Analisis Sentimen Menggunakan Support Vector Machine dan Maximum Entropy

Sentiment Analysis Using Support Vector Machine and Maximum Entropy Method

Ramadhan WP[1], Astri Novianty S.T.,M.T[2],Casi Setianingsih S.T.,M.T[3]
[1,2,3]Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Telkom University
Bandung, Indonesia

[1]ramaprakoso@students.telkomuniversity.ac.id, [2]astrinov@telkomuniversity.ac.id ,
[3]setiacasie@telkomuniversity.ac.id

Abstrak

Jumlah data mengalami pertumbahan yang sangat cepat dalam era sekarang ini. Data dapat berupa *text*, gambar, suara, dan video. Media sosial menjadi salah satu faktor pertumbuhan data, setiap orang berekspresi, beropini dan mengeluh di media sosial.

Dalam tugas akhir ini dilakukan analisis sentimen menggunakan dua metode yaitu *Support Vector Machine* dan *Maximum Entropy*. Langkah pertama adalah dilakukan penambangan data menggunakan twitter API dengan *keyword* masing-masing adalah nama calon dalam pilkada DKI. Setalah mengumpulkan data, dilakukan proses *preprocessing*, setelah proses *preprocessing* dilakukan pengambilan fitur pada setiap *tweet*, fitur yang didapatkan kemudian dikumpulkan menjadi sebuah list fitur. List fitur kemudian ditransformasikan menjadi *feature vector* dengan bentuk *binary* kemudian ditransformasikan menggunakan metode Tf-idf. Dataset terdiri dari 2 data yaitu *training* dan *testing*. *Training* diberikan label secara manual. Untuk pengujian performa algoritma digunakan metode *K-Fold Cross Validation*.

Hasil pengujian adalah akurasi yang diperoleh mencapai rata-rata 75% dengan komposisi data *training* dan data uji 90:10. *Kernel* yang paling optimal dalam pengujian adalah *kernel* linier. Perubahan jumlah *folding* tidak berpengaruh terhadap tingkat akurasi. Metode *Support Vector Machine* lebih baik digunakan daripada *Maximum Entropy*, untuk melakukan analisis sentimen.

Kata kunci : *machine learning*, *support vector machine*, *maximum entropy*, *text mining*, *opinion mining*

Abstrak

Data amount becomes rapidly increased in today's era. Data can be in form of text, picture, voice, and video. Social media is one factor of the data increase as everybody expresses, gives opinion, and even complains in social media.

Sentiment analysis is done in this final project in two methods. They are Support Vector Machine and Maximum Entropy. The first step is data collection using API twitter with each candidate names of *Pilkada DKI*. The collected data then becomes input for preprocessing step.The next step is extracting each tweet's feature to be listed. The list of features is transformed into feature vector in binary form and transformed again using Tf-idf method. Dataset consists of two kinds of data, training and testing. Training is labeled manually. K-Fold Cross Validation is used to test algorithm performance.

Based on the result of the test, accuracy obtained reaches 75% in average with composition of training data and testing data by 90:10. The most optimal Kernel used is the linear one. The changing of folding amount gives no impact to the accuracy level. Support Vector Machine method is better used than Maximum Entropy to do the sentiment analysis.

Keyword *: Support Vector Machine, Text Mining, Maximum Entropy*

1. Pendahuluan

Peran media sosial seperti *Facebook* dan *Twitter*, saat ini menjadi sebuah media untuk memberikan kritik terhadap kegiatan politik [1]. *Sentiment analysis* atau bisa yang disebut *opinion mining* adalah sebuah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi terhadap sebuah entitas seperti produk, jasa, organisasi, individu, masalah, peristiwa topik dan attribut mereka [2].

Banyak metode yang digunakan untuk melakukan analisis sentimen seperti *Naive Bayes*, *Maximum Entropy* dan *Support Vector Machine*, dan lain-lain [10]. Teknik *machine learning* menggunakan satu *set* pelatihan dan tes ditetapkan untuk klasifikasi [10]. Menurut Pang [4], metode menggunakan *machine learning* dapat memberikan hasil yang lebih baik, namun *klasifier supervised learning* membutuhkan banyak data latih yang telah diberikan label.

