

**ALGORITMA *SUPPORT VECTOR MACHINE* (SVM) DAN *NAÏVE BAYES*
PADA ANALISIS SENTIMEN SOSIAL MEDIA TWITTER
(Studi Kasus Kuliah Luring Setelah Pandemi Covid-19)**

SKRIPSI

Diajukan untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Matematika



Oleh:

Zahra Agusfiyanti Nurlaila

NIM 1905657

**PROGRAM STUDI MATEMATIKA
DEPARTEMEN PENDIDIKAN MATEMATIKA
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA**

2023

LEMBAR HAK CIPTA

**ALGORITMA *SUPPORT VECTOR MACHINE* (SVM) DAN *NAÏVE BAYES*
PADA ANALISIS SENTIMEN SOSIAL MEDIA TWITTER
(Studi Kasus Kuliah Luring Setelah Pandemi Covid-19)**

Oleh:

Zahra Agusfiyanti Nurlaila

NIM 1905657

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Matematika pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

© Zahra Agusfiyanti Nurlaila 2023
Universitas Pendidikan Indonesia

Hak Cipta dilindungi undang-undang.

Skripsi ini tidak boleh diperbanyak seluruhnya atau sebagian, dengan dicetak ulang, fotokopi, atau cara lainnya tanpa izin dari penulis.

LEMBAR PENGESAHAN

ZAHRA AGUSFIYANTI NURLAILA

**ALGORITMA SUPPORT VECTOR MACHINE (SVM) DAN NAÏVE BAYES
PADA ANALISIS SENTIMEN SOSIAL MEDIA TWITTER
(Studi Kasus Kuliah Luring Setelah Pandemi Covid-19)**

Disetujui dan disahkan oleh :

Pembimbing I



Dr. Dadan Dasari, M.Si.
NIP. 196407171991021001

Pembimbing II



Dr. Lukman, S.Si., M.Si.
NIP. 196801281994021001

Mengetahui,

Ketua Program Studi Matematika



Dr. Kartika Yulianti, S.Pd., M.Si.
NIP. 199207282005012001

ABSTRAK

Setelah melewati masa pandemi Covid-19, berbagai sektor dalam kehidupan manusia banyak mengalami perubahan, termasuk dalam kegiatan pendidikan. Hal tersebut membuat berbagai pendapat atau tanggapan mengenai kegiatan pendidikan banyak dituangkan dalam media sosial. Untuk mengetahui sentimen tanggapan masyarakat tersebut perlu dilakukan analisis sentimen dengan algoritma *machine learning*. Metode yang digunakan dalam penelitian ini adalah *Support Vector Machine*, *Naïve Bayes Classifier*, dan *Lexicon Based*. SVM digunakan untuk proses klasifikasi analisis sentimen dan mencari nilai akurasi terbaik menggunakan *kernel linear*. *Naïve Bayes* digunakan untuk proses klasifikasi analisis sentimen dan sebagai metode perbandingan terhadap SVM. *Lexicon Based* digunakan untuk menentukan kelas sentimen positif, netral dan negatif pada data. Hasil penilaian dari 3522 data *tweet* diperoleh 766 *tweet* (21.7%) positif, 254 *tweet* netral (7.2%), dan 2502 *tweet* negatif (71%). Metode klasifikasi SVM memiliki tingkat akurasi sebesar 83% , presisi sebesar 78%, dan *recall* sebesar 55. Sedangkan metode *Naïve Bayes* memiliki tingkat akurasi sebesar 73%, nilai presisi 58% dan *recall* sebesar 34%. Berdasarkan hasil analisis sentimen ini, dapat disimpulkan bahwa performa metode SVM lebih baik dalam mengklasifikasi data dibandingkan *Naïve Bayes*.

Kata Kunci: Analisis Sentimen, Komentar, Opini, *Twitter*, *Support Vector Machine*, SVM, *Naïve Bayes*, *Lexicon Based*, Kuliah, Luring

ABSTRACT

After going through the Covid-19 pandemic, various sectors in human life experienced many changes, including in educational activities. This causes various opinions or responses regarding educational activities to be poured on social media. To find out the sentiment of the public's response, it is necessary to carry out sentiment analysis with a machine learning algorithm. The methods used in this research are Support Vector Machine, Naïve Bayes Classifier, and Lexicon Based. SVM is used for the sentiment analysis classification process and to find the best accuracy value using a linear kernel. Naïve Bayes is used for the sentiment analysis classification process and as a comparison method against SVM. Lexicon Based is used to determine the positive, neutral and negative sentiment classes in the data. The results of the assessment of 3522 tweet data obtained 766 positive tweets (21.7%), 254 neutral tweets (7.2%), and 2502 negative tweets (71%). The SVM classification method has an accuracy rate of 83%, a precision of 78%, and a recall of 55. Meanwhile the Naïve Bayes method has an accuracy rate of 73%, a precision value of 58% and a recall of 34%. Based on the results of this sentiment analysis, it can be concluded that the performance of the SVM method is better in classifying data than Naïve Bayes.

