JITE (Journal of Informatics and Telecommunication Engineering)

Available online <http://ojs.uma.ac.id/index.php/jite> DOI : 10.31289/jite.v3i2.3153

Indonesian Text Dataset for Determining Sentiment Classification Using Mechine Learning Approach

Indra Edy Syahputra¹⁾*, Tulus¹⁾ & Syahril Efendi¹⁾

1) Prodi S2 Teknik Informatika, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Indonesia

*Email: indraedysyahputra@gmail.com

Abstrak

Kemajuan teknologi informasi mendorong munculnya informasi tekstual tanpa batas dengan penggunaan media *online* berkembang begitu cepat sehingga munculnya kebutuhan dalam penyajian informasi tanpa mengurangi nilai dari informasi yang disajikan. Pada dasarnya konsep dari dataset merupakan bentuk umum dari hampir setiap disiplin ilmu, dimana dataset memberikan informasi dasar yang empiris untuk kegiatan penelitian. Analisis sentimen dilakukan untuk melihat pendapat atau perasaan terhadap suatu permasalahan atau mengidentifikasi serta mengklasifikasikan kecenderungan informasi dari permasalahan tersebut. Analisis dataset dalam menentukan klasifikasi sentimen merupakan model dari klasifikasi sentimen yang memiliki relevansi dengan dataset dengan penggunaan teknik pembelajaran mesin dengan pengawasan yang belajar dari pengalaman untuk memprediksi keluaran dari data input berlabel dan output dari pembelajaran mesin. Hasil percobaan dan pengujian yang telah dilakukan pada teknik pembelajaran mesin dengan pengawasan dapat melakukan klasifikasi sentimen pada teks tweet dengan baik dan tingkat keakurasian masih dapat ditingkatkan ke arah yang lebih baik dengan data yaitu *baseline* 100 (hari) dan 83 (minggu), *naivebayes* 100 (hari) dan 82 (minggu), *maxent* 100 (hari) dan 83 (minggu), dan *SVM* 100 (hari) dan 83 (minggu).

Kata Kunci: *Mechine learning, klasifikasi, sentimen, dataset.*

Abstract

Advances in information technology encourage the emergence of unlimited textual information with the use of online media developing so rapidly that the emergence of the need for information presentation without reducing the value of the information presented. Basicaly the concept of the dataset is a general form of almost every discipline, where the dataset provides empirical basic information for research activities. Sentiment analysis is done to see opinions or feelings about a problem or identify and classify information trends from the problem. The dataset analysis in determining sentiment classification is a model of sentiment classification that has relevance to the dataset with the use of machine learning techniques with supervision that learns from experience to predict output from labeled input data and output from machine learning. The results of experiments and tests that have been carried out on machine learning techniques with supervision can classify sentiments in the tweet text properly and the level of accuracy can still be improved to a better direction with data namely baseline 100 (days) and 83 (weeks), naivebayes 100 (days) and 82 (weeks), maxent 100 (days) and 83 (weeks), and SVM 100 (days) and 83 (weeks).

Keywords: *Mechine learning, classificatin, sentiment, dataset.*

How to Cite: Syahputra, I, E. Tulus. Efendi, S. (2020). Indonesian Text Dataset for Determining Sentiment Classification Using Mechine Learning Approach. JITE (Journal of Informatics and Telecommunication Engineering). 3(2): 192-201.

I. PENDAHULUAN

Pada dasarnya konsep dari suatu dataset merupakan bentuk umum dari hampir setiap disiplin ilmu, dimana dataset memberikan informasi dasar yang empiris untuk kegiatan penelitian. Meletakkan pondasi dan definisi yang jelas dari tujuan analisis dataset untuk memprediksi klasifikasi sentimen, merupakan hal yang paling penting dan mendasar pada penelitian ini. Tujuannya adalah bagaimana membangun model dan spesifikasi dari klasifikasi sentimen yang memiliki relevansi dengan dataset yang digunakan. Teknik yang dipakai dalam melakukan analisis dataset teks berbahasa Indonesia, adalah teknik pembelajaran mesin dengan pengawasan yang dapat digambarkan sebagai salah satu metode pembelajaran dari pengalaman yang digunakan untuk memprediksi keluaran dari data input yang berlabel, dan output dari pembelajaran mesin dengan pengawasan adalah memberikan sampel data yang tepat. (Harrington, 2012).

