KLASIFIKASI ALGORITMA TF DAN NEUTRAL NETWORK DALAM SENTIMEN ANALISIS

Amril Mutoi Siregar
Universitas Buana Perjuangan Karawang
amrilmutoi@ubpkarawang.ac.id

ABSTRACT

Nowadays social media has become one of the tools to express idea or opinion. They are more active expressing it on social media instead of speaking directly. Twitter is the most popular among them to express idea, also share news, picture, music and etc. Twitter users are increasing significantly each year as the result the information grows in same way. Due too much information flow, people get difficulties to make sure or clarify the news. For example, Looking for the information about a figure who will participate in a Pilkada. There are many researchers analyze subjectively and haven't given the maximum result yet. This research is trying to clarify information and divided them into positive, negative and neutral information. It is using TF algorithm and Neutral Network as the tools. The dataset is taken from a figure' twitter which is participate in Pilkada. And the result shows that accuracy 66.92%, positive precision 67.80%, negative precision 64.29%, neutral precision 73.33%, and positive recall 80%, negative recall 70%, neutral recall 36.67%.

Keyword: *TF algorithm, neural network, tweeter, sentiment, text mining.*

ABSTRAK

Media sosial merupakan salah satu alat yang gunakan pengguna untuk mengekspresikan pendapat. Sekarang ini pengguna lebih aktif memberikan pendapat dengan media sosial daripada menyampaikan secara langsung. Twitter adalah salah satu media paling popular untuk menyampaikan baik pendapat, berita, gambar, music dan lain lain. Setiap tahunnya pengguna Twitter mengalami peningkatan sehingga informasi yang ada juga semakin meningkat. Informasi yang semakin meningkat menyebabkan pengguna yang ingin mencari suatu informasi tertentu mengalami kesulitan. Untuk mengatasi masalah tersebut diperlukan klasifikasi informasi. Misalnya untuk menilai suatu tokoh politik yang ikut pilkada. Banyak peneliti melakukan analisa sentiment tapi hasilnya belum maksimal. Penelitian saat ini melakukan klasifikasi informasi berupa positif, negative dan netral, dengan menggunakan algoritma TF dan Neural Network. Dataset diambil dari twitter seorang tokoh dalam pilkada. Berdasarkan hasil pengujian yang telah dilakukan didapatkan hasil nilai accuracy 66.92%, precision positif 67.80%, precision negative 64.29%, precision netral 73.33%, dan recall positif 80%, recall negative 70%, recall netral 36.67%.

Kata kunci: *algoritma term frequency, Neural Network, Sentimen, Twitter, text mining.*

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Saat ini internet telah memberikan dampak pada kehidupan manuasi, dampaknya berupa penyebaran informasi yang semakin berkembang pesat bahkan percepatannya yang luar biasa. Kategori informasi hampir tersedia di jaringan internet. Informarsi tersebut beruba *website, microblog, blog,* sosial media. Contoh media yang popular untuk bertukar informasi adalah Twitter. *Twitter* adalah salah satu jejaring sosial yang memungkinkan pengguna dapat berbagi informasi dalam berupa teks dengan jumlah karakter yang dibatasi (Phuvipadawat, et

al, 2010). Selain jejaring sosial banyak pengguna memanfaatkan sebagai sarana iklan, berita dan sebagainya. Pengguna yang mendapat informasi disebut dengan *follower* (Kwak, at all, 2010). Banyaknya informasi yang tersedia pada *Twitter* dapat dianalisa dengan melakukan kategorisasi informasi berdasarkan ketegorisasi tertentu (Weissbock, et al, 2013). Algoritma untuk klasifikasi dan kategorisasi saat banyak dilakukan peneliti adalah *Algoritma Klasifikasi Naïve Bayes*, KNN, *Neural Network* (NN), *Support Vector Machine* (SVM), *Decision Tree, Linear Least Squares Fit* (LLFS), dan lain lain (Yuan, L, 2010).

