matrix

SVM	SVM SGD
	

Setelah membandingkan nilai pada Tabel 8 yang memiliki *max features* 30000 maka dilanjutkan dengan membandingkan hasil *Confusion Matrix* untuk *max features* 50000 yang dapat dilihat pada tabel 9 berikut.

Tabel 9 Confusion Matrix

SVM	SVM SGD
	

Dari hasil diatas dapat dilihat bahwa nilai dari *Confusion matrix* untuk *max features* 50000 sedikit berbeda jika untuk *True Positive* SVM SGD lebih besar dari *max features* 30000 namun untuk hasil lainnya berkurang sedikit. Jadi dapat disimpulkan bahwa jika *max features* lebih besar maka hasil akan lebih berkurang karena max feature membatasi *term* terhadap dataset yang menjadi objek penelitian.

4. Kesimpulan

Kesimpulan dari eksperimen ini dari fakta hasil adalah pengujian klasifikasi menggunakan SVM dengan SGD memiliki nilai ketepatan yang hampir sama satu sama lain, nilai accuracy untuk SVM adalah 0.8762 dan untuk SGD adalah 0.87404 atau 87.620% dan 87.404% untuk max features 10000 dan untuk Confusion Matrix didapatkan hasil yang cukup memuaskan dengan kesalahan klasifikasi dapat dilihat pada label False Positive dan False Negative masing-masing memiliki persentase kurang dari 10% atau kurang dari 2000 kata dan untuk label True Positive dan True Negatif hamprr menyentuh 50% dengan nilai antara 40% - 45%.

5. Saran

Saran untuk penelitian selanjutnya adalah:

1. Perlu adanya penelitian tentang parameter terbaik yang dapat digunakan untuk mendapatkan hasil klasifikasi yang lebih baik.
2. Karena pada penelitian ini menggunakan bahasa Python maka perlu adanya penggunakan bahasa pemrograman lain atau menggunakan tools yang telah tersedia untuk mempengaruhi hasil dalam proses klasifikasi.
3. Menggunakan SVM dengan kernel lain sepert rbf untuk memberikan perbandingan dengan linier SVM yang telah digunakan dalam penelitian ini.

Daftar Pustaka

[1] Sing,V.K., Piryani,R., Uddin, A., Waila, P., 2013, Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), Kottayam, 2013

[2] Chandani. V., 2015, Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film, *Journal of Intelligent System(JIS),* Vol. 1, No 1.

[3] Rahutomo, F., Saputra, P. Y., Miftahul, 2018, Implementasi Twitter Sentiment Analysis Untuk Review Film Menggunakan Algoritma Support Vector Machine. *Journal Informatika Polinema(JIP),* Vol. 4, No. 2, Hal 93 - 99

[4] Hidayat A.N, 2015, Sentimen Analisis Terhadap Wacana Politik Pada Media Massa Online Menggunakan Algoritma Support Vector Machine dan Naïve Bayes. ***Jurnal Elektronik Sistem Informasi dan Komputer (JESIK)* , Vol. 1, No. 1**

[5] Habibi, M , 2017, Analisis Sentimen dan Klasifikasi Komentar Mahasiswa Pada Sistem Evaluasi Pembelajaran Menggunakan kombinasi KNN Berbasis Cosine Similarity dan Supervised Model. *Tesis,* Program Pasca Sarjana Ilmu Komputer, Univ. Gadjah Mada, Yogyakarta

[6] Saleh, M.R., Martín-Valdivia, M.T., Montejo-Ráez A., Ureña-López, L.A, 2011, Experiments With SVM to Classify Opinions in Different Domains, *Expert Systems with Applications*, Volume 38, Issue 12, Hal. 14799-14804.

[7] Moraes, R., Valiati, J.F., Neto,W.P.G., 2012, Document-Level Sentiment Classification: An Empirical Comparison Between SVM and ANN, *Expert Systems with Applications*, Volume 40, Issue 2, Hal. 621-633.

[8] Ahmad, M., Aftab, S., Ali, I., 2017, Sentiment Analysis of Tweets using SVM, *International Journal of Computer Applications*, Vol 5, Hal 25-29

[9] Huq, M.R., Ali, A., Rahman, A., 2017, Sentiment Analysis on Twitter Data using KNN and SVM, *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 6, Hal. 19 - 25.

[10] Wongso, R., Luwinda, F.A., Trisnajaya, B.C.,Rusli, O., Rudy, 2017, News Article Text Classification in Indonesian Language, *Procedia Computer Science,* Vol. 116, Hal. 137-143

[11] Vishal A.K., Sonawane .S.S, 2016, Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications*, Vol.139 ,No. 11

[12] Putranto, H.A., Setyawati ,O, Wijono, 2016, *Pengaruh Phrase Detection dengan POS-Tagger terhadap Akurasi Klasifikasi Sentimen Menggunakan Sentimen.*

[13] Lu,Shuxia., Jin, Zhao., 2017, Improved Stochastic Gradient Descent Algorithm for SVM, *International Journal of Recent Engineering Science(IJRES).* Vol 4, Issue 4, Hal. 28 – 31.

[14] Lu,Shuxia., Jin, Zhao., 2016, Map Reduce – based SVM Ensemble with Stochastic Gradient Descent, *International Journal of Recent Engineering Science(IJRES).* Vol 5, Issue 12, Hal. 206 – 211.

[15] Mustafa, M.S., Ramadhanm M.R, 2017, Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier, *Creative Information Technology Journal (CITEC Journal),* No. 2, Vol. 4, Hal. 151-162.

[16] Santoso, E.B., Nugroho, A., 2019, Analisis Sentimen Calon Presiden Indonesia 2019 Berdasarkan Komentar Publik di Facebook*, Jurnal Eksplora Informatika,* No.1, Vol. 9, Hal. 60-69