Klasifikasi sentimen merupakan suatu teknik bagaimana mengelompokkan sebuah teks berbahasa Indonesia ke dalam kategori yang tepat dengan bentuk berupa label atau kelas yang dikategorikan ke dalam tiga bagian, yaitu positif, negatif, dan neutral berdasarkan keseluruhan

pendapat atau informasi. Pada penelitian ini akan dijelaskan bagaimana melakukan analisis dataset teks berbahasa Indonesia, sehingga dataset tersebut digunakan sebagai data latih, dan data pengujian dalam klasifikasi sentimen pada pembelajaran mesin. Dataset tesk berbahasa Indonesia yang dimaksud di dalam penelitian ini adalah data yang diambil dari teks tweet twitter yang dikumpulkan berdasarkan *keyword* atau kata kunci, dan kemudian data teks twitter tersebut dikategorikan berdasarkan label atau kelas positif, negatif, dan neutral.

Beberapa penelitian klasifikasi sentimen yang telah dilakukan diantaranya adalah penelitian yang dilakukan oleh Janadhana, (2012) dalam penelitian analisis sentimen teks tweet dan opini data mining, dimana dalam penelitian ini diperkenalkan pendekatan secara otomatis untuk mengklasifikasikan sentiment dari pesan teks twitter dan algoritma yang digunakan adalah *Naïve Bayes*, *SVM*, dan *MaxEnt*. Berikutnya adalah penelitian yang dilakukan oleh (Esiyok & Albayrak, 2015) dengan penelitian pelacakan teks tweet untuk memprediksi sentimen pada tren pemasaran, dimana penelitian ini menyajikan klasifikasi sentimen untuk empat merek perusahaan yang terdiri dari

Facebook, Twitter, Apple, dan Microsoft dengan sistem berbasis web, algoritma yang digunakan adalah *Naïve Bayes* dan *Maximum Entropy*. Berikutnya adalah (Benamara, Cesarano & Reforgiato, 2006) dengan penelitian, representasi N-Gram berbasis dimensi rendah dalam mengklasifikasikan dokumen. Penelitian ini menjelaskan tentang model *The Bag Words* (BOW), teknik yang digunakan di dalam penelitian ini adalah *Latent Semantic Analysis* (LSA), dan *Latent Dirichlet Allocation* (LDA). Kemudian (Basu & Shepherd, 2002) dalam jurnal "Analisis Sentimen Twitter untuk Teks Berbahasa Indonesia menggunakan algoritma *Maximum Entropy* dan *Support Vector Machine* (SVM), kemudian pengklasifikasian *Tweet Twitter* menggunakan fitur unigram (Pang, *et al*, 2002), dan pengklasifikasian data yang dilatih menggunakan teknik pembelajaran mesin dengan pengawasa. (Richert & Coelho, 2013), mengklasifikasi-kan sentimen pada *Twitter* menggunakan *Emoticon* sebagai label untuk mengklasifikasikan sentimen positif dan negatif untuk mengurangi ketergantungan pada teknik pembelajaran mesin. Serta penelitian oleh (Tong & Koller, 2001) klasifikasi teks vektor menggunakan *Support Vector Machine* (SVM) berbasis

webmining sption fosil analisis penyebaran penyakit tropis, penelitian ini bertujuan untuk mengembangkan model sistem klasifikasi teks menggunakan algoritma SVM untuk mengklasifikasikan informasi tekstual pada website. (Dergiades, 2012).