Dengan berkembangnya pengguna media sosial maka berkembang pula, data dihasilkan yang berbentuk teks, saat ini telah memcapai jumlah data yang sangat besar. *Twitter* salah satu menjadi sumber data teks yang potensial mengingat, pengguna dan *follower*nya yang sangat banyak. Data yang dihasilkan memiliki karakteristik yang tidak terstruktur dan banyak *noise* yang terkandung didalamnya contohnya tag, @, #, %, $ yang menjadikan lebih rumit untuk dipahami makna dari tulisan dan maksudnya. Untuk mengeksplor data digunakan Teknik *text mining* yang memiliki peranan penting dibidang ilmu data *mining*, dengan mengaplikasikan aplikasi aplikasi proses dalam *text mining*. Maka diperoleh pola pola data, tren dan ekstraksi dari *knowledge* yang potensial dari suatu data teks (E. Junianto 2014). Diantara proses yang dilakukan dalam *text mining* adalah klasifikasi teks. Klasifikasi teks dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu. Salah satu masalah berkaitan dengan *text classification* yang ditemukan pada *twitter* orang terhadap pemilihan suatu tokoh pilkada, dan ditemukan banyak jenis twiiter seperti *positive, negative* dan netral. Hal dapat tersebut harapkan dapat membantu dalam memberikan informasi kepada orang tingkat sentiment orang terhadap tokoh atau topik yang dibicarakan.

1.2 Penelitian Terdahulu

Penelitian ini mempunyai keterkaitan beberapa peneliti sebelumnya dalam klasifikasi Analisa sentimen, berupa data yang tidak terstruktur menyebabkan banyak atribut yang tidak relevan. Jika semua attribut digunakan akan mengurangi *performance* klasifikasi (Wang, et al., 2013). Penelitian sebelumnya melakukan klasifikasi Positif, negative dan netral (Chandani, V et al. 2015).

1.3 Tinjauan Pustaka

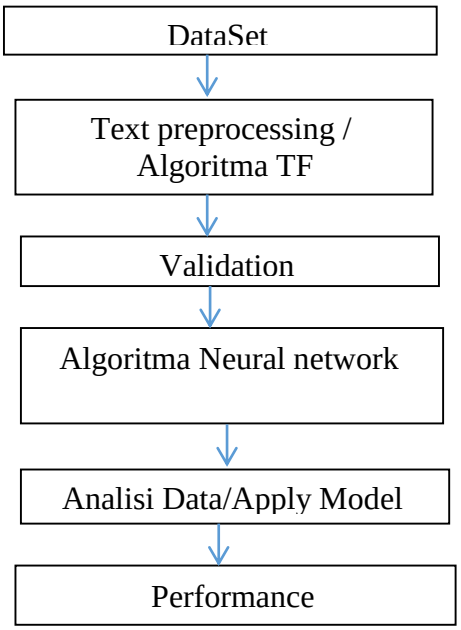
Teks mining merupakan metode penambangan dilakukan oleh komputer untuk mendapatkan sesuatu yang baru, sesuatu tidak diketahui sebelumnya atau menemukan suatu informasi yang tersirat secara implisit yang berasal dari informasi yang diekstrak secara otomatis dari sumber-sumber data teks yang berbeda-beda (Feldman, et al, 2006). *Teks mining* digunakan untuk analisis informasi pengambilan keputusan yang berupa teks, *image* dan lainya.

2. METODOLOGI

Untuk Penelitian ini di usulkan untuk mencari nilai akurasi yang paling maksimal dengan ekstraksi fitur *Term frequency (TF*) dengan algoritma *Neural Networks (NN*). Teks proses berguna untuk mempersiapkan dokumen dalam teks yang tidak terstruktur menjadi terstruktur, setelah terstruktur dilanjutkan tahapan selanjutnya. Tahapan *text processing* dalam *text mining* adalah sebagai berikut:

1. Tokenisasi adalah proses pemisahan kata. Penggalan kata kata disebut dengan token atau *term* (kata), (Manning, et al).
2. *Transform cases* adalah proses untuk merubah bentuk kata-kata, pada proses ini karakter dijadikan menjadi huruf kecil atau *lower case* semua.
3. *Filter Token* adalah proses pengambilan kata-kata yang dianggap penting dari hasil tokenisasi (Langgeni et al., 2010).

4. *Stem/steming* adalah proses perubahan bentuk dari kata-kata menjadi kata dasa. Perubahan bentuk kata dasar akan disesuaikan dengan struktur Bahasa yang digunakan proses *stemming* (Langgeni et al, 2010).
5. *Stopwords* adalah tahapan untuk menghilangkan kata kata yang sering muncul tapi tidak terlalu berpengaruh terhadap ekstraksi Analisa *sentiment.* Seperti kata petunjuk waktu, tanya dan lain lain (Langgeni et al, 2010).



Gambar 1 Metode Penelitian yang digunakan

2.1 Dataset

Penelitian ini menggunakan dataset yang *download* dari *twitter* merupakan situs pertemanan bersifat *online*. Data yang diambil berhubungan dengan *tweet* orang terhadap tokoh pada saat menjelang Pilkada, *sentiment* yang diklasifikasikan adalah positif, negative dan netral. Fitur yang digunakan terdiri dari *sentiment*, *tweet*. Dataset yang dikumpulkan sebanyak 130 *tweet* dari berbagai akun.

Tabel 1 Contoh Dataset yang digunakan

Tweet	Sentimen
persiapan tim inti besok 25.8 relawan perempuan @basuki_btp &djarot 1putaran menang.amin (sorry tdpamit) http…	Positif
@kevin_gitara: #Salam2Jari juga pak @basuki_btp ✌□ https://t.co/KC3em03orM	Positif
@KholikTuti: Saya pernah ke rumah nya pak @basuki_btp di Pantai Mutiara-Pluit... Beliau orang nya ramah tdk seperti yg di beritakan yg…	Positif
@ladyttiq: @triwul82 @basuki_btp Hanya orang2 cerdas & yg mau KERJA, yg mampu menjadikan cuti mjd hal yg produktif & bermanfaat. #salam2…	Positif
#PojokSatu Disebut Air Mata Buaya, Ini Kata #Ahok https://t.co/SnulSbECdH #nasional #kasusahok #sidangahok	Negatif
* Felix Siauw : Orang Yang Biasa Berkata Kasar Dan Kotor, Linangan Air Mata (Ahok) Tak Ada Harganya https://t.co/E3aMzfsNWV	Negatif
* PP Muhammadiyah: Eksepsi Ahok Tidak Berdasar Hukum https://t.co/VmFUUXcMc5 https://t.co/vi24N6R2Vu	Negatif

@hayked: Ini timses paslon no 1, nge tagline jg mulut utk kesatuan.seolah2 mau nyalahin ahok. Tp lupa, Suriah, Libya, Mesir dll ga ada a…	Netral
@kompascom: Ahok yang Berwajah Sedih Dipeluk Kakak Angkatnya Seusai Sidang. https://t.co/qmELapRHOu	Netral
@liputan6dotcom: Kesaksian Warga Saat Ahok Bicara Surat Al Maidah di Pulau Seribu https://t.co/pUqRi2mubG https://t.co/zyl3cjMjKK	Netral

2.2 *Text Preprosessing*

Text preprocessing merupakan tahap awal dalam text mining untuk merubah data yang tidak terstruktur menjadi data terstruktur (Perdana, et al, 2013). Proses ini dilakukan dengan tahapan tokenisasi, *transform caces*, *filter tokenisasi*, *stopwords* dan *stemming.*

2.3 *Term Frequency* (TF)

Dalam penelitian ini menggunakan pembobotan untuk mendapatkan bobot setiap kata pada setiap dokumen untuk menentukan sentiment dari isi *tweet.* Adapun algoritma yang digunakan dengan persamaan berikut.

$$Tf_{i,j}\,\frac{Tf_{i,j}}{\max(Tf_{i,j})}$$

Keterangannya:

