# ANALISIS SENTIMEN TERHADAP CALON PRESIDEN 2019 DENGAN *SUPPORT VECTOR MACHINE* DI TWITTER

**[1]Digna Tata Lukmana, [2]Sri Subanti, [3]Yuliana Susanti**
[1,2,3]Program Studi Statistika, FMIPA Universitas Sebelas Maret, (0271)669376
e-mail: tasktata99@gmail.com

**Abstrak**

Indonesia merupakan negara demokrasi dengan penduduk terbanyak keempat dan pengguna *Twitter* terbesar kelima di dunia. Pemilihan presiden 2019 di Indonesia menjadi suatu topik yang menarik bagi para pengguna *Twitter*. *Tweet* masyarakat yang berkaitan dengan para calon presiden dapat digunakan untuk melihat gambaran opini masyarakat terhadap para calon presiden. Banyaknya jumlah *tweet* yang masuk mengenai para calon presiden mendorong perlunya metode yang membantu untuk melihat opini masayarakat secara efektif. Salah satu metode yang dapat digunakan untuk mengklasifikasikan opini masyarakat secara efektif adalah *Support Vector Machine*. Metode ini akan mengklasifikasikan apakah suatu opini masyarakat akan termasuk dalam sentimen positif atau negatif dengan mencari *hyperlane* terbaik dari kedua kelas klasifikasi. Penambahan fungsi *Kernel* pada *Support Vector Machine* berguna untuk mengatasi data yang tidak terpisah secara *linier*. Hasil dari klasifikasi didapatkan akurasi sebesar 86%.

**Kata Kunci:** Pemilihan Presiden 2019, Analisis Sentimen, Klasifikasi, *Support Vector Machine*.

**Abstract**

Indonesia is a democracy with the fourth largest population and the fifth largest Twitter user in the world. The 2019 Presidential Election 2019 in Indonesia became an interesting topic for Twitter users. Public tweets about presidential candidates can be used to see the public opinion on presidential candidates. The large number of tweets that come in regarding presidential candidates encourages the need for methods that help to see people's opinions effectively. One method that can be used to classify public opinion effectively is Support Vector Machine. This method will classify whether a public opinion will be included in positive or negative sentiments by finding the best hyperlane from both classification classes. Adding Kernel function to Support Vector Machine is useful to resolve data that is not linearly separable. The results of the classification obtained the accuracy of 86%.

**Keywords:** Presidential Election 2019, Sentimen Analysis, Classification, Support Vector Machine.

## PENDAHULUAN

Indonesia merupakan negara yang memiliki jumlah penduduk terbesar keempat di dunia dan menganut sistem demokrasi. Salah satu penerapan demokrasi di Indonesia adalah adanya pemilihan umum (Pemilu) untuk memilih wakil-wakil rakyat. Tahun 2019 akan menjadi tahun diadakannya pemilu serentak dalam memilih presiden serta wakilnya atau pemilihan presiden (Pilpres), dan pemilihan legislatif (Pileg).

Jumlah Daftar Pemilih Tetap (DPT) dalam Pilpres 2019 mencapai 195 juta pemilih. Hal yang menarik dari pemilu 2019 ini adalah komposisi DPT yang sebagian besar didominasi generasi muda yang berusia 20-34 tahun. Penduduk yang berusia 20-34 tahun akan disebut secara sederhana sebagai kelompok milenial (Howe & Stratus, 2007). Pada tahun 2019 jumlah penduduk yang berusia milenial, diproyeksi sebanyak 24,32% dari total populasi Indonesia yang mencapai 267 juta jiwa (BPS, 2018). Menurut Asosiasi Penyedia Jasa Internet Indonesia (APJII), komposisi generasi milenial mendominasi penggunaan internet. Pengguna internet di Indonesia mencapai 143,26 juta orang, dari angka tersebut 95% adalah pengguna media sosial. Indonesia menempati peringkat 4 pengguna *Facebook* terbesar dan peringkat 5 pengguna *Twitter* terbesar di dunia. *Twitter* menjadi ajang bagi para Calon Presiden 2019 dalam menaikkan popularitasnya. Pilpres 2019 menjadi suatu topik yang menarik bagi para pengguna *Twitter*.