II. STUDI KASUS

Machine Learning

Pembelajaran mesin atau *machine learning*, merupakan bidang ilmu komputer yang fokus pada pengembangan dan implementasi sistem yang bertujuan untuk meningkatkan kemampuan mesin atau komputer dalam menghadapi lebih banyak data (Pustejovsky & Stubbs, 2013). Mengutip perkataan dari *Herbert Simmon*, pemenang hadiah nobel dalam bidang ilmu ekonomi bahwa "*Belajar adalah proses dimana suatu sistem meningkatkan kinerja dari pengalaman*". Di bidang teknologi bahasa dan perhitungan *linguistic*, topik yang paling penting dalam pembelajaran mesin salah satunya adalah yang berhubungan dengan suasana hati atau *mood*, pengaruh atau klasifikasi sentimen teks atau ucapan.

Machine learning merupakan design dan studi tentang kecerdasan perangkat lunak, menggunakan pengalaman masa lalu untuk membuat keputusan di masa

depan. Dalam artian pembelajaran mesin adalah suatu program yang belajar dari data (Heckeling, 2014). Tujuan dari pembelajaran mesin untuk menjeneralisasikan, mempengaruhi aturan yang tidak diketahui dari aturan aplikasi.

Salah satu dari pembelajaran mesin adalah penyaringan email yang dikategorikan sebagai spam dengan cara mengamati ribuan email yang sebelumnya sudah dikategorikan sebagai spam atau bukan spam. Dimana penyaringan email yang dikategorikan sebagai spam merupakan proses pembelajaran mesin untuk mengklasifikasikan email kedalam spam atau bukan spam. Pada prinsipnya pembelajaran mesin menggunakan beberapa strategi dan teknik yang dapat digunakan untuk melakukan pendekatan terhadap suatu masalah.

Naïve Bayes Classifier (NBC)

Naïve Bayes merupakan salah satu algoritma pembelajaran mesin yang paling elegan dan sangat praktis dalam penggunaannya. Bukanlah suatu yang naif ketika melihat kinerja dari algoritma ini. Pendekatan dengan menggunakan algoritma pembelajaran mesin *naïve bayes* terbukti cukup kuat untuk tidak mengabaikan fitur yang tidak relevan dan

dapat belajar dengan cepat dan memprediksi suatu sentimen dengan tidak memerlukan banyak penyimpanan (Richert & Coelho, 2013).

Dikarena pada probabilitas tidak dapat diperkirakan secara langsung maka perlu digunakan teknik yang menemukan model *bayes* sebagai berikut:

(Richert & Coelho, 2013).

$$P(A).P(B|A)=P(B).P(A|B) \quad (1)$$

Jika dilakukan substitusi pada A dengan probabilitas kedua fitur $F1$ dan $F2$ yang terjadi dari B sebagai kelas C akan membantu untuk mengambil probabilitas dari contoh data yang memiliki kelas tertentu.

$$P(F1, F2) \cdot (P(C | F1, F2) = P(C) \cdot P(F1, F2 | C) \quad (2)$$

Hal ini dapat memungkinkan untuk mengekspresikan probabilitas $P(C | F1, F2)$ dengan cara lainnya sebagai berikut:

$$P(C | F1, F2) = P(C) \cdot P(F1, F2 | C) P(F1, F2) \quad (3)$$

III. METODE PENELITIAN

Sesuai dengan tujuan penelitian yaitu analisis dataset teks berbahasa Indonesia untuk menentukan klasifikasi sentimen menggunakan pendekatan pembelajaran mesin dengan pengawasan, maka pada bab ini akan dimulai dengan pembahasan tentang gambaran analisis dataset teks

berbahasa Indonesia dan gambaran studi kasus algoritma pembelajaran mesin dengan pengawasan yang akan dibahas.

Analisis dataset teks berbahasa Indonesia tidaklah semudah yang digambarkan, ini dikarenakan data yang digunakan merupakan data tweet pada platform twitter yang memiliki karakteristik yang unik, dengan panjang karakter 140 atau kurang. Dan atribut-atribut yang dimiliki oleh tweet menjadikan hal yang berbeda, belum lagi penggunaan model bahasa dari media. Selain itu tingkat frekuensi kesalahan pengejaan dan penggunaan bahasa yang tidak memiliki standar. (Chandani, Wahono & Purwanto, 2015).

Berikut adalah persiapan data sebelum digunakan sebagai dataset pelatihan dan pengujian.