$Tf_{i,j}$ = Jumlah kata dalam dokumen

$Max\ (Tf_{i,j})$ = Jumlah semua kata dalam dokumen

2.4 *Validation*

Validation digunakan untuk memperkirakan akurasi sebuah model yang akan tampil, Operator Validasi memiliki dua subproses: sub proses pelatihan dan subproses Pengujian. Sub proses pelatihan digunakan untuk melatih model. Model yang dilatih kemudian diterapkan didalam sub proses pengujian. Kinerja model diukur selama tahap pengujian. *Input Example Set* dipartisi menjadi subset k dengan ukuran sama. Dari subset k, satu subset dipertahankan sebagai kumpulan data uji (yaitu masukan dari subproses pengujian). Sisa k - 1 sisanya digunakan sebagai kumpulan data pelatihan (yaitu masukan subproses pelatihan). Proses validasi silang kemudian diulang k kali, dengan masing-masing subset k digunakan sama sebagai data uji. Hasil k dari iterasi k adalah rata-rata (atau gabungan lainnya) untuk menghasilkan estimasi tunggal. Nilai k dapat disesuaikan dengan menggunakan jumlah parameter lipatan. Evaluasi kinerja model pada set uji independen menghasilkan perkiraan kinerja yang baik pada kumpulan data yang tidak terlihat.

2.5 *Algoritma Neural Network*

Algoritma Neural Network salah satu algoritma atau model untuk klasifikasi baik untuk *Datamining*, *Text Mining*, *image* dan lain lain. *Neural network* cocok digunakan untuk *machine learning* untuk meniru neurofisiologi seperti otak manusia melalui perhitungan sederhana neuron dalam sistem yang saling berhubungan. Dengan nilai bobot jaringan secara random (angka -0.1 sampai dengan 1.0). Persamaan input untuk titik berdasarkan nilai input dan bobot jaringan pada data training, untuk menghitung persamaan input pada training adalah sebagai berikut :

$$Input_{\ j} = \sum_{i=1}^{n} O_i W_{ij} + \Theta_j$$

Keterangan :

O_i = sebagai Output simpul i dari layer sebelumnya

Wij =sebagai Bobot relasi dari simpul pada layer sebelumya ke simpul j

Θj = sebagai bias

Berdasarkan input dari langkah 2, selanjutnya membangkitkan output untuk simpul menggunakan fungsi aktifasi sigmoid :

$$Output = \frac{1}{1+e^{-Input}}$$

Untuk menghitung nilai error antara nilai yang diprediksi dengan nilai yang sebenarnya mengunakan rumus sebagai berikut:

Error j = Output j · (1-Output j) · (Target j – Output j)

Keterangan:
Output j = Sebagai Output actual dari simpul j
Target j = sebagai nilai target yang sudah diketahui pada data *training.*

Setelah nilai Error dihitung, selanjutnya dibalik *layer* sebelumnya *(back propagation*). Untuk menghiutng nilai Error pada *hiden layer*, menggunakan rumus sebagai berikut:

$$Error\ j = \text{Output j}\ (1 - \text{Output j}) \sum_{n=1}^{n} Error_k\ W_{jk}$$

Keterangan:
Output j = Sebagai Output actual dari simpul j
Error k = Sebagai Error simpul k
Wjk = Sebagai bobot relasi dari simpul j ke simpul k pada layer berikutnya.

Untuk Nilai Error yang dihasilkan dari langkah sebelumnya digunakan untuk memperbaharui bobot relasi menggunakan rumus sebagai berikut:

Wij = Wij + 1. Error j. Output I

Keterangan:
Wij = Sebagai bobot relasi daru unit I pada layer sebelumnya ke unit j
1 = Sebagai *Learning rate* (Konstanta, Nilai antara 0 – 1)
Error j = Sebagai Error pada output *layer* simpul j
Output i = Sebagai *Output* dari simpul i

2.6 *Apply Model*

Apply Model pertama dilatih pada *Example Set* oleh Operator lain, yang sering merupakan algoritma pembelajaran. Model ini bisa diaplikasikan pada *Example Set* yang lain. Biasanya, tujuannya adalah untuk mendapatkan prediksi pada data yang tidak terlihat atau untuk mentransformasikan data dengan menerapkan model *preprocessing. Example Set* dimana model diterapkan, harus kompatibel dengan Atribut model. Ini berarti, bahwa *Example Set* memiliki jumlah, urutan, jenis, dan peran Atribut yang sama seperti *Example Set* yang digunakan untuk menghasilkan model.