Pang (2002) membandingkan tiga teknik *machine learning* untuk klasifikasi sentimen. Teknik tersebut adalah *Naïve Bayes Classifier* (NBC), *Maximum Entropy* dan *Support Vector Machine* (SVM). Hasil yang didapat adalah SVM memberikan akurasi tertinggi yaitu sebesar 82,7%. Kumar dan Sebastian

(2012) melakukan penelitian mengunakan metode berbasis *corpus* untuk menemukan orientasi semantik dari kata sifat, dan metode berbasis kamus untuk menemukan orientasi semantik kata kerja. Kesimpulan yang didapatkan adalah perkembangan situs *microblogging* seperti *Twitter* memberikan kesempatan baru untuk mengaplikasikan teori dan teknologi yang dapat digunakan untuk penambangan sentimen atau emosi masyarakat. Nurirwan, d.k.k (2015) meneliti sentimen masyarakat terhadap Jokowi. Data yang digunakan adalah data mengenai Jokowi yang didapatkan dari *Facebook*, *Twitter*, dan *blog* politik, pada 9 Juli hingga 20 Oktober 2014. Penelitian ini membandingkan metode SVM dengan metode NBC pada 6 kali percobaan. Hasil dari keenam percobaan tersebut menunjukkan SVM cenderung memberikan hasil yang lebih baik dengan akurasi tertinggi mencapai 88%.

SVM mampu mengklasifikasikan data dengan cara mencari *hyperplane* terbaik dari data. Penerapan algoritma klasifikasi SVM dalam analisis sentimen dengan data yang diambil dari *Twitter* diharapkan dapat mengklasfikasikan *tweet* kedalam sentimen positif atau negatif. Berdasarkan uraian diatas, permasalahan yang diteliti adalah membahas akurasi SVM dalam mengklasifikas *tweet* dan melihat hasil visualisasi sentiment positif dan negatif.

## METODE PENELITIAN

Penelitian ini menggunakan data yang diperoleh dengan kode akses *Twitter* API melalui proses *crawling* di R, dari 18 Januari hingga 18 Februari 2019. Kata kunci yang diganakan dalam pengambilan data adalah nama kedua calon presiden, yaitu "Jokowi" dan "Prabowo". Data yang diambil adalah berupa *tweet* masyarakat yang memuat kata kunci tersebut.

### *Text Mining*

*Text Mining* adalah proses menganalisis teks atau bahasa untuk mengekstrak informasi berharga dalam rangka mencapai tujuan tertentu (Witten, 2005). Sama seperti *data mining*, *text mining* juga bertujuan untuk mencari pola pada teks. Beberapa aplikasi dalam *text mining* antara lain:

1. *Text Categorization*, untuk menentukan satu atau lebih kategori yang belum didefinisikan pada sebuah teks.
2. *Spell and Grammar Cheking*, untuk mengecek ejaan kata dan menyarankan kata alternatif.
3. *Topic Modeling*, untuk menemukan topik yang ada pada kumpulan dokumen.
4. *Information Extraction*, menemukan informasi relevan terhadap pertanyaan pengguna.
5. *Question Answering*, berkaitan dengan *artificial intelegence* biasa terdapat pada aplikasi *auto-chat* dan *google assistant*.
6. *Sentimen Analisis*, mengidentifikas sentimen dan opini pada teks.

### *Analisis Sentimen*

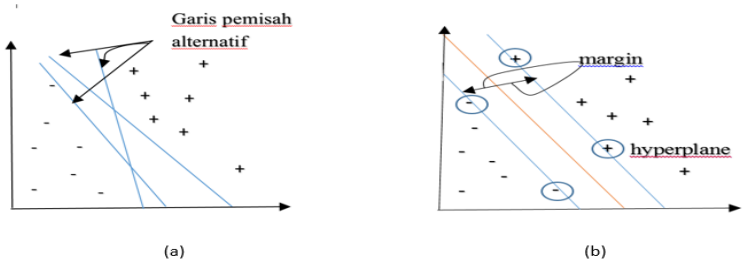
Analisis sentimen merupakan bagian dari *text mining* yang sangat sering dilakukan. Analisis sentimen adalah bidang ilmu yang menganalisis opini orang-orang, sentimen, evaluasi, dan emosi terhadap produk, layanan, individu, organisasi, masalah, topik, peristiwa tertentu (Liu, 2012). Terdapat 3 tahap dalam melakukan analisis sentimen, yaitu *preprocessing*, pembobotan dan klasifikasi.