Keywords: *Sentiment Analysis, Comments, Opinion, Twitter, Support Vector Machine, SVM, Naïve Bayes, Lexicon Based, Lecture, Offline*

DAFTAR ISI

LEMBAR HAK CIPTA	i
LEMBAR PENGESAHAN	ii
LEMBAR PERNYATAAN	iii
KATA PENGANTAR.....	iv
UCAPAN TERIMA KASIH.....	v
ABSTRAK	vi
<i>ABSTRACT</i>	vii
DAFTAR ISI	viii
DAFTAR TABEL	xi
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN	xiv
BAB I PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah	3
1.3. Tujuan Penelitian	3
1.4. Manfaat Penelitian.....	4
1.5. Batasan Penelitian.....	4
BAB II KAJIAN TEORI	5
2.1. Analisis Sentimen.....	5
2.1.1. Tipe Analisis Sentimen	6
2.1.2. Cara Kerja Analisis Sentimen	7
2.2. Machine Learning.....	9
2.2.1. Jenis Machine Learning	9
2.3. <i>Text Mining</i>	12
2.4. <i>Text Preprocessing</i>	12
2.5. <i>Lexicon Based</i>	13
2.6. Klasifikasi	14

2.6.1.	<i>Support Vector Machine</i>	15
2.6.2.	<i>Naïve Bayes</i>	17
2.6.2.1.	Teorema Bayes	17
2.6.2.2.	Metode Naïve Bayes	17
2.7.	Pembobotan TF-IDF	19
2.8.	Sistem Evaluasi	19
2.9.	<i>Big Data</i>	21
2.10.	Media Sosial.....	21
2.11.	<i>Twitter</i>	22
2.11.1.	Peran <i>Big Data</i> dalam <i>Twitter</i>	22
2.12.	<i>Python</i>	23
2.13.	Penelitian Terdahulu.....	23
BAB III METODE PENELITIAN		25
3.1.	Metode Pengumpulan Data.....	25
3.1.1.	Studi Pustaka.....	25
3.1.2.	<i>Crawling Data</i>	25
3.2.	Pengolahan Data	26
3.2.1.	<i>Data Preprocessing</i>	27
3.2.2.	Pelabelan Data.....	31
3.2.3.	Pembobotan TF-IDF	32
3.2.4.	Klasifikasi Menggunakan <i>Support Vector Machines</i>	33
3.2.5.	Klasifikasi Menggunakan <i>Naïve Bayes</i>	34
3.3.	Evaluasi Sistem	35
3.4.	Diagram Alur Penelitian	36
BAB IV HASIL DAN PEMBAHASAN		38
4.1.	Pengambilan Data (<i>Crawling Data</i>).....	38
4.2.	<i>Data Preprocessing</i>	40
4.2.1.	Cek Duplikat.....	41
4.2.2.	<i>Case Folding</i>	42

4.2.3.	Cleaning Data	43
4.2.4.	Normalisasi	44
4.2.5.	<i>Tokenizing</i>	45
4.2.6.	<i>Remove Stopwords</i>	46
4.2.7.	Stemming	47
4.3.	Pelabelan Data.....	48
4.4.	<i>Modelling</i>	52
4.4.1.	<i>Data Split</i>	52
4.4.2.	Pembobotan TF-IDF	52
4.5.	Pengujian Model Klasifikasi.....	59
4.6.	<i>Sistem Evaluasi</i>	60
4.6.1.	<i>Confussion Matrix</i> Model SVM.....	60
4.6.2.	<i>Confusion Matrix</i> Model <i>Naïve Bayes</i>	61
4.6.3.	Perbandingan Performansi SVM dan <i>Naïve Bayes</i>	63
4.7.	Visualisasi	64
BAB V KESIMPULAN		66
5.1.	Kesimpulan	66
5.2.	Saran	66
DAFTAR PUSTAKA		68
LAMPIRAN		73