2.7 Evaluasi *Performance*

Untuk mengevaluasi performa dalam penelitian ini dengan menampilkan tingkat akurasi dalam *sentiment,* evaluasi dengan menggunakan *confusion matrix* adalah menghasilkan nilai *accuracy*, *precision*, dan *recall* untuk ke tiga kelas positif, negatiif dan netral. *Accuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han and Kamber, 2006). Sedangkan *precision* dan *confidence* adalah proposrsi masalah diprediksi atau klasifikasi yang positif yang sebenarnya. Dan *recall* atau *sensitifity* adalah proporsi kasus positif sebenarnya yang diprediksi positif secara benar (Power, 2011).

Tabel 2 Model *Confusion Matrix*

Klasifikasi	Diklasifikasikan sebagai	
	+	-
+	True Positives	False Negatives
-	False Positives	True Negatives
	True Netral	False Netral

Untuk rincian perhitungan positives, negatives dan netral menggunakan rumus sebagai berikut: (Han and Kamber, 2006)

$$\text{Sensitifity} = \frac{t\_pos}{pos}$$

$$\text{Specifity} = \frac{t\_neg}{neg}$$

$$\text{Specifity} = \frac{t\_netral}{netral}$$

$$\text{Precision Positives} = \frac{t\_pos}{t\_pos+f\_pos}$$

$$\text{Precision Negatives} = \frac{t\_Neg}{t\_Neg+f\_Neg}$$

$$\text{Precision Netral} = \frac{t\_netral}{t\_netral+f\_netral}$$

$$\text{accuracy} = \frac{posistif}{(pos+neg+Netral)} + \frac{negatif}{(pos+neg+Netral)} + \frac{netral}{(pos+neg+Netral)}$$

Keterangan:

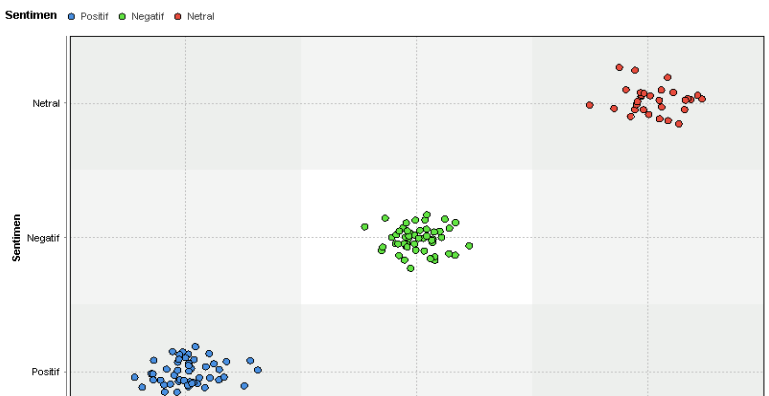
t_pos = Jumlah *true positives*
t_neg = *Jumlah true negatives*
t_Netral = *Jumlah true Netral*
Pos = *Jumlah record positive*
neg = *Jumlah record negative*
Netral = *Jumlah record netral*
f_pos = *Jumlah false positives*
f_neg = *Jumlah false Negatives*
f_netral = *Jumlah false Netrals*

3. ANALISA DAN PERANCANGAN SISTEM

Dalam penelitian ini dilakukan menggunakan computer dengan spesifikasi Laptop dengan CPU intel core i5, RAM 4 Gb, Dengan Sistem Operasi *Microsoft Windows 10 Propessional 64 bit*, dan Aplikasi *Rapidminer studio 7.3.*

Tabel 3 Hasil Accuracy, Precision dan Recall
Accuracy: 66.92% +/- 5.76% (mikro: 66.92%)