1. *Preprocessing* merupakan tahap persiapan data yang bertujuan agar mempermudah proses pengolahan data. *Preprocessing* memfokuskan pada data *cleaning* & *cleansing*, termasuk menghilangkan *noise* di data, mengatasi struktur data yang tidak baik, dan informasi yang hilang. Berikut tahapan yang ada dalam *preprocessing* data:
    1) *Cleansing*, yaitu membersihkan data dari *noise* seperti *hashtag*, *username*, url, dan tanda baca.
    2) *Case folding*, merupakan tahap mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar yang konsisten secara keseluruhan (dalam hal ini huruf kecil).
    3) Menghapus *Stopword*, merupakan tahap untuk membuang kata-kata yang tidak penting seperti "yang", "di", "ke" dan seterusnya.
    4) *Stemming*, merupakan tahap untuk merubah kata-kata dalam kalimat menjadi kata dasar.
    5) Tokenisasi, merupakan proses memecah kalimat menjadi kata-kata.
    6) *Filtering*, merupakan tahap mengambil kata-kata penting dari hasil token dengan cara membuang kata-kata yang tidak penting.

2. Pembobotan, setelah melalui *preprocessing* data yang berupa kata-kata diubah menjadi vektor dan diberikan nilai pembobot untuk setiap kata agar bisa dihitung dan diolah dengan algoritma klasifikasi. Pembobotan dilakukan pada setiap kata memiliki kepentingan yang berbeda-beda dalam dokumen. Metode *Term Frequency* (TF) menentukan bobot *term* pada suatu dokumen berdasarkan jumlah kemunculannya dalam dokumen tersebut.
3. Klasifikasi, dalam mengklasifikasi diperlukan input berupa himpunan data latih yang berlabel atau memiliki atribut kelas untuk memetakan ke output yang berupa model klasifikasi

### *Support Vector Machine* (SVM)

*Support Vector Machine* memaksimalkan batas *hyperplane (maximal margin hyperplane*), seperti yang diilustrasikan pada Gambar 1. Berikut




**Gambar 1. SVM berusaha mencari *hyperplane* terbaik**

Gambar 1 (a) menunjukkan *hyperplane* pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan pola terdekat dari masing-masing kelas. Pola yang paling dekat ini disebut sebagai *support vector*. Garis *solid* pada Gambar 1 (b) menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan titik merah dan kuning yang berada dalam lingkaran hitam adalah *support vector*. Data dinotasikan sebagai $x \in R^d$ dengan $R^d$ adalah ruang vektor , dan $d > 1$. Sedangkan label masing-masing kelas dinotasikan $y_i \ni \{-1,+1\}$ untuk $i = 1,2,\dots,n$. Diasumsikan kedua kelas dapat terpisah secara sempurna oleh *hyperplane* berdimensi $d$, yang didefinisikan :

$$\vec{w}.\vec{x} + b = 0 \quad (1)$$

$$\vec{w}.\vec{x} + b \leq -1 \quad (2)$$

$$\vec{w}.\vec{x} + b \geq +1 \quad (3)$$

$\vec{w}$ adalah *vector* bobot, $\vec{x}$ adalah vektor data (*input*) dan $b$ adalah bias. *Pattern* $\overrightarrow{x_i}$ yang termasuk kelas -1 (sampel negatif) dapat dirumusukan sebagai *pattern* yang memenuhi (2), dan *pattern* $\overrightarrow{x_i}$ yang termasuk kelas +1 (sampel positif) dapat dirumusukan sebagai *pattern* yang memenuhi (3). *Margin* terbesar dapat ditemukan dengan memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$ dengan $\|w\|$ adalah vektor normal. Masalah ini dapat dipecahkan dengan berbagai teknik komputasi, diantaranya *lagrange multiplier* yang dinyatakan seagai (4)

$$L(w,b,a) = \frac{1}{2}\|w\|^2 - \sum_i^l a_i \left(y_i\left((\vec{w}.\vec{x} + b) - 1\right)\right) \quad (4)$$

Nilai optimal dari (4) dapat dihitung dengan meminimalkan $L$ terhadap $w$ dan $b$, dan memaksimalkan $L$ terhadap $a_i$, dengan memperhatikan sifat bahwa pada titik optimal *gradient L = 0* sehingga didapatkan (5) dan (6) sebagai pemisah bidang.