DAFTAR TABEL

Tabel 2.1 Contoh hasil pelabelan teks menggunakan kamus <i>Lexicon</i>	14
Tabel 2.2 Fungsi Kernel.....	16
Tabel 2.3 Confusion Matrix 2x2.....	20
Tabel 3.1 <i>Sample Crawling Data Tweet</i>	25
Tabel 3.2 Confusion Matrix 3x3.....	35
Tabel 4.1 Hasil <i>Preprocessing Case Folding</i>	42
Tabel 4.2 Hasil <i>Preprocessing Cleaning Data</i>	43
Tabel 4.3 Hasil <i>Preprocessing Normalisasi</i>	44
Tabel 4.4 Hasil <i>Preprocessing Tokenizing</i>	45
Tabel 4.5 Hasil <i>Preprocessing Remove Stopwords</i>	46
Tabel 4.6 Hasil <i>Preprocessing Stemming</i>	47
Tabel 4.7 Hasil <i>Scoring</i> pada Data <i>Tweet</i>	49
Tabel 4.8 Hasil <i>Labelling</i> pada <i>Scoring Data Tweet</i>	49
Tabel 4.9 Contoh <i>Tweet</i> Setelah <i>Preprocessing</i> yang akan dinilai bobotnya	53
Tabel 4.10 <i>Corpus</i> dari Beberapa <i>Tweet</i>	53
Tabel 4.11 Perhitungan TF.....	54
Tabel 4.12 Perhitungan DF.....	55
Tabel 4.13 Perhitungan IDF.....	56
Tabel 4.14 Perhitungan TF-IDF.....	57
Tabel 4.15 <i>Confusion Matrix Support Vector Machine</i>	60
Tabel 4.16 <i>Confusion Matrix Naïve Bayes</i>	61
Tabel 4.17 Perbandingan Performansi Model SVM dan Naïve Bayes.....	63
Tabel 4.18 Hasil Perbandingan Akurasi Penelitian	64

DAFTAR GAMBAR

Gambar 2.1 Support Vector Machine	15
Gambar 3.1 Flowchart Proses Cek Duplikat	27
Gambar 3.2 <i>Flowchart</i> Proses <i>Case Folding</i>	28
Gambar 3.3 <i>Flowchart</i> Proses <i>Cleaning Data</i>	28
Gambar 3.4 <i>Flowchart</i> Proses Normalisasi	29
Gambar 3.5 <i>Flowchart</i> Proses <i>Tokenizing</i>	30
Gambar 3.6 <i>Flowchart</i> Proses <i>Remove Stopwords</i>	30
Gambar 3.7 <i>Flowchart</i> Proses <i>Stemming</i>	31
Gambar 3.8 Tahapan Pelabelan Data.....	32
Gambar 3.9 Alur Penelitian Model SVM.....	33
Gambar 3.10 Alur Penelitian Model <i>Naïve Bayes</i>	34
Gambar 3.11 Diagram Alur Penelitian	36
Gambar 4.1 Pencarian Kata Kunci Dalam Aplikasi Twitter.....	38
Gambar 4.2 Kode Program <i>Crawling</i> Data	39
Gambar 4.3 Hasil Pengumpulan Data dengan <i>Crawling</i>	39
Gambar 4.4 Hasil <i>Crawling</i> dari Dua Kata Kunci.....	40
Gambar 4.5 <i>Import Library</i> yang akan Digunakan dalam Proses Pengolahan Data ...	41
Gambar 4.6 Kode Program dan tahap Cek Duplikat	41
Gambar 4.7 Kode Program tahap <i>Case Folding</i>	42
Gambar 4.8 Kode Program tahap <i>Cleaning Data</i>	44
Gambar 4.9 Kode Program tahap Normalisasi	45
Gambar 4.10 Kode Program tahap <i>Tokenizing</i>	46
Gambar 4.11 Kode Program tahap <i>Remove Stopwords</i>	47
Gambar 4.12 Kode Program Proses <i>Stemming</i>	48
Gambar 4.13 Hasil Data <i>Preprocessing</i>	48
Gambar 4.14 Kode Program Pelabelan Data.....	50
Gambar 4.15 Hasil Perhitungan <i>Labelling</i> Data <i>Tweet</i>	51

Gambar 4.16 <i>Pie Chart</i> Hasil Perhitungan <i>Labelling Data Tweet</i>	51
Gambar 4.17 Pembagian Data Latih dan Data Tes.....	52
Gambar 4.18 Hasil dari Data Split.....	52
Gambar 4.19 Kode Program tahap Pembobotan TF-IDF	59
Gambar 4.20 Hasil Pengujian SVM Kernel Linear	59
Gambar 4.21 Hasil Pengujian Model <i>Naïve Bayes</i>	60
Gambar 4.22 Diagram Perbandingan Performansi Model SVM dan <i>Naïve Bayes</i> ...	63
Gambar 4.23 Visualisasi <i>Wordcloud</i>	64
Gambar 4.24 10 Kata Paling Sering Muncul dalam <i>Tweets</i>	65