Tanpa diberikan data latih yang telah diberi label, *supervised learning* tidak dapat bekerja [5]. Pemilihan metode *Support Vector Machine* dipilih karena memiliki kemampuan generalisasi dalam mengklasifikasikan suatu *pattern* [11].

2. Dasar Teori

2.1. *Twitter*

Twitter adalah layanan sosial media dan mikroblogging daring yang memungkinkan pengguna memposting, mengirim dan membaca sebuah pesan berbasis teks hingga 140 karakter. *Twitter* didirikan oleh Jack Dorsey pada tahun 2006 dan menjadi salah satu layanan media sosial populer di dunia. *Twitter* mengalami pertumbuhan yang cukup pesat hingga bulan Januari 2013 jumlah pengguna mencapai angka 500 juta yang terdaftar di *twitter*, 200 juta diantaranya adalah pengguna aktif. Di Indonesia sendiri menurut Dick Costolo saat kunjungan di Indonesia 2014 pengguna *Twitter* di Indonesia mencapai 50 juta, berdasarkan data terakhir 2014 pengguna aktif *Twitter* 284 juta. Pada bulan Juni 2012 Jakarta menempati urutan pertama dengan jumlah postingan 10.6 juta *tweet* [3].

2.2. Analisis Sentimen

Analisis sentimen atau yang biasanya disebut dengan *opinion mining*, merupakan penentuan emosi dibalik serangkaian kata-kata,yang digunakan untuk memperoleh pemahaman tentang sikap, pendapat dan emosi yang diekspresikan secara online. Analisis sentimen sangat berguna dalam pemantauan media sosial karena memungkinkan kita untuk memperoleh gambaran tentang opini publik yang lebih luas dari suatu topik-topik tertentu.

2.3. Ekstraksi Fitur

Ekstrasi fitur adalah proses mentrasformasikan input data kedalam sebuah set fitur [8]. Fitur adalah objek dari sebuah pola yang kuantitasnya dapat diukur, pengklasifikasiannya berdasarkan masing-masing dari fitur-fitur tersebut. Kemampuan dari proses *machine learning* sangat tergantung pada fitur-fiturnya sehingga sangat penting memilih tujuan ekstraksi fitur [6].

2.4. *Support Vector Machine*

Support Vector Machine (SVM) merupakan salah satu metode klasifikasi dengan menggunakan *machine lear*ning (*supervised learning*) yang memprediksi kelas berdasarkan model atau pola dari hasil proses *training* [7]. Dengan melakukan *training* menggunakan data inputan dalam bentuk numerik dan pembobotan dengan *Tf-idf* akan didapatkan sebuah pola yang nantinya akan digunakan dalam proses pelabelan. Nilai atau pola yang dihasilkan dari Metode *Support Vector Machine* sebenarnya adalah sebuah garis pemisah yang disebut dengan *Hyperplane*, dimana pada kasus ini garis tersebut berperan dalam memisahkan *tweet* dengan sentimen positif (berlabel 1) dengan *tweet* yang memiliki sentimen negatif (berlabel 0). Berikut adalah rumus perhitungan *hytperplane* [7] :

$$f:w.x+b=0 \qquad (2\text{-}1)$$

Dimana :

w = parameter *hyperplane* yang dicari (garis yang tegak lurus antara garis *hyperplane* dan titik *support vector*)

x = data input SVM (x1 = index kata, x2= bobot kata)

b = parameter *hyperplane* yang dicari (nilai bias)