$$w = \sum a_i y_i x_i \quad (5)$$

$$b = y_k - w^T x_k \quad (6)$$

### *Kernel Trick*

Pada umumnya masalah dalam domain dunia nyata jarang yang bersifat *linear separable* dan kebanyakan bersifat *non linear*. Fungsi *Kernel* dapat digunakan untuk menyelesaikan masalah SVM *non linear*. Data SVM *non linear* $\vec{x}$ dipetakan oleh fungsi $\phi(\vec{x})$ ke ruang *vector* yang berdimensi lebih tinggi. Pemetaan ini dilakukan dengan menjaga topologi data, dalam artian dua data yang berjarak dekat pada *input space* akan berjarak dekat juga pada *feature space*, sebaliknya jika dua data yang berjarak jauh

pada *input space* maka akan berjarak jauh juga pada *feature space*. Selanjutnya proses pembelajaran pada SVM hanya bergantung pada *dot product* dari data yang sudah ditransformasikan pada ruang baru yang berdimensi lebih tinggi, yaitu $\phi(x_i).\phi(x_j)$. Karena pada umumnya transformasi $\phi$ tidak diketahui maka perhitungan *dot product* dapat digantikan dengan fungsi *kernel* $x_i , x_j$ yang mendefinisikan secara implisit fungsi $\phi$ transformasi tersebut. Dengan kata lain, *Kernel Trick* adalah hasil kali dalam (*inner product*) pada ruang fitur, yang diformulasikan sebagai

$$K(x_i , x_j) = \phi(x_i).\phi(x_j) \quad (7)$$

dengan $x_i , x_j$ adalah pasangan data training dan $\phi$ *adalah* fungsi pemetaan dari *inner space* ke *feature space.*

**Evaluasi Model**

Evaluasi terhadap suatu model klasifikasi umumnya dilakukan terhadap sebuah data uji dengan ukuran tertentu yang tidak digunakan dalam data pelatihan. Model klasifikasi yang dibuat adalah pemetaan dari suatu baris data dengan output sebuah prediksi kelas atau target dari data tersebut. Klasifikasi yang memiliki output dua kelas dinamakan klasifikasi biner. Kedua kelas tersebut biasa dipresentasikan dengan {0,1}, {+1,-1} atau {positif, negatif}. Terdapat empat kemungkinan yang terjadi pada proses klasifikasi suatu baris data (Fawcett, 206)

1. *True positive* (TP), yaitu jika data positif dan diprediksi positif.
2. *False positive* (FP), yaitu jika data positif dan diprediksi negatif.
3. *True negative* (TN), yaitu jika data negatif dan diprediksi negatif.
4. *False negative* (FN), yaitu jika data negatif dan diprediksi positif.

Salah satu ukuran yang dapat digunakan untuk mengevaluasi model klasifikasi, adalah akurasi. Akurasi adalah jumlah proporsi prediksi yang benar.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (8)$$

***Wordcloud***

*Wordcloud* merupakan visualisasi dari output dalam sentimen analisis yang mrnggambarkan karakterisitik dari teks. Wordcloud berupa kumpulan kata-kata yang berbeda ukuran besar hurufnya. Semakin besar tampilan suatu kata dalam *wordcloud* maka semakin besar pula frekuensi kemunculnya (Latifah, 2018).

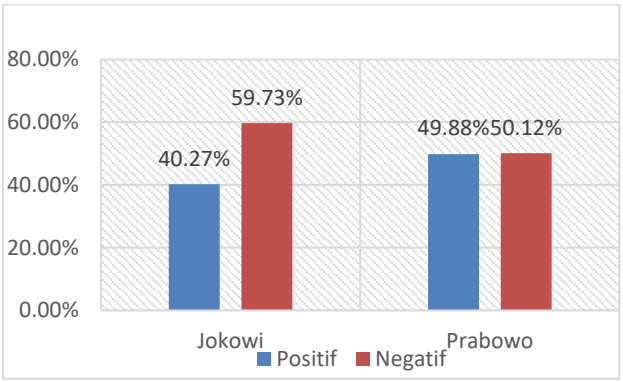
**Tahap Analisis**

Adapun tahapan-tahapan yang dilakukan dalam penelitian adalah

1) Mengumpulkan *tweet* secara berkala dengan proses *crawling* di R dan memberikan label positif atau negatif pada setiap *tweet*, dan menghapus *tweet* yang tidak sesuai dengan konteks penelitian.
2) Melakukan *preprocessing* untuk membersihkan data dan memberikan bobot untuk setiap kata dengan TF.
3) Melakukan klasifikasi menggunakan metode SVM dengan fungsi *Kernel* dan mengealuasi model yang diperoleh untuk melihat akurasi dihasilkan metode SVM.
4) Membuat visualisasi *wordcloud* dan mengintrepetasikan hasil untuk melihat topik apa saja yang ada pada sentiment positif maupun negatif untuk masing-masing calon presiden.