DAFTAR LAMPIRAN

Lampiran 1 Pengujian Parameter Optimal Naïve Bayes.....	73
Lampiran 2 Pengujian Parameter Optimal SVM	74
Lampiran 3 Source Code Visualisasi.....	75
Lampiran 4 <i>Slang Words</i>	76
Lampiran 5 Hasil Data <i>Crawling</i>	84

DAFTAR PUSTAKA

- Admin LP2M. (2022, February 21). *Analisis Sentimen (Sentiment Analysis) : Definisi, Tipe dan Cara Kerjanya*. Lembaga Penelitian Dan Pengabdian Masyarakat (LP2M) Universitas Medan Area.
- Alencar, P., & Cowan, D. (2018). *The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review Ivens Portugal*.
- Alfi, M., Reynaldi, R., & Sibaroni, Y. (2015). *Analisis Sentimen Review Film pada Twitter menggunakan Metode Klasifikasi Hybrid SVM, Naïve Bayes, dan Decision Tree*.
- Arini, A.-, Wardhani, L. K., & Octaviano, D.-. (2020). Perbandingan Seleksi Fitur Term Frequency & Tri-Gram Character Menggunakan Algoritma Naïve Bayes Classifier (Nbc) Pada Tweet Hashtag #2019gantipresiden. *KILAT*, 9(1), 103–114. <https://doi.org/10.33322/kilat.v9i1.878>
- Camelia, D. (2023, February 16). *Apa Itu Analisis Sentimen : Pengertian, Tipe, Dan Cara Kerjanya*. Kazee.Id.
- Darwis, D., Shintya Pratiwi, E., Ferico, A., & Pasaribu, O. (2020). PENERAPAN ALGORITMA SVM UNTUK ANALISIS SENTIMEN PADA DATA TWITTER KOMISI PEMBERANTASAN KORUPSI REPUBLIK INDONESIA. In *Jurnal Ilmiah Edutic* (Vol. 7, Issue 1).
- Esti. (2023, August 9). *Mengenal Apa Itu Analisis Sentimen, Tipe dan Cara Kerjanya*. Mekari Qontak.
- Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi. *Teknika*, 8(1), 11–16. <https://doi.org/10.34148/teknika.v8i1.95>

- Fauzi, A. (2020). Studi Perbandingan Metode Analisis Naive Bayes Classifier dengan Support Vector Machine untuk Analisis Sentimen (Studi Kasus: Tweet Berbahasa Indonesia tentang Covid-19). Yogyakarta.
- Fauzi, R. R. (2022) Analisis Sentimen Dampak Ekonomi Masyarakat Indonesia Akibat Pandemi Covid-19 Pada Media Sosial Twitter Menggunakan Metode Naive Bayes Classifier, Support Vector Machine Dan Lexicon Based. Jakarta.
- Feinerer, I., Hornik, K., & Meyer, D. (2008). *Journal of Statistical Software Text Mining Infrastructure in R*. <http://www.jstatsoft.org/>
- Feldman, R., & Sanger, J. (2007). *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge University Press.
- Ghahramani, Saeed. (2005). *Fundamentals of probability, with stochastic processes*.
- Han, J., & Kamber, M. (2006). *Data Mining Concept and Techniques*.
- Hendry, J. (2012, Mei 30). Prior Probability Vs Posterior Probability. Retrieved 8 5, 2023 from scribd.com: <https://www.scribd.com/doc/95263438/Prior>
- Kashina, M., Lenivtceva, I. D., & Kopanitsa, G. D. (2020). Preprocessing of unstructured medical data: The impact of each preprocessing stage on classification. *Procedia Computer Science*, 178, 284–290. <https://doi.org/10.1016/j.procs.2020.11.030>
- Kurniasari, D. (2021, May 6). *Algoritma Supervised Learning vs Unsupervised Learning Cari Tahu Perbedaannya Disini*. DQLAB AI Powered Learning.
- Kwok, L., & Yu, B. (2013). Spreading Social Media Messages on Facebook An Analysis of Restaurant Business-to-Consumer Communications. *Cornell Hospitality Quarterly*, 54, 84–94. <https://doi.org/10.1177/1938965512458360>
- LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature* 521, 436–444.

- Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Mukhtar, N., Khan, M. A., & Chiragh, N. (2018). Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. *Telematics and Informatics*, 35(8), 2173–2183. <https://doi.org/https://doi.org/10.1016/j.tele.2018.08.003>
- Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, 4, 51–62. <https://doi.org/10.20544/horizons.b.04.1.17.p05>
- Nindito, H. (2016). TEORI TEXT MINING DAN WEB MINING. <https://sis.binus.ac.id/2016/12/15/teori-text-mining-dan-web-mining/>
- Novakovic, J., Veljovic, A., Ilić, S., Papić, Ž. M., & Milica, T. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, 7, 39–46. <https://api.semanticscholar.org/CorpusID:1586327>
- Nugroho, A. S., Witarto, A. B., & Handoko, D. (2003). *Support Vector Machine-Teori dan Aplikasinya dalam Bioinformatika 1*. <http://asnugroho.net>
- P, R. D. L., Faticah, C., & Purwitasari, D. (2017). Deteksi Gempa Berdasarkan Data Twitter Menggunakan Decision Tree, Random Forest, dan SVM. *Jurnal Teknik ITS*, 6(1), 153–158.
- Pujianto, U., & Ristanti, P. Y. (n.d.). *Jurusan Teknik Elektro, Universitas Negeri Malang, Indonesia | Maret 2019 U. Pujianto, Putri Yuni Ristanti | Perbandingan kinerja metode C4.5 dan Naive Bayes dalam klasifikasi ...* (Vol. 29). <http://journal2.um.ac.id/index.php/tekno>
- Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*.

- Rania, D. (2022, June 28). *Apa Itu Big Data? Arti, Konsep, Contoh, dan Manfaatnya*. Rumah Web.
- Rish, I. (2006). *An empirical study of the naive Bayes classifier*.
- Robertson, S. (2004). Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation - J DOC*, 60, 503–520. <https://doi.org/10.1108/00220410410560582>
- S. N. & F. K. Asiyah, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine dan K-Nearest Neighbor," *Jurnal Sains & Seni*, pp. 317-322, 2016.
- Sammut, C., & Webb, G. (2010). Encyclopedia of Machine Learning. In *Encyclopedia of Machine Learning: , ISBN 978-0-387-30768-8. Springer Science+Business Media, LLC, 2010*. <https://doi.org/10.1007/978-0-387-30164-8>
- Schmidhuber, J. (2014). *Deep Learning in Neural Networks: An Overview*. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Sulardi, N., & Witanti, A. (2020). Sistem Pakar Untuk Diagnosis Penyakit Anemia Menggunakan Teorema Bayes. *Jurnal Teknik Informatika*, 1(1), 19–24. <https://doi.org/10.20884/1.jutif.2020.1.1.12>
- Tiara, Sabariah, M. K., & Effendy, V. (2015). Sentiment analysis on Twitter using the combination of lexicon-based and support vector machine for assessing the performance of a television program. *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, 386–390. <https://api.semanticscholar.org/CorpusID:18189720>
- Wahyudi, R., Kusumawardhana, G., Purwokerto, A., Letjend, J., Soemarto, P., Purwanegara, K., Purwokerto, T., & Banyumas, K. (2021). Analisis Sentimen pada review Aplikasi Grab di Google Play Store Menggunakan Support Vector Machine. *JURNAL INFORMATIKA*, 8(2). <http://ejournal.bsi.ac.id/ejurnal/index.php/ji>

- Winiarti, S. (2013). Pemanfaatan dalam penentuan penyakit tht. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699.
- Wira, J., & Putra, G. (2020) *Pengenalan Konsep Pembelajaran Mesin dan Deep Learning Edisi 1.4*.
- Xu, J., Zhang, Y., & Miao, D. (2020). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, 507, 772–794. <https://doi.org/https://doi.org/10.1016/j.ins.2019.06.064>
- Xu, J., Zhang, Y., Wu, Y., & Wang, J. (2015). *Citation Sentiment Analysis in Clinical Trial Papers Article in AMIA*. <http://www.ncbi.nlm.nih.gov/pmc/>
- Zoqi Sarwani, M., & Mahmudy, W. (2015). *Analisis twitter untuk mengetahui karakter seseorang menggunakan algoritma naïve bayess classifier*.
- Zukhrufillah, I. (2018). *Gejala Media Sosial Twitter Sebagai Media Sosial Alternatif* (Vol. 1, Issue 2). <https://id.wikipedia.org/wiki/Twitter>