Gambar 1. Proses Persiapan Analisis Dataset Teks Berbahasa Indonesia

Agar dapat membangun sebuah model sistem yang dapat digunakan

sebagai pembelajaran mesin pada perangkat lunak komputer yang mampu memahami bahasa alami manusia dalam bentuk teks yang dapat digunakan untuk menentukan klasifikasi sentimen dengan menggunakan teknik pembelajaran mesin dengan pengawasan atau *supervised learning*.

Pada penelitian ini, pendekatan permasalahan yang digunakan dalam analisis dataset untuk menentukan klasifikasi sentimen dengan teknik pembelajaran mesin dengan pengawasan menggunakan pendekatan fitur ekstraktor yang unigrams dan bigrams dengan kata kunci positif dan negatif. Serta membangun kerangka yang memperlakukan pengklasifikasian dan fitur ekstraktor sebagai dua komponen yang berbeda.

Gambar 2. Feature List Generator

Adapun proses pengumpulan dataset yang berupa data tweet twitter adalah sebagai berikut:

Gambar 3. Proses Pengumpulan Data *Tweet*

Berikut adalah diagram pada metode penelitian dengan menggunakan teknik pembelajaran mesin dengan pengawasan untuk menentukan klasifikasi sentimen.

Gambar 4. Metodologi Penelitian

Penelitian ini akan menjelaskan lebih dekat model probabilitas algoritma *naivebayes* yang digunakan dan bagaimana menerapkan konsep algoritma *naivebayes* di dalam melakukan klasifikasi teks pada penerapan sistem yang akan dirancang pada penelitian ini. Proses dilakukan dengan memilih algoritma pembelajaran mesin dimuali dengan baseline, maxent, *naivebayes*, dan SVM.

Untuk dapat memahami bagaimana klasifikasi *naivebayes* bekerja disini akan dijelaskan secara singkat intisari dari konsep atau aturan bayes yang dikembangkan oleh *Thomas Bayes* (1701-1761). Model probabilitas yang dirumuskan sangat sederhana dapat dituliskan dalam kata-kata yang sederhana (Pustejovsky, & Stubbs, 2013). Adapun rumusnya adalah sebagai berikut:

$$\text{Probabilitas akhir} = \frac{\text{Probabilitas kondisional} \times \text{Probabilitas awal}}{\text{Bukti}}$$

Bila dilakukan perhitungan dari nilai probabilitas dari sampel dataset yang tersebut pada tabel diatas, dapat diasumsikan sebagai berikut:

$$P(\text{Class}|\text{Pos}_i), \text{Pos}_i = [\text{pak}, \text{ada}, \text{yang}, \text{harus}] \quad (4)$$

Dimana:

Posisi menjadi fitur vektor dari sampel dataset $i, i \in \{D1, D2, \dots, Dn\}$

$D1j$ menjadi notasi dari kelas $j, j \in \{a, b, \dots, k\}$ dan $P(\text{Pos}|D1)$ merupakan probabilitas observasi pada sampel dataset yang memiliki kelas $D1$.

Maka notasi umum dari probabilitas akhir atau posterior probability dapat dituliskan sebagai berikut:

$$P(\text{Class}|\text{Data}) = \frac{P(\text{class}|\text{data}) \times P(\text{data})}{P(\text{class})} \quad (5)$$

Fungsi objektif dalam *naivebayes* probability adalah untuk memaksimalkan probabilitas akhir yang diberikan data pelatihan untuk dapat merumuskan

keputusan dengan aturan yang cerdas, dan tepat.

$$\text{Prediksi kelas label} \leftarrow \underset{j=1, \dots, m}{\operatorname{argmax}} P(\text{class} | \text{Data}) \quad (6)$$

Dari data sampel pada tabel 3.1 tersebut diatas, maka penelitian ini membuat suatu aturan keputusan berdasarkan probabilitas akhir sebagai berikut:

Klasifikasi sentimen dikatakan positif jika $(\text{Pos} | \text{Data}) \geq P(\text{Neg} | \text{Data}) \geq P(\text{Neu} | \text{Data})$, maka akan dikelompokkan pada sentimen positif.