**HASIL DAN PEMBAHASAN**

Data yang berhasil di *crawling* dari media sosial Twitter berkaitan dengan kata kunci tersebut adalah sebanyak 20.000 *tweet* dimana 10.000 tweet berkaitan dengan kata kunci “Jokowi” dan 10.000 *tweet* berkaitan dengan kata kunci “Prabowo”. Gambaran umum data dapat dilihat dari Gambar 2 berikut.




**Gambar 2. Grafik Proporsi Sentimen**

Gambar 2 terlihat proporsi sentiment positif yang diperoleh Jokowi lebih sedikit dari sentimen negatif. Pelabelan dilakukan pada 10.000 *tweet* yang berkaitan dengan kata kunci "Jokowi", diperoleh data sebanyak 2.774 *tweet* yang mengandung sentimen positif dan negatif. Sebanyak 1.117 *tweet* berlabel positif dan 1.657 *tweet* berlabel negatif. Proporsi sentiment positif yang diperoleh Prabowo lebih sedikit dari sentimen negatif seperti yang terlihat pada Gambar 2. Pelabelan dilakukan pada 10.000 *tweet* yang berkaitann dengan kata kunci "Prabowo", diperoleh data sebanyak 1.632 *tweet* yang mengandung sentimen positif dan negatif. Sebanyak 814 *tweet* berlabel positif dan 818 *tweet* berlabel negatif.

Seletah data yang telah dilabeli dilakukan *preprocessing*, maka dilakukan klasifikasi dengan algoritma SVM *non-linier* dengan fungsi *Kernel* di aplikasi R. Setelah klasifikasi maka dilakukan evaluasi model untuk melihat kebaikan model. Evaluasi model dilakukan dengan metode *cross validation* dengan proporsi 80% data *training* dan 20% data *testing.*

**Tabel 1. Akurasi Model SVM dengan Fungsi *Kernel***

| Data | *Training* | Testing |
|---|---|---|
| Jokowi | 0,8719 | 0,8682 |
| Prabowo | 0,8723 | 0,8637 |

Tabel 1 menujukkan akurasi data *testing* sebesar 0,8682 artinya persentase model SVM dengan fungsi *Kernel* melakukan prediksi dengan benar adalah sebesar 86,82% pada data Jokowi. Sedangkan pada data Prabowo, akurasi data *testing* sebesar 0,8637 artinya persentase model SVM dengan fungsi *Kernel* melakukan prediksi dengan benar adalah sebesar 86,37%. Model dikatakan cukup bagus karena tidak *underfitting* maupun *overfitting*. Hal tersebut terlihat dari nilai akurasi data *testing* dan *training* yang tidak terpaut jauh.


**Gambar 3. *Wordcloud* sentimen positif Joko Widodo**

Gambar 3 menampilkan visulalisasi output sentiment positif Jokowi, pada gambar tersebut terlihat kata yang paling sering muncul yaitu kata "bangun", "desa", "jalan", "dana", "ekonomi", "pemerataan",

"konektivitas" dan lain-lain. Berdasarkan kata-kata yang muncul pada *wordcloud* di atas, dapat disimpulkan sentimen positif masyarakat berkaitan dengan pembangunan infrastruktur di Indonesia.


**Gambar 4. *Wordcloud* sentimen negatif Joko Widodo**

Gambar 4 menampilkan visulalisasi output sentimen negatif Jokowi, pada gambar tersebut terlihat kata yang paling sering muncul yaitu kata "hutan", "bakar", "bohong", "data", "gambut", dan lain-lain. Berdasarkan kata-kata yang muncul pada *wordcloud* di atas, dapat kita simpulkan beberapa sentimen negatif masyarakat berkaitan dengan pernyatan Jokowi mengenai kebakaran hutan yang tidak sesuai data pada debat kedua.