	True Positif	True Negatif	True Netral	Class Precision
Pred. Positif	40	13	6	67.80%
Pred. Negatif	7	36	13	64.29%
Pred. Netral	3	1	11	73.33%
Class recall	80.00%	72.00%	36.67%	



Gambar 2. Hasil Klasifikasi sentimen

Confussion matrix dalam penelitian ini adalah menggunakan table matriks, dataset terdiri dari tiga kelas yaitu positif, negative dan Netral. Setelah pengujian diperoleh hasil dimasukan berupa *confusion matrix*, untuk menghitung nilai – nilai performance klasifikasi yaitu *sensitivity (recall), Specifity, precision,* dan *accuracy*. Performance accuracy penelitian diatas 66.92%, rata rata *precision* adalah *precision positif + precision Negative + precision netral* dibagi 3 *class* yaitu 68.5%, rata *recall* adalah *recall positif + recall negatif + recall netral* dibagi 3 *class* yaitu 62.89% jadi hasil penelitian kurang maksimal. Permasalahan yang ditemukan adalah pada tahap *preprocessing* yang tidak maksimal, *wordnet* dalam bahasa Indonesia belum tersedia. Dan karakteristik bahasa Indonesia sangat unik karena terdapat kata kata yang berimbuhan seperti *prefix* dan *suffix* sehingga algoritma yang butuhkan berbagai macam untuk menghilangkan *prefix* dan *suffix* untuk mendapatkan kata dasar, untuk jadikan dasar penentuan *sentiment analysis*.

4. PENUTUP

4.1 Kesimpulan

Penelitian ini menampilkan hasil klasifikasi *textmining* yang diterapkan dalam *sentiment analysis* yaitu untuk mengklasifikasi sentimen-sentimen yang terbagi 3 macam yaitu positif, negatiF dan netral. Penelitian saat ini dapat disimpulkan sebagai beriku t:

1. Hasil penelitian ini dengan tingkat akurasi 66.92%.
2. Untuk penelitian dengan menggunakan pembobotan local TF dengan *neural network* kurang maksimal dalam sentiment analysis dalam bahasa Indonesia, dikarenakan belum tersedia *wordnet* sebagai standard bagi para peneliti.

4.2 Saran

Untuk penelitian berikutnya perlu menggunakan dataset yang lebih besar sebagai data *training* guna mendapatkan hasil pembelajaran lebih maksimal, dan perlu membuat *wordnet* sebagai dasar untuk menghitung tingkat *sentiment.*

DAFTAR PUSTAKA

Bramer, Max. (2007). *Principles of data mining*. Vol. 180. Springer.

Chandani V et al. (2015). *Komparasi algoritma klasifikasi machine learning dan feature selection pada analisa sentimen film.*

E. Junianto. (2014). *Penerapan Particle Swarm Optimization untuk seleksi fitur pada klasifikasi dokumen berita menggunakan naïve bayes classifier*. Program Pascasarjana Magister
ilmu computer STMIK Nusa Mandiri : Jakarta.

Felman, et al. (2006). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.* Cambridge University Press.

Kwak, et all. (2010). *What is Twitter a social Network or a News Media*. www *2010, pp.591-600.*

Langgeni, et al. (2010). *Clustering Artikel Berita Berbahasa Indonesia*. *SemnasIF, 1-10.*

Manning, et al. *Introduction to Information Retrieval.*

Perdana, et al. (2013). *Pengkategorian Pesan Singkat Berbahasa Indonesia Pada jejaring sosial twitter dengan klasifikasi naïve bayes. Pp, 1-12.*

Phuvipadawat, et al. (2010). *Breaking news detection and tracking in Twitter. Proceedings WI-IAT, pp 120-123.*

Wang, et al. (2013). *Sample cutting method for imbalanced text sentiment classification based on BRC Knowledge Based System, 37, 451-461*.

Weissbock, at al.(2013). *Using external information for classifying tweets. BRACIS, pp, 1-5.*

Yuan, L. (2010). *An Improved Naïve bayes text Classification Algorithm in Chinese Infromation Processing.* Jiaozuo.
` 1``````````````````````````````````````