Support vector machine (SVM) merupakan metode pembelajaran mesin dengan pengawasan atau supervised learning adalah metode yang populer untuk menyelesaikan permasalahan pengklasifikasian dan regresi. Pada penelitian ini menggunakan LIBSVM yang merupakan librari SVM untuk mengklasifikasikan suatu teks ke dalam kelas positif, negatif, dan neutral.

Penggunaan SVM memerlukan dua langkah melatih dataset untuk memperoleh suatu model, dan menggunakan model yang diperoleh untuk memprediksi informasi yang berkaitan dengan data uji dataset (Chang & Lin, 2011). Tujuan penggunaan algoritma SVM adalah untuk menghasilkan suatu model berdasarkan dataset pelatihan yang digunakan untuk memprediksi nilai-nilai

target dataset pengujian dari data atribut (Chen, & Tian, 2009). Formulasi SVM yang digunakan dalam penelitian ini adalah *C-support vector classification*. Dimana vektor-vektor $X_i \in R^n, i=1, \dots, l$, yang masing-masing memiliki label 1 atau 2, dan sebuah vektor indikator $y \in R^n$, sedemikian rupa sehingga $y_i \in \{1, -1\}$, dimana dalam menyelesaikan *primal optimization* sebagai berikut:

$$\begin{aligned} \min_{w, b, \epsilon} \quad & \frac{1}{2} W^T W + C \sum_{i=1}^l \epsilon_i \\ \text{Subject to} \quad & y_i (W^T \phi(X_i) + b) \geq 1 - \epsilon_i \\ & \epsilon_i \geq 0, i = 1, \dots, l \end{aligned} \quad (7)$$

IV. HASIL DAN PEMBAHASAN

Pada tahap bagian ini akan dilakukan serangkaian penelitian, percobaan, serta pembahasan terhadap proses analisis dataset teks berbahasa Indonesia untuk menentukan klasifikasi sentimen menggunakan pendekatan pembelajaran mesin dengan pengawasan. Pada tahap ini akan dilakukan beberapa percobaan terhadap beberapa data sampel yang dijadikan dataset baik digunakan untuk data pelatihan maupun digunakan untuk pengujian data. Bagian ini juga akan dijelaskan penggunaan algoritma pembelajaran mesin dengan pengawasan (*supervised learning*) dan melakukan optimalisasi serta evaluasi untuk menentukan suatu keputusan

mengklasifikasikan sentimen berdasarkan data yang telah dilakukan observasi dan pengujian.

Akhir pembahasan pada bagian ini akan di jelaskan kinerja dari masing-masing algoritma pembelajaran mesin dengan pengawasan yang berkaitan dengan tugas-tugas pembelajaran mesin dengan pengawasan pada sisi kelemahan dan keunggulan dari masing-masing algoritma dengan menampilkan hasil output dari masing-masing percobaan.

Percobaan ini merupakan proses pelatihan algoritma pembelajaran mesin dengan pengawasan *naïve bayes*. Sebagai ilustrasi berikut adalah sampel dataset tweet twitter yang digunakan untuk melakukan percobaan pada algoritma pelatihan *naivebayes*. Sampel dataset tweet dengan total tweet 71 yang terdiri dari tiga kelas positive 22 tweet, negative 27 tweet, dan neutral 22 tweet.

Dari percobaan yang dilakukan diatas sangat diperlukan untuk menemukan suatu data dari dokumen-dokumen tidak terstruktur untuk dapat memberikan informasi yang dibutuhkan. Tujuan adalah agar dapat memberikan kebutuhan informasi yang relevan antara dokumen dan data test yang sedang di analisis.

Salah satu sistem yang baik adalah sistem yang memungkinkan pengguna dapat menentukan secara cepat dan akurat apakah dokumen yang digunakan dapat memenuhi kebutuhan data yang sedang dianalisis. Maka untuk itu agar dapat memenuhi tujuan yang ingin dicapai perlu diperhatikan bagaimana menyusun dokumen atau data yang ditampilkan secara terstruktur dari dokumen yang memiliki tingkat relevansi tinggi sampai kepada dokumen yang memiliki tingkat relevansi rendah. Penyusunan dokumen tersebut merupakan suatu teknik yang disebut dengan perengkingan dokumen.

Tabel 1. Kondisional *Sampling Dataset*

N _{doc}	Term (W _j)	FITUR (X _i)											Class	N _{class}	Prior
		D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11			
22	D1	1	1	1	1	0	0	0	0	0	0	0	Pos	11	0.31
	D2	1	1	1	1	1	1	1	1	1	1	0	Pos	11	
	D3	1	1	1	1	0	0	0	0	0	0	0	Pos	11	
	D4	1	1	1	1	0	0	0	0	0	0	0	Pos	11	
27	D5	1	1	1	1	1	1	1	1	1	1	1	Neg	11	0.38
	D6	1	1	1	1	1	1	1	0	0	0	0	Neg	11	
	D7	1	1	1	1	1	1	1	1	1	0	0	Neg	11	
	D8	1	1	1	1	1	1	1	1	0	0	0	Neu	11	
22	D9	1	1	1	1	1	1	1	1	1	1	1	Neu	11	0.31
	D10	1	1	1	0	0	0	0	0	0	0	0	Neu	11	
71		10	10	10	9	6	6	6	5	4	3	2		110	1.00

Tabel 2. Kondisional Probabilitas Sampel Dataset

KONDISIONAL PROBABILITI											
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	
0.023	0.023	0.023	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.000	
0.023	0.023	0.023	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
0.023	0.023	0.023	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	
0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.000	0.000	0.000	
0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.000	0.000	
0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.000	0.000	0.000	
0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	0.023	
0.023	0.023	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
0.20	0.20	0.20	0.18	0.11	0.11	0.11	0.10	0.08	0.06	0.04	

Tabel 3. Posterior Probabilitas Sampel Dataset

D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	Total
0.007	0.007	0.007	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028
0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.000	0.070
0.007	0.007	0.007	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028
0.007	0.007	0.007	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028
0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.063
0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.000	0.040
0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.000	0.000	0.052
0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.000	0.000	0.000	0.056
0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.077
0.007	0.007	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.021
0.067	0.067	0.067	0.059	0.038	0.038	0.038	0.033	0.026	0.020	0.013	0.465

Tabel 4. Pengujian Klasifikasi Teks Berbahasa Indonesia

Data Test	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	Class
DT	Punya	Atau	Cuma	Kebetulan	Saja	Jelas	Pak	0	0	0	0	?
	1	1	1	1	1	1	2	0	0	0	0	

Berikut adalah hasil percobaan ujicoba sistem berdasarkan percobaan dengan menggunakan keyword atau kata kunci KPK, Kapolri, dan Ahok secara real time pada tweet twitter dengan periode per hari dan per minggu.

Gambar 6. Base Klasifikasi KPK

Gambar 7. Base Klasifikasi KPK Analisis Per Minggu

Gambar 8. Naivebayes Klasifikasi KPK

Gambar 9. Naivebayes Klasifikasi KPK

Gambar 10. Maxent Klasifikasi KPK

Gambar 11. Maxent Klasifikasi KPK Per Minggu

Gambar 12. LIBSVM Klasifikasi KPK

Gambar 13. LIBSVM Klasifikasi KPK Per Minggu

V. SIMPULAN

Dalam penelitian ini menghasilkan beberapa kesimpulan yaitu dengan beberapa percobaan dan pengujian pada teknik pembelajaran mesin dengan pengawasan dapat melakukan klasifikasi sentimen pada teks tweet dengan baik dan tingkat ke akurasian masih dapat

ditingkatkan ke arah yang lebih baik. Hasil percobaan dan pengujian yang telah dilakukan pada teknik pembelajaran mesin dengan pengawasan dapat melakukan klasifikasi sentimen pada teks tweet dengan baik dan tingkat keakurasian masih dapat ditingkatkan ke arah yang lebih baik dengan data yaitu *baseline* 100 (hari) dan 83 (minggu), naivebayes 100 (hari) dan 82 (minggu), maxent 100 (hari) dan 83 (minggu), dan SVM 100 (hari) dan 83 (minggu).

DAFTAR PUSTAKA

- Basu, A., Watters, S., & Shepherd, M., (2002). Support Vector Machines for Text Categorization. *Proceesings of the 36th Hawai International Conference on System Sciences* (HICSS'03), 0-7695-1874-5/03 © 2002 IEEE.
- Benamara, F., Cesarano, C., & Reforgiato, D., (2006). Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone. *ICWSM'2006 Boulder, CO USA*.
- Chandani, V., Wahono, S.R., Purwanto., (2015). Komparasi Algoritma Klasifikasi Machine Learning dan Fature Selection pada Analisis Sentimen Review Film. *Journal of Intelligent System*, Vol. 1, No. 1, February 2015 ISSN : 2365 – 3982.
- Dergiades, T., (2012). Do Investors Sentiment Dynamics Affeck Stock Return. Evidence from the USA Economy. *Economucs Letters*, 116 (3), 404-407. Doi: 10.1016/j.econlet.2012.04.018.
- Chang, C.C., & Lin J.C., (2011). LIBSVM: A Library for Support Vector Machine. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Article No.27, Volume 2 Issue 3, April 2011. DOI: 10.1145/1961189.1961199.
- Chen, J., Huang, H., & Tian, S., (2009). Feature Selection for Text Classification with Naïve Bayes. *Expart System Application*, 36 (3), 5432-5435.
- Heckeling, G., 2014. Mastering Machine Learning with Scikit-Learn: Apply effective learning algoritms to real-word problems using scikit-learn. First Published: Oktober 2014. ISBN: 978-1-78398-836-5. Production Reference: 1221014. Published by Packt Publishing Ltd, Livery Place, 35 Livery Street, Birmingham B3 2PB, UK.
- Esiyok, C., & Albayrak, S., (2015). Twitter Sentiment Tracking for Predicting Marketing Trends. © Springer International Publishing Switzerland 2015. F.Hopfgartnet (ed), *Smart Information Systems, Advances in Computer Vision and Pattern Recognition*, DOI: 10.1007/978-3-319-14178-7_2.
- Harrington, P., (2012). Machine Learning in Action. Jeff Bleiel. ISBN: 9781617290183. Special Sales Department, Manning Publication Co, 20 Baldwin Road PO Box 261 Shelter Island, NY 11964. Printed in the United States of America.
- Janadhana, R., (2012). Twitter Sentiment Analysis and Opinion Mining. Dapartmen of Computer Science, University of North Carolina at Chapel.
- Pang, O.B., & Lee, L., (2008). Opining Mining and Sentimen Analysis. *Foundations and Trend® in Information Retrieval*, Vol. 2, Nos. 1 – 2 (2008) 1 – 135 DOI: 10.1561 / 15000000001.
- Pang, O.B., Lee, L., Vaithyanathan, S., (2002). Thumbs up? Sentiment Classification using Machine Learning Tecniques. *Appers in Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv:cs/0205070v1, [cs.CL] 28 May 2002
- Richert, W., & Coelho, P.L., (2013). Building Machine Learning System with Python. Master the art of machine learning with Python and build effective machine learning systems with this intensive hand-on guide. ISBN: 978-1-78216-140-0. First Published: July 2013. Production Reference: 1200713. Published by Packt Publishing Ltd. Livery Place. 35 Livery Street. Birmingham B3 2PB, UK.
- Pustejovsky, J., & Stubbs, A., (2013). Natural Language Annotation for Machine Learning. Editors: Steele, J., & Blanchette, M. First Edition: Oktober 2012. ISBN: 978-1-449-30666-3, [LSI]. Publish by O'Reilly Media, Inc., 1005. Gravenstein Highway North, Sabastopol, CA 95472. Copyright © 2013 Pustejovsky, J., & Stubbs, A. All rights reserved. Printed in the United States of America.
- Tong, S., & Koller, D., (2001). Support Vector Machine Active Learning with Application to Text Classification. *Journal of Machine Learning Research*, Submitted 10/01; Published 11/01, 45-66.