f = fungsi *Hyperplane*

2.5. *Maximum Entropy*

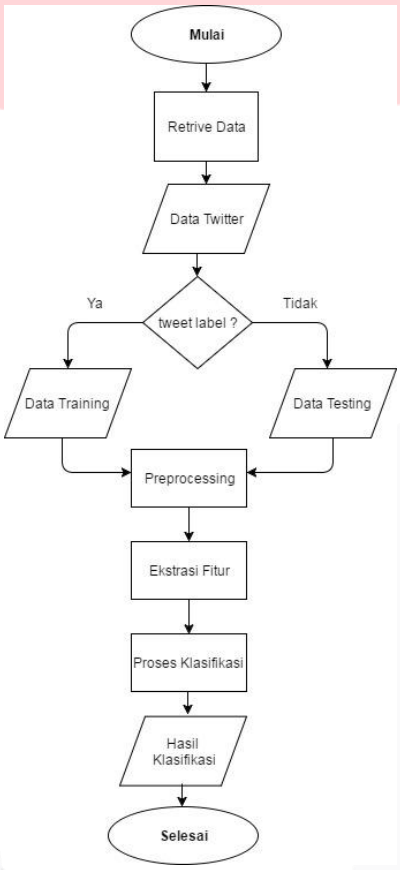
Merupakan metode klasifikasi berbasiskan probabilitas yang termasuk dalam kelas model eksponensial [19]. Prinsip dari *maximum entropy* mencari distribusi p(a|b) yang akan memberikan nilai *entropy* maksimum. Pada (MacKay,2003), *Maximum Entropy* didefiniskan sebagai rata-rata nilai informasi yang maksimum untuk suatu himpunan kejadian X dengan distribusi nilai probabilitas yang seragam [9]. Yang dimaksud dengan distribusi nilai probabilitas seragam adalah distribusi yang menggunakan faktor ketidakpastian yang minimum atau dapat disebut sebagai distribusi yang memakai asumsi sedemikian mungkin. Dengan menggunakan asumsi yang minimal, maksimal distribusi yang didapatkan merupakan distribusi yang paling mendekati kenyataan [9]. Pencarian distribusi probabilitas yang paling memberikan nilai *entropy* yang maksimum dilakan dengan tujuan distribusi probabilitas terbaik mendekati kenyataan [9].

2.6. Validasi dan Evaluasi

Validasi yang digunakan dalam penilitian ini adalah *k-fold cross validation*. Dalam *k-fold cross validation,* data awal dipartisi secara acak menjadi sebuah k *subset (fold),* yaiktu D_1,D_2,D_3,...,D_k, yang masing-masing berukuran sama. Proses *training* dan *testing* dilakukan sebanyak k kali [7].

3. Perancangan Sistem

Sistem yang dibangun pada penelitian ini adalah sistem yang dapat menganalisis sentimen yang terdapat pada media sosial *Twitter* mengenai pilkada DKI. Dalam penelitian ini analisis diambil berdasarkan sebuah *tweet*. Kumpulan dari *tweet* berguna sebagai data *training* dengan sebuah *label* dan kemudian akan dilakukan suatu pengujian dengan data uji. Hasil dari performansi algoritma *Support Vector Machine* dan *Maximum Entropy* mengenali *tweet* positif dan negatif menjadi fokus dari penelitian.



Gambar 3.1 *Flowchart* Proses Secara Umum

Berikut penjelasan dari gambaran proses :

1. Pengambilan data mentah dalam bentuk *tweet* menggunakan API dari *twitter* kemudian data disimpan dalam bentuk *csv*. Untuk data yang berlabel adalah data *training* dan data yang tidak berlabel adalah data *testing*.
2. Pada data *testing* ataupun data *training* dilakukan *preprocessing*. Proses dari *preprocessing* meliputi menghapus URL, tanda baca, penghapusan *stopwords*, merubah kata *slang* menjadi baku dan *stemming*.
3. Kemudian dilakukan proses ekstraksi fitur pada *tweet* yang sudah bersih hasil *preprocessing*. Proses ekstrasi fitur meliputi penglompokan kata dengan metode *Bag Of Words* dan pembobotan fitur dengan Tf-idf.
4. *Tweet* yang sudah berupa kumpulan fitur dalam *Bag of Words* dan sudah diberikan sebuah bobot menggunakan Tf-idf diklasifikasi menggunakan metode *Support Vector Machine* dan *Maximum Entropy*.
5. Proses klasifikasi menghasilkan sebuah nilai *recall*, presisi, akurasi dan hasil sentimen dari *tweet*.

4. Pengujian

4.1. *Dataset*

Dalam melakukan pengujian, dataset dibagi menjadi 2 yaitu data *training* dan data *testing*. Untuk kategorinya, data *training* dan *testing* berisi data campuran antara data dari pasangan a,b, dan c, digabungkan menjadi satu. Kemudian data digabungkan menjadi satu kemudian dibagi sesuai dengan skenario. Didalam sistem data uji diberikan label secara manual sedangkan data uji tidak diberikan label, sehingga sistem dapat memprediksi apakah nilai *tweet* uji positif atau negatif. Untuk menghitung performansi sistem menggunakan sistem *cross validation*, untuk menilai atau menvalidasi keakuratan sebuah model yang dibangun berdasarkan *dataset* tertentu.

Tabel 4.1 Rincian *Dataset* Untuk Pengujian Skenario

Data	Jumlah Data Training	Jumlah Data Testing
50:50	753 tweet	753 tweet
60:40	904 tweet	602 tweet
70:30	1054 tweet	452 tweet
80:20	1205 tweet	301 tweet
90:10	1356 tweet	151 tweet

Tabel 4.2 Jumlah Sentimen Data Training

Komposisi Data	Jumlah Data Training	Sentimen	
		Positif	Negatif
50:50	753 tweet	361	392
60:40	904 tweet	442	462
70:30	1054 tweet	522	532
80:20	1205 tweet	598	607
90:10	1356 tweet	682	672

4.2. Skenario Pengujian

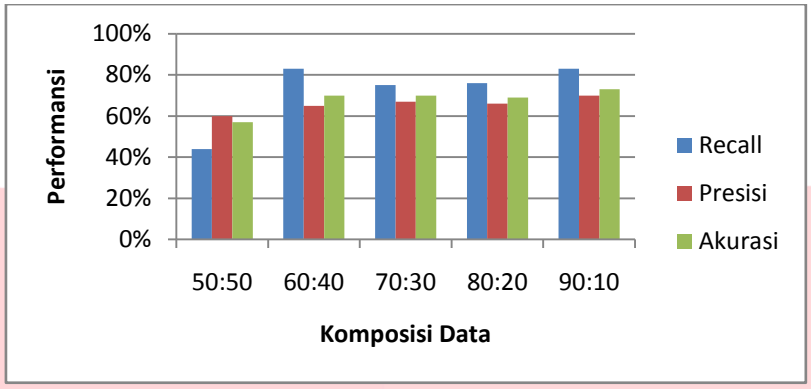
1. Pengujian pertama menggunakan perbandingan komposisi data antara data *training* dan *testing*.
2. Pengujian kedua untuk metode *Support Vector Machine* diberikan beberapa macam kernel seperti *linear*, *rbf* dan polinomial. Komposisi data yang digunakan adalah mengambil dari komposisi paling optimal dari skenario pertama. Akan dicari kernel dengan tingkat akurasi, presisi dan *recall* tertinggi.
3. Pengujian ketiga karena *cross validation* menggunakan k-fold, pengujian dibagi menjadi 5-Fold. Untuk metode *Support Vector Machine* data yang digunakan adalah komposisi paling optimal dari skenario pertama dan kernel digunakan *kernel* yang paling optimal dalam skenario kedua. Dari pengujian dicari *fold* ke berapa dengan nilai akurasi, presisi dan *recall* tertinggi.

4.3. Analisa Hasil Pengujian

Pada bagian ini akan dipaparkan hasil dari pengujian dan analisis terhadap hasil yang telah didapatkan dari proses pengujian.

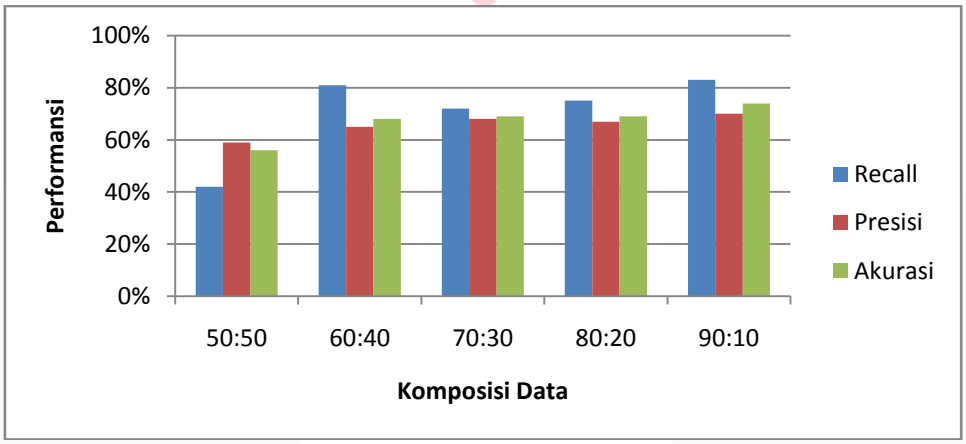
4.3.1. Analisis Terhadap Komposisi Data Latih dan Data Uji

Dalam pengujian skenario pertama dilakukan pengujian dengan perbadingan komposisi data *training* dan data *testing*. Dari perbandingan komposisi data akan dilihat yang paling optimal.



Gambar 3.2 Grafik Hasil Pengujian *SVM* dengan Perbandingan Komposisi Data

Dalam pengujian, algoritma *SVM* dengan perbandingan komposisi 90:10 memiliki hasil yang paling optimal mencapai 76% nilai akurasi. Dari pengujian dapat dilihat semakin banyak jumlah data training, hasil perhitungan semakin tinggi.

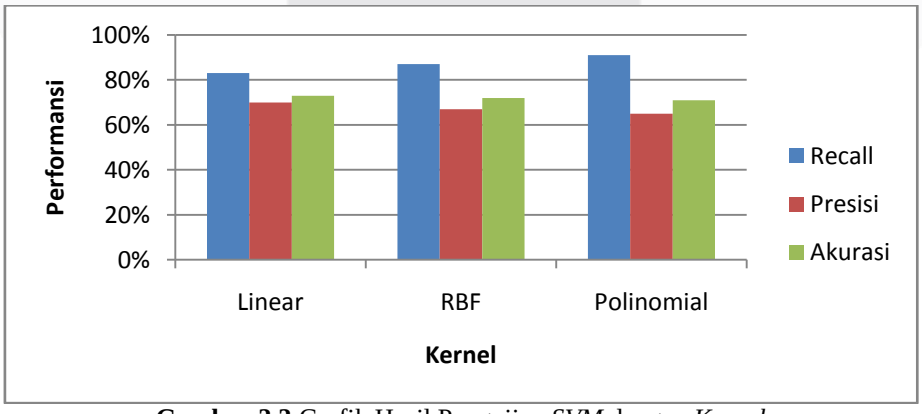


Gambar 3.3 Grafik Hasil Pengujian *ME* dengan Perbandingan Komposisi Data

Pengujian *Maximum Entropy* dengan menggunakan perbandingan komposisi, diperoleh hasil sama dengan pengujian pada metode *Support Vector Machine*. Perbandingan 90:10 memiliki nilai perhitungan paling optimal, dengan nilai 74%.

4.3.2. Analisis *Kernel Support Vector Machine*

Dalam analisis kedua, digunakan komposisi data yang paling optimal dari pengujian pertama. Masing-masing akan dibandingan akurasi, presisi dan *recall* berdasarkan *kernel* dari metode *Support Vector Machine*. Untuk komposisi data digunakan skema 90:10.

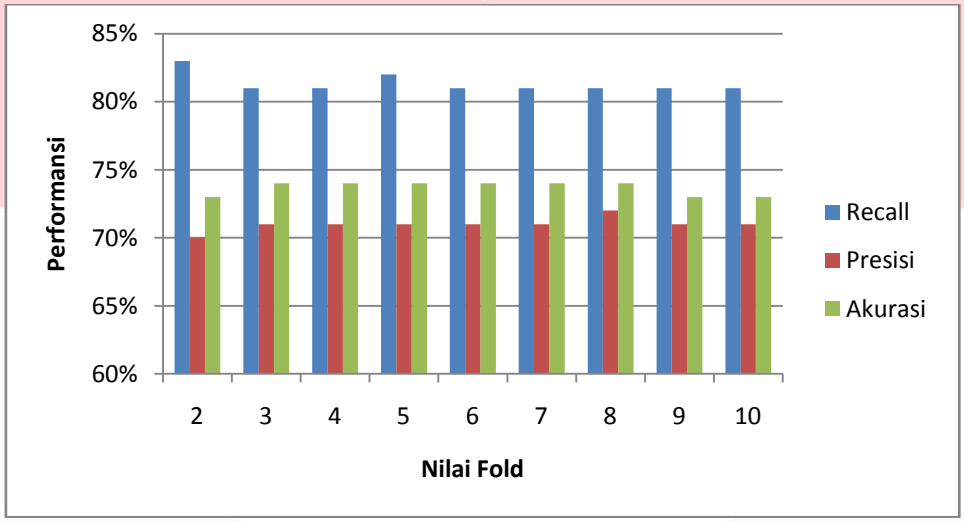


Gambar 3.3 Grafik Hasil Pengujian *SVM* dengan *Kernel*

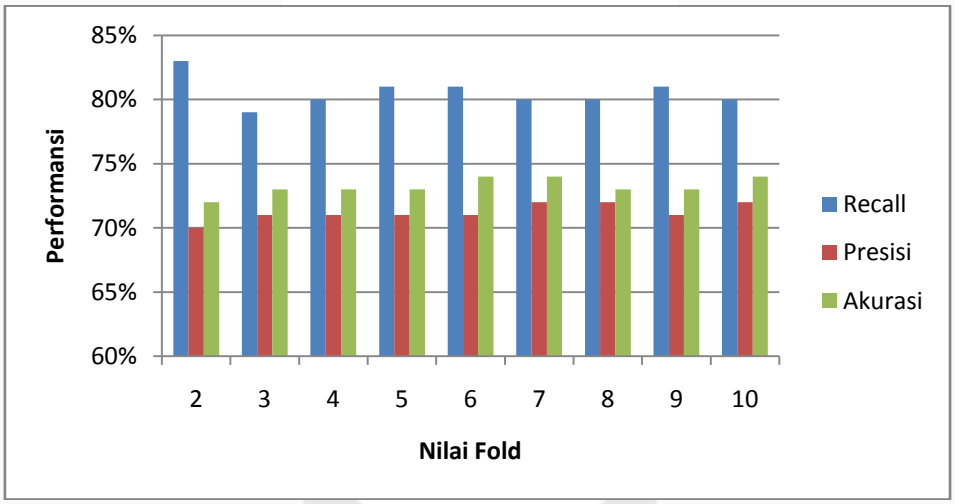
Dari *kernel* yang diujikan, *kernel* linier memiliki akurasi lebih baik daripada kedua. Dalam studi-studi analisis teks *kernel* linier lebih sering digunakan daripada *kernel* yang lain.

4.3.3. Analisis Berdasarkan Pembagian Nilai *K-Fold*

Dalam analisis ketiga, dilakukan perubahan nilai *fold*. Nilai *fold* dirubah dari 2-5 *fold*. Untuk analisis pada metode *Support Vector Machine*, digunakan *kernel linear* dan komposisi data dengan skema 90:10. Untuk *Maximum Entropy* hanya digunakan skema data 90:10 karena paling optimal nilai akurasinya.



Gambar 3.4 Grafik Hasil Pengujian *SVM* Terhadap Jumlah *Fold*



Gambar 3.5 Grafik Hasil Pengujian *ME* Terhadap Jumlah *Fold*

Dari hasil pengujian, semakin besar nilai *fold* mempengaruhi perolehan akurasi dari masing-masing metode. *Maximum Entopry* dapat mencapai akurasi tertingginya sampai dengan 74%.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil pengujian dan analisa yang telah dilakukan sebelumnya , maka diambil beberapa kesimpulan sebagai berikut :

1. Dari kedua percobaan, parameter yang paling berpengaruh dalam mendapatkan hasil yang baik adalah komposisi data *training* dan *testing*. Semakin banyak data *training* dibandingkan dengan jumlah data *testing*, hasil dari akurasi yang diperoleh semakin tinggi.
2. Dalam percobaan *kernel Support Vector Machine* yang paling optimal adalah menggunakan *kernel linear*. Dengan rata-rata nilai akurasi, presisi dan *recall* mencapai 75%.

3. Jumlah *folding* tidak mempengaruhi performansi dari masing-masing metode.
4. Dari kedua metode *Support Vector Machine* lebih unggul karena memiliki metode *kernel* untuk menyelesaikan masalah *linier* ataupun *non-linier* dan hasil mencapai rata-rata 75%.

5.2. Saran

Saran untuk pengembangan lebih lanjut dari Tugas Akhir ini sebagai berikut :

1. Menambahkan teori penanganan negasi untuk lebih memvalidasi *tweet* dan mencegah keambiguan dalam *tweet.*
2. Manambahkan teori POS-*Tagging* untuk mencegah ambigutias dari setia kata yang menjadi fitur.
3. Menambahkan kamus dalam kbba untuk menangani kata-kata yang *slang* untuk menambahkan frekuensi fitur kata.
4. Membandingkan dengan metode *unsupervised*, dikarena metode *supervised* membutuhkan label untuk data *training* dan harus ahli yang melakukan validasi *dataset*. Metode *unsupervised* diharapkan dapat secara mandiri belajar dikarenakan data yang ada dalam media sosial sangatlah banyak.

DAFTAR PUSTAKA

[1] Ibrahim Mochamad Ibrahim, Omar Abdillah, Alfan F.Wicaksono, Mirna Adriani.***Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Result in A TwitterNation***.1

[2] Liu, B.***Sentiment Analysis and Opinion Mining***.Morgan & Claypool Publishers. 2

[3] Semiocast.2012.***Twitter reachs half a bilion accounts More than 140 million in the U.S.***.4 Maret 2016<http://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US>; 3

[4] Pang, B., Lee, L., & Vithyanathan, S. (2002). ***Thumbs Up ? SentimentClassification Using Machine Learning Techniques***. Proceedings of The ACL-02 conference on Empirical methods in natural language processing (pp. 79-86). Stroudsburg: Association for ComputationalLinguistic 4

[5] McCallum, A., Freitag, D., dan Pereira, F.2000. ***Maximum Entropy Markov Models for Information Extraction and Segmentation***. Proc. ICML 2000, pp. 591–598, Stanford, California.

[6] Zainuddin, Nurulhuda dan Ari Selamat.2014. **Sentimen Analysis Using Support Vector Machine".*International Conference on Computer, Communication, and Control Technology***. 12

[7] Novantirani, Anita. 2013. Analisis Sentimen Pada *Twitter* Mengenai Penggunaan Transportasi Umum Darat Dengan Kota dengan Metode Support Vector Machine. Bandung : Tugas Akhir Universitas Telkom. 17

[8] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines Information Retrieval in Practice*. Addison Wesley, 2009. 18

[9] Anggreini, Dyta. 2008. Klasifikasi Topik Menggunakan Metode Naive Bayes dan Maximum Entropy Pada Artikel Media Massa dan Abstrak Tulisan . Depok : Skripsi Universitas Indonesia. 19

[10] Ley, Z., Riddhiman, G., Mohamed, D., Meichun, H., & Bing, L. (2011). "Combining lexicon-based and learning-based methods for twitter sentiment analysis". *HP Laboratories*, Technical Report HPL-2011, 89. 27

[11] Nugroho, A.S., Witarto, A.B. dan Handoko, D. 2003, ***Application of Support Vector Machine in Bioinformatics***. Proceeding of Indonesian Scientific Meeting in Central Japan, Gifu-Japan, December 20, 2003.