**Gambar 5. *Wordcloud* Sentimen Positif Prabowo Subianto**

Gambar 5 menampilkan visulalisasi output sentiment positif Prabowo, pada gambar tersebut terlihat kata yang paling sering muncul yaitu kata "debat", "menang", "kualitas", "percaya", "rival", dan lain-lain. Berdasarkan kata-kata yang muncul pada *wordcloud* di atas, dapat kita simpulkan beberapa sentimen positif masyarakat berkaitan dengan elektabilitas Prabowo sebagai oposisi yang berkualitas.


**Gambar 6. *Wordcloud* sentimen negatif Prabowo**

Gambar 6 menampilkan visulalisasi output sentimen negatif Prabowo, pada gambar tersebut terlihat kata yang paling sering muncul yaitu kata "*unicorn*", "data", "luas", "hektar", "bingung",

"debat", dan lain-lain. Berdasarkan kata-kata yang muncul pada *wordcloud* di atas, dapat kita simpulkan beberapa sentimen negatif masyarakat berkaitan dengan pernyataan Prabowo mengenai data luas daerah dan *unicorn* yang tidak sesuai data pada debat kedua.

## SIMPULAN DAN SARAN

Berdasarkan pembahasan dapat dilihat bahwa algoritma klasifikasi SVM dengan fungsi *Kernel* dapat melakukan klasifikasi dengan akurasi sebesar 86,82% untuk *tweet* dengan kata kunci "Jokowi" dan 86,27% untuk untuk *tweet* dengan kata kunci "Prabowo". *Tweet* dengan kata kunci "Jokowi" mengenai pembangunan infrastruktur di Indonesia termasuk dalam sentimen positif, sedangkan *tweet* mengenai pernyatan Jokowi tentang kebakaran hutan yang tidak sesuai data pada debat kedua, termasuk dalam sentimen negatif. *Tweet* dengan kata kunci "Prabowo" mengenai elektabilitas Prabowo sebagai oposisi yang berkualitas termasuk dalam sentimen positif, sedangkan *tweet* mengenai pernyatan Prabowo tentang data luas daerah dan *unicorn* yang tidak sesuai data pada debat kedua, termasuk dalam sentiment negatif. Penelitian selanjutnya disarankan untuk melakukan perbandingan algoritma klasifikasi lainya seperti *Maximum Entropy* dan NBC.

## DAFTAR PUSTAKA


APJII. (2017). *Penetrasi & Perilaku Pengguna Internet Indonesia*. https://apjii.or.id/survei2017. Diakses pada 20 Desember 2018

BPS. (2018). *Proyeksi Penduduk Indonesia 20015-2045 Hasil SUPAS 2015*. Jakarta. Badan Pusat Statistik.

Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters 27.8, pp. 861–874.

Hadi, A.F., Bagus, B.C.W., dan Hasan, M. (2017). *Text Mining pada Media Sosial Twitter Studi Kasus: Masa Tenang Pilkada DKI 2017 Putaran 2*. Surabaya. Universitas Airlangga. Seminar Nasional Matematika dan Aplikasinya.

Howe, N. and Strauss, W.(2007). *The Next 20 Years: How Customer and Workforce Attitude Will Evolve*. Hardvard Business Review. July-August 2007

Kumar, A., and Sebastian, T.M. (2012). *Sentimen Analysis on Twitter*. Delhi. Delhi Techonlogical University. International Journal of Computer Science Issue, Vol.9 Issue 4. Pages 371-378.

Latifah, E.F. (2018). *Perbandingan Kinerja Machine Learning Berbasis Algoritma Support Vector Machine dan Naïve Bayes*. Yogyakarta. Universitas Islam Indonesia

Liu, B. (2012).*Sentimen Analysis and Opinion Mining*.Morgan & Claypool Publishers.

Nurhuda, Sihwi, Doewes. (2013). *Analisis Sentimen terhadap Calon Presiden Indonesia 2014 berdasakan Opini dari Twitter dengan Metode Naïve Bayes Classifier*. Jurnal ITSMART. Vol 2 No 2 Desember 2013. Halaman 35-42.

Pang, B. (2002). *Thumb Up? Sentiment Classification Using Machine Learning Techiques*. Prociding of the Conference on Empirical Method in Natural Language Processing. 70-86.

Witten, I.H. (2005). *"Text mining." in Practical handbook of internet computing*, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida.