REVIEW ATAS ANALISIS SENTIMEN PADA TWITTER SEBAGAI REPRESENTASI OPINI PUBLIK TERHADAP BAKAL CALON PEMIMPIN

Ahmad Afief Amrullah[1], Ahmad Tantoni[2], Nahrowi Hamdani[3], Rahmat Taufik R.L. Bau[4], Muhammad Rafiqudin Ahsan[5], Ema Utami[6]
Program Studi Magister Teknik Informatika, STMIK AMIKOM Yogyakarta
Jl Ringroad Utara, Condongcatur, Depok, Sleman, Yogyakarta, Indonesia 55283
Telp. (0224) 8311668
E-mail: aseptilena@gmail.com, ahmad.tantoni@students.amikom.ac.id, dani.hamdan29@gmail.com, rahmat.taufik.277@gmail.com, zzfiq@ymail.com, ema.u@amikom.ac.id

ABSTRAK

Mengamati perkembangan Opini Publik kini dapat dengan mudah dilakukan dengan mengamati Media Sosial.). Termasuk ke dalam Media Sosial terdapat Jejaring Sosial atau mungkin lebih akrab di telinga kita nama Facebook dan Twitter sebagai contoh jejaring sosial yang paling populer di Indonesia khususnya. Pada akhirnya, fenomena ini dapat dimanfaatkan bagi elite politik atau pada makalah ini kami batasi pada bakal calon gubernur, dengan memanfaatkan sistem informasi untuk mengelola data sentimen pada jejaring sosial sebagai representasi opini publik terhadap bakal calon sebagai bahan pertimbangan yang selanjutnya disebut Sistem Analisis Sentimen Publik. Namun demikian, masalah baru timbul saat sistem analisis sentimen tersebut dihadapkan pada kenyataan di Twitter. Terdapat beberapa masalah yang cukup urgen (tujuan) yaitu terkait dengan adanya duplikasi post, duplikat akun, BOT (program pengirim post otomatis) dan uraian preprocessing dalam berbagai versi subprosesnya. Oleh karena itu dibutuhkan suatu cara untuk menyaring data tweet yang memiliki salah satu bahkan ketiga masalah tersebut. Makalah ini membahas tentang metode yang dapat digunakan untuk memproses data tweet agar setidaknya ketiga masalah tersebut dapat diatasi. Sejauh penelusuran kami, salah satu proses dari sistem analisis sentimen publik yang terkait dengan hal ini adalah preprocessing.

Kata Kunci: opini publik, analisis sentimen, jejaring sosial, twitter, preprocessing

1. PENDAHULUAN

Di era berkembang pesatnya teknologi informasi seperti sekarang ini, perkembangan Opini Publik kini dapat dengan mudah dilakukan dengan mengamati Media Sosial. Media Sosial sendiri dapat diartikan sebagai bagian dari *New Media* (Internet), sebuah media untuk bersosialisasi satu sama lain, dilakukan secara *online* yang memungkinkan manusia untuk saling berinteraksi hampir tanpa dibatasi ruang dan waktu (Rustian dalam Ahsan: 17). Termasuk ke dalam Media Sosial terdapat Jejaring Sosial atau mungkin lebih akrab di telinga kita nama Facebook dan Twitter sebagai contoh jejaring Sosial yang paling populer di Indonesia khususnya.

Mengingat begitu populernya Facebook dan Twitter, sehingga dewasa ini dimungkinkan hampir setiap lapisan masyarakat Indonesia memiliki akun kedua Jejaring Sosial tersebut atau paling tidak akun Facebook yang dinilai lebih *user friendly*. Sehingga Opini Publik yang berkembang di masyarakat akan dengan mudahnya terangkat ke Jejaring Sosial. Secara

PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & CALL FOR PAPERS UNISBANK (SENDI_U) KE-2 Tahun 2016
Kajian Multi Disiplin Ilmu dalam Pengembangan IPTEKS untuk Mewujudkan Pembangunan Nasional Semesta Berencana (PNSB) sebagai Upaya Meningkatkan Daya Saing Global
ISBN: 978-979-3649-96-2

sederhana menilik Opini Publik yang berkembang di masyarakat dapat dilakukan dengan melihat langsung berbagai status serta komentar atas suatu fenomena (contohnya kebijakan elite politik) yang bertebaran di Jejaring Sosial. Sehingga tidak jarang Media Massa kini menampilkan berbagai komentar pengguna Jejaring Sosial atas suatu fenomena sebagai komponen dari berita yang diangkatnya.

Pada akhirnya, fenomena ini dapat dimanfaatkan bagi elite politik atau pada makalah ini kami batasi pada bakal calon pemimpin, dengan memanfaatkan sistem informasi untuk mengelola data sentimen pada jejaring sosial sebagai representasi opini publik terhadap bakal calon sebagai bahan pertimbangan yang akan diambil. Di antara media sosial yang ada, Twitter menjadi pilihan utama dalam analisis sentimen publik ini. Beberapa alasannya adalah (1) pesaingnya dengan jumlah pengguna yang jauh lebih banyak, Facebook memiliki kelemahan yang cukup signifikan pengaruhnya untuk dianalisis - 1) Facebook menggunakan Graph API. Ada beberapa keterbatasan dalam menggunakan facebook yang paling besar berpengaruh adalah keterbatasan pada konten publik. 2) kontrak dengan DataSift (perusahaan) untuk memperoleh data tingkatan topik anonim. Hal ini akan memungkinkan untuk mengintai obrolan yang bersifat pribadi, akan tetapi dengan cara ini kita tidak dapat benar-benar mengetahui siapa yang mengatakan apa atau seseorang yang mengatakan suatu hal.(quora.com, Jan 2013), (2) data dari Twitter telah terbukti dalam kontestasi politik PilPres 2014 yang memenangkan pasangan Jokowi-JK- "Berdasarkan data 10 hari terakhir di media sosial twitter, Jokowi-JK dibicararakan sebanyak 770.491, sedangkan Prabowo-Hatta sebanyak 709.294," ujar Direktur Katapedia Deddy Rahman di Jakarta, Jumat (4/7). Menurutnya, angka tersebut menunjukkan, Jokowi-JK meraih 52,07 persen. Sementara Prabowo Subianto-Hatta Rajasa sebanyak 47,93 persen (republika.co.id, 14 Juni 2014) dan (3) Twitter menjadi salah satu media utama dalam aliran data opini publik terkait politik.

Namun demikian, masalah baru timbul saat sistem analisis sentimen tersebut dihadapkan pada kenyataan di Twitter. Terdapat beberapa masalah yang cukup urgen yaitu terkait dengan adanya duplikasi post, duplikat akun dan BOT(program pengirim post otomatis). Danny menyampiakan bahwa analisa Sentigram(salah satu platform analisis sentimen yang sukses dalam prediksi PilPres 2014) sudah mempertimbangkan adanya akun-akun robot atau BOT pada 6 *platform* tersebut. Dalam mengambil data, Sentigram telah memverifikasi akun tak jelas tersebut. Sehingga data yang diambil lebih valid, pengguna akun pada platform lebih jelas. Ia memaparkan tim Sentigram telah mengetahui akun robot melalui pola dan aktivitas yang diposting pada platform. Pertama akun yang postingannya menduplikasi akun lain, jadwal posting, sampai pola perubahan sentimen akun. “Akun yang duplikat, copy paste nggak kita ambil. Kita juga lihat polanya, kapan bikin akun, pola jadwal posting, sampai bagaimana awalnya mereka positif dan kemudian sering jadi negatif. Pokoknya yang BOT nggak kita ambil" jelas Danny. (Danny, teknologi.news.viva.co.id, 2014). Oleh karena itu dibutuhkan suatu cara untuk menyaring data tweet yang memiliki salah satu bahkan ketiga masalah tersebut. Makalah ini membahas tentang metode yang dapat digunakan untuk memproses data *tweet* agar setidaknya ketiga masalah tersebut dapat diatasi. Sejauh penelusuran kami, salah satu proses dari sistem analisis sentimen publik yang terkait dengan hal ini adalah *preprocessing*. Oleh karena itu pembahasan pada makalah ini akan dibatasi pada *preprocessing*.

2. TUJUAN

Tujuan dari penelitian ini adalah untuk mengidentifikasi:

- keterkaitan masalah duplikasi *tweet* dan BOT yang penting bagi kasus dengan *preprocessing*
- uraian *preprocessing* dalam berbagai versi subprosesnya.

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

3. KAJIAN PUSTAKA

3.1 Twitter

Twitter adalah situs web dimiliki dan dioperasikan oleh Twitter, Inc., yang menawarkan jaringan sosial berupa *microblog*. Disebut *microblog* karena situs ini memungkinkan penggunanya mengirim dan membaca pesan blog seperti pada umumnya namun terbatas hanya sejumlah 140 karakter yang ditampilkan pada halaman profil pengguna. Twitter memiliki karakteristik dan format penulisan yang unik dengan simbol ataupun aturan khusus.Pesan dalam Twitter dikenal dengan sebutan *tweet*.(L. Zhang, 2011)

3.2 Analisis Sentimen

Analisis sentimen, yang disebut juga dengan opinion mining, merupakan salah satu cabang ilmu dari data mining yang bertujuan untuk menganalisis, memahami, mengolah, dan mengekstrak data tekstual yang berupa opini terhadap entitas seperti produk, servis, organisasi, individu, dan topik tertentu(B. Liu, 2012). Analisis ini digunakan untuk mendapatkan suatu informasi tertentu dari suatu kumpulan data yang ada. Analisis sentimen berfokus pada pengolahan opini yang mengandung polaritas, yaitu memiliki nilai sentimen positif ataupun negatif.(Novantirani, 2014)

3.3 Teori Elite Politik

Dalam Kamus Besar Bahasa Indonesia elite berarti orang-orang terbaik atau pilihan disuatu kelompok. Pengertian lain dari elite adalah kelompok kecil orang-orang terpandang atau berderajat tinggi seperti kaum bangsawan, cendikiawan dan lainnya. Sehingga dapat disimpulkan bahwa elite politik merupakan orang yang berkuasa dan terpilih dalam suatu kelompok. Teori elite dapat diklasifikasikan ke dalam dua status yaitu elite dalam struktur kekuasaan dan elite diluar struktural. Elite dalam struktur kekuasaan dapat diartikan sebagai anggota legislatif yang memiliki kemampuan dan kecakapan untuk mewakili masyarakat yang memilihnya. Sedangkan elite yang berada diluar struktural yaitu elite masyarakat yang mampu mempengaruhi masyarakat dalam mendukung ataupun menolak kebijakan elite berkuasa. Sehingga diperlukan menjalin komunkasi dengan elite masyarakat dalam upaya mendapatkan dukungan. Namun elite masyarakat dapat menggunakan popularitas yang dimilikinya untuk berkompetisi dengan elite dalam struktur (Marsland, 2009).

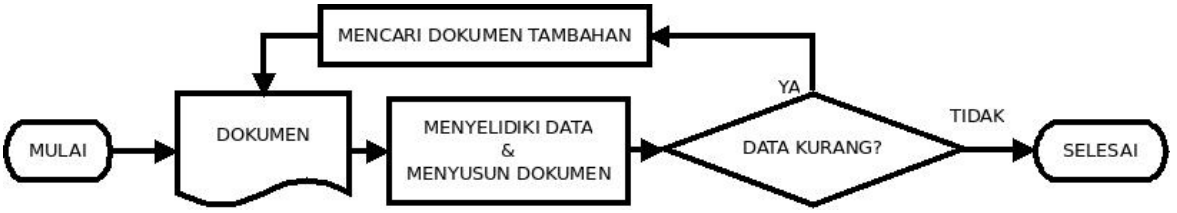
4. METODE

Metode yang digunakan dalam penelitian ini adalah kajian dokumentasi. Dokumen diartikan sebagai suatu catatan tertulis / gambar yang tersimpan tentang sesuatu yang sudah terjadi. Dokumen merupakan fakta dan data tersimpan dalam berbagai bahan yang berbentuk dokumentasi. Sebagian besar data yang tersedia adalah berbentuk surat-surat, laporan, peraturan, catatan harian, biografi, simbol, artefak, foto, sketsa dan data lainya yang tersimpan. Dokumen tak terbatas pada ruang dan waktu sehingga memberi peluang kepada peneliti untuk mengetahui hal-hal yang pernah terjadi untuk penguat data observasi dan wawancara dalam memeriksa keabsahan data, membuat interpretasi dan penarikan kesimpulan.(Djaelani, 2013)

Kajian dokumen dilakukan dengan cara menyelidiki data yang didapat dari dokumen, catatan, file, dan hal-hal lain yang sudah didokumentasikan. Metode ini relatif mudah dilaksanakan dan apabila ada kekeliruan mudah diganti karena sumber datanya tetap. Dengan membuat panduan / pedoman dokumentasi yang memuat garis-garis besar data yang akan dicari akan mempermudah

ISBN: 978-979-3649-96-2 *Unisbank Semarang, 28 Juli 2016*

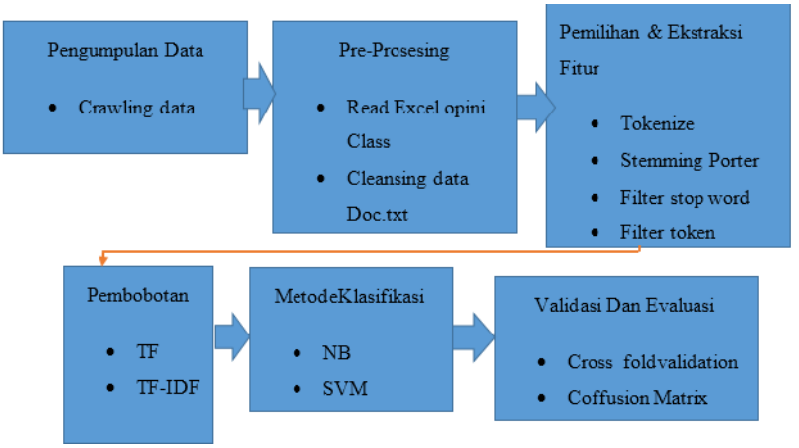
kerja di lapangan dalam melacak data dari dokumen satu ke dokumen berikutnya.(Djaelani, 2013)



Gambar 1. Metode Penelitian

5. HASIL DAN PEMBAHASAN

Terdapat banyak penelitian yang telah membahas tentang analisis sentimen serta bentuk sistemnya. Pemahaman terkait bentuk sistem ini menjadi penting dikarenakan posisi dari preprocessing yang menjadi fokus pada pembahasan makalah ini. Berikut beberapa ulasan mengenai bentuk sistem analisis sentimen yang terdapat pada beberapa penelitian. Dalam penelitian Hidayat(2015), wacana politik pada media masa online menggunakan algoritma support vector machine dan naive bayes.



Gambar 2. Desain Eksperimen oleh Hidayat(2015)

Hidayat(2015) memberikan keterangan bahwa sebelum melakukan komparasi/kombinasi dataset, dilakukan *text proccesing* terlebih dahulu *text proccesing* bertujuan untuk mempersiapkan dokumen teks yang tidak terstruktur menjadi data terstruktur yang siap digunakan untuk procesing data selanjutnya.

Adapun beberapa tahap-tahap dan implementasi *text proccesing* yang meliputi:

1) Tokenize merupakan proses untuk memisah-misahkan kata. Pemotongan kata tersebut yang sering disebut token term.
2) Filter

Token merupakan pengambilan/menyaring sebuah kata yang berkarakter misal di input nilai karakter 3 maka panjang dalam sebuah karakter kata akan di filter menjadi panjang 3 karakter sesuai panjang karakter yang diinputkan.

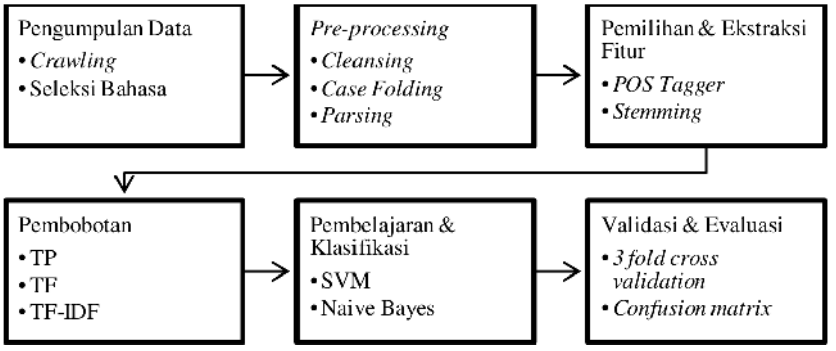
ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

3) Stemming yaitu proses menghilangkan kata-kata yang tidak penting dalam teks namun sering meuncul yang tidak memiliki pengaruh apapun dalam proses ekstraksi sentimen suatu *preview*. Misalnya kata yang termasuk kata penunjuk waktu dan kata tanya.

Sheela(2016) dalam penelitiannya, mengajukan sistem dengan modul-modul sebagai berikut. 1) *Data Streaming*, 2) *Preprocessing*, 3) *Sentiment Polarity Analysis*, 4) *Visualization*, 5) *Evaluation Metrics*

Sheela(2016) menjelaskan terkait *preprocessing*, dalam fase ini, *tweets* tersedia dalam bentuk data teks dan setiap baris berisi sebuah *tweet*. Awalnya kami membersihkan atau menghapus *retweets* karena ini akan memicu bias dalam proses klasifikasi. Kita perlu untuk menghapus tanda baca dan simbol lainnya yang tidak masuk akal karena dapat mengakibatkan inefisiensi dan dapat memengaruhi keakuratan dari proses secara keseluruhan.

Gambar 1 adalah metodologi yang digunakan dalam penelitian yang dilakukan oleh Nur & , 2011.



Gambar 3. Metode Penelitian Analisis Sentimen oleh Nur et al.(2011)

Tahapan yang dilakukan dari dokumen *pre-processing* adalah sebagai berikut(Nur et al., 2011):

1) *Cleansing*, yaitu proses membersihkan dokumen dari kata yang tidak diperlukan untuk mengurangi noise. Kata yang dihilangkan adalah karakter HTML, kata kunci, ikon emosi, *hashtag* (#), *username* (@*username*), url (http://situs.com), dan email (nama@situs.com).

2) *Case folding*, yaitu penyeragaman bentuk huruf serta penghapusan angka dan tanda baca. Dalam hal ini yang digunakan hanya huruf latin antara a sampai dengan z.

3) *Parsing*, yaitu proses memecah dokumen menjadi sebuah kata. Hal ini sesuai dengan fitur digunakan yaitu unigram.

Berikut adalah proses pemilihan dan ekstraksi fitur yang akan digunakan sebagai dasar proses klasifikasi(Nur et al, 2011).

1) *Part of Speech* (POS) *Tagger*, yaitu proses memberikan kelas pada kata. Kelas kata yang dipilih adalah kata sifat (*adjective*), kata keterangan (*adverb*), kata benda (*noun*) dan kata kerja (*verb*), sesuai dengan penelian Liu(2010), bahwa keempat jenis kata di atas merupakan jenis kata yang paling banyak mengandung sentimen. Penentuan kelas kata berdasarkan Kamus Besar Bahasa Indonesia (KBBI).

2) *Stemming*, bertujuan mengurangi variasi kata yang memiliki kata dasar sama. Seperti pada proses POS *Taggger*, proses *stemming* dilakukan dengan menggunakan bantuan KBBI. Dari

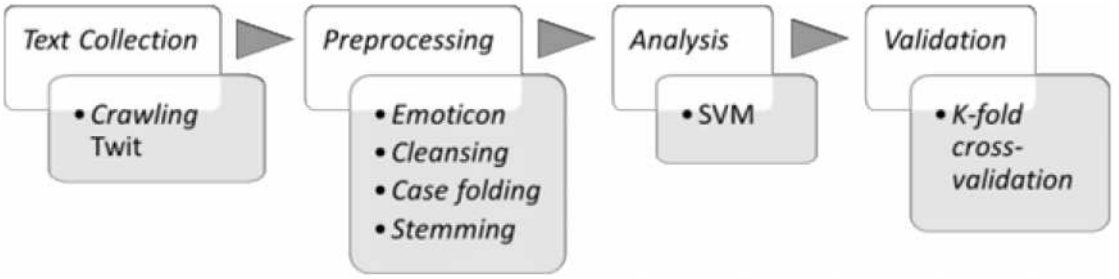
ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

proses pemilihan dan ekstraksi fitur ini maka didapatkan 5 (lima) buah dataset yang berbeda (Tabel 1).

Tabel 1. Label data oleh Nur et al.(2011)

Dataset Label	Deskripsi
TwOri	Merupakan data *tweet* yang sudah dilakukan *preprocessing*.
TwLex	Merupakan data *tweet* yang sudah dilakukan *preprocessing* dan dilakukan filtering dengan menghapus kata-kata yang tidak ada di dalam KBBI.
TwLexRoot	Merupakan data *tweet* yang sudah dilakukan *preprocessing* dan dilakukan filtering dengan menghapus kata-kata yang tidak ada di dalam KBBI dan dilakukan proses stemming, sehingga hanya berupa kumpulan dari kata dasar.
TwLexAAVN	Merupakan data *tweet* yang sudah dilakukan *preprocessing* dan dilakukan filtering dengan menghapus kata-kata yang tidak ada di dalam KBBI dan hanya diambil kata yang memiliki kelas *adjective, adverb, verb* dan *noun.*
TwLexAAVNRoot	Merupakan data *tweet* yang sudah dilakukan *preprocessing* dan dilakukan filtering dengan menghapus kata-kata yang tidak ada di dalam KBBI dan dilakukan proses stemming, sehingga hanya berupa kumpulan dari kata dasar, dan hanya diambil kata yang memiliki kelas *adjective, adverb, verb* dan *noun.*

Terdapat penelitian lain dari Monarizqa, et al.(2014) yang secara garis besar terdiri dari dua bagian yakni proses konfigurasi metode analisis sentimen dengan tahapan seperti Gambar 4 dan proses pembuatan aplikasi.



Gambar 4. Bagan Proses Konfigurasi Metode Analisis Sentimen oleh Monarizqa, et al.(2014)

5.1 Proses Konfigurasi Metode Analisis Sentimen

Monarizqa, et al.(2014) dalam penelitiannnya memberikan keterangan sebagai berikut.

1) *Text Collection*: dilakukan dengan menggunkan streaming API dengan penambahan filter geolocation

2) *preProcessing*: Berbeda dengan penelitian berjudul "*Twitter Sentiment Classification using Distant Supervision*" (A. Go, R. Bhayani & L. Huang, 2009) yang menempatkan emotikon sebagai *noise*, pada penelitian ini emotikon digunakan sebagai elemen pembobotan. Dilakukan penyaringan pada data yang diperoleh dengan menggunakan tabel konversi emotikon menurut penelitian "Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini Terhadap Tokoh Publik" (I. Sunni & D. H. Widyantoro, 2012) yang telah dimodifikasi seperti pada Tabel 2:

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

Tabel 2. Tabel Konversi Emotikon

Emotikon	Konversi	Polaritas	Label
>:] :-) :) :o) :] :3 :c) :> =] 8) =) :} \(^.^)/ \(?)/ (~^.^)~ ^_^ ^^ ^ ^ :^)	Emohappy	Positif	1
>:D :-D :D :)) =)) 8-D) 8D x-D xD =-D =D =-3 =3	Emolaugh	Positif	1
>:[:-(:(:-c) :c :-< :< :-[:[:{ >.> <.< >.<	Emosad	Negatif	-1

Setelah itu dilakukan *cleansing* untuk mengganti mention dengan "*usernameaa*", hashtag dengan "*hashtagaa*", menghapus tautan dan simbol. *Case folding* digunakan untuk menyeragamkan jenis karakter. *Stemming* adalah proses untuk melakukan ekstraksi kata dasar, menggunakan pustaka Lucene (Lucene 3.1.0 API) *class IndonesianStemmer* (F. Tala, 2003).(Monarizqa, et al., 2014)

Pada proses ini didapatkan sejumlah 1.358.417 (14%) twit mengandung emotikon dengan 215.211 twit mengandung emosad (negatif) dan 1.143.206 twit mengandung emolaugh dan emohappy (positif). Sebagai model, hanya diambil 175.000 twit saja dengan 8.750 twit positif dan 8.750 twit negatif yang kemudian tiap twitnya diproses seperti pada Tabel 3:(Monarizqa, et al., 2014)

Tabel 3. Contoh Hasil Preprocessing oleh Monarizqa et al.(2014)

Jenis *preprocessing*	Hasil
Tidak ada	@Babang_iecal aku terima keputusan perusahaan tapi sangat terpukul juga dengan keputusan itu :(#akurapopo (at Yandhy Home) â€ https://t.co/KXcJ0r3Nnj
Cleansing	usernameaa aku terima keputusan perusahaan tapi sangat terpukul juga dengan keputusan itu hashtagaa at Yandhy Home
Case folding	usernameaa aku terima keputusan perusahaan tapi sangat terpukul juga dengan keputusan itu hashtagaa at yandhy home
Stemming	usernameaa ima putus usaha pukul putus hashtagaa at yandhy home
Konversi tiap kata ke indeks menjadi baris .dat	-1 14517:1.000000 58648:1.000000 61784:1.000000 112156:1.000000 113052:1.000000 144881:1.000000 151342:1.000000

Keseluruhan twit kemudian dijadikan dalam satu file .dat dan dilakukan pengujian menggunakan SVM dengan bantuan SVM *light* . Validasi yang digunakan adalah 7-*fold cross validation*.

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

Tabel 4. Tabel Hasil Konfigurasi oleh Monarizqa et al.(2014)

Model	*Precision*	*Recall*	*Accuracy*
0	71.60%	72.52%	72.81%
1	75.56%	71.86%	74.30%
2	74.50%	70.94%	73.33%
3	76.44%	71.34%	74.67%
4	75.91%	68.34%	73.32%
5	74.91%	69.57%	73.13%
6	74.73%	67.86%	72.46%
Average	**74.81%**	**70.35%**	**73.43%**

Mengacu pada Tabel 4, ternyata konfigurasi tersebut yakni perpaduan antara beberapa proses *preProcessing* dan algoritme SVM menghasilkan akurasi 73,43%. Akurasi ini sudah dianggap baik sehingga konfigurasinya digunakan untuk core aplikasi (Monarizqa et al., 2014).

Nilai akurasi dari konfigurasi algoritme SVM dengan *preProcessing* yang diterapkan penelitian ini untuk analisis sentimen pada teks berupa twit berbahasa Indonesia adalah sebesar 73.43%. Konfigurasi ini kemudian menjadi core aplikasi analisis sentimen. Aplikasi analisis sentimen dibangun dengan menggunakan Twitter API untuk mengambil data *realtime*, pustaka Lucene sebagai alat stemming pada core aplikasi, SVMlight, dan Java. Ketika konfigurasi diterapkan dalam aplikasi analisis sentimen yang dibuat, nilai akurasi ketika menggunakan kata kunci "Jokowi" sebesar 68%, "Prabowo" sebesar 56%, "kalimilk" sebesar 70% dan "sunmor" sebesar 74% (Monarizqa et al., 2014).

Beberapa hal yang dapat dilakukan dalam *preProcessing* agar hasil prediksi dapat lebih baik antara lain mendeteksi dan menghilangkan twit duplikasi (spam), mereduksi huruf beruntun dan menghilangkan angka. Ektraksi topik perlu dilakukan pada bagian analisis agar prediksi dapat lebih tepat(Monarizqa et al., 2014).

Saputra, et al(2015) melakukan penelitian terhadap data sentimen Jokowi. Pada penelitian ini terdapat pengurangan data yang cukup besar, sehingga pencarian data Jokowi dapat dilakukan dengan efektif dan efisien. Total keseluruhan data yang dipakai dalam penelitian ini untuk data positif sebanyak 59, data netral sebanyak 60, dan data negatif sebanyak 58. Pada penelitian ini, dilakukan *preProcessing* yang berbeda pada masing-masing data, teknik *preProcessing* yang dilakukan pada penelitian ini terdapat pada kolom teknik *preProcessing*, teknik *preProcessing* sebagai berikut.

Teknik *preProcessing* A
TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = Yes ; Tokenizer = N-Gram ; dan Emoticon = Yes.
Teknik *preProcessing* B
TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = Yes ;

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

Tokenizer = N-Gram ; dan Emoticon = No.ANALISIS SENTIMEN DATA PRESIDEN JOKOWI

Teknik *preProcessing* C

TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = No ; Tokenizer = N-Gram ; dan Emoticon = Yes.

Teknik *preProcessing* D

TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = No ; Tokenizer = Unigram ; dan Emoticon = Yes.

Teknik *preProcessing* E

TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = No ; Tokenizer = Bigram ; dan Emoticon = Yes.

Teknik *preProcessing* F

TF-IDF = Yes ; Lowercase = Yes ; minTermFreq = 1 ; Normalize all data ; Stopwords = No ; Tokenizer = Trigram ; dan Emoticon = Yes.

Agar tabel cukup satu halaman, dilakukan penyingkatan terhadap judul tabel, singkatan tersebut adalah sebagai berikut.(Saputra et al, 2015)

LC = Lowecase, MTF = minTermFreq, N = normalize, SW = Stopwords, Met = Metode yang digunakan, T = Tokenizer, TPre = Teknik *preProcessing*, Emo = Emoticon, NB = Naive Bayes, TTB = Time taken to build model (Waktu yang dibutuhkan untuk membangun model).

5.1.1 Data Sudah Dinormalisasi (Percobaan 1)

Percobaan 1 adalah data yang sudah dilakukan normalisasi tetapi belum dilakukan stemming dapat dilihat pada Tabel 5 (Saputra et al., 2015).

Tabel 5. Data Sudah Dinormalisasi oleh Saputra, et(2015)

TPre	Met	TF-IDF	LC	MTF	N	SW	T	Emo	TTB(s)	Hasil (%)
A	NB	Yes	Yes	1	1	Yes	N-Gram	Yes	1,08	83,0508
A	SVM	Yes	Yes	1	1	Yes	N-Gram	Yes	0,66	83,0508
B	NB	Yes	Yes	1	1	Yes	N-Gram	No	1,13	80,226
B	SVM	Yes	Yes	1	1	Yes	N-Gram	No	0,55	79,096
C	NB	Yes	Yes	1	1	No	N-Gram	Yes	1,14	86,4407
C	**SVM**	**Yes**	**Yes**	**1**	**1**	**No**	**N-Gram**	**Yes**	**0,5**	**88,7006**
D	NB	Yes	Yes	1	1	No	Unigram	Yes	0,23	88,1356
D	SVM	Yes	Yes	1	1	No	Unigram	Yes	0,22	88,1356
E	NB	Yes	Yes	1	1	No	Bigram	Yes	0,5	67,7966
TPre	Met	TF-IDF	LC	MTF	N	SW	T	Emo	TTB(s)	Hasil (%)
E	SVM	Yes	Yes	1	1	No	Bigram	Yes	0,74	66,6667
F	NB	Yes	Yes	1	1	No	Trigram	Yes	0,59	42,3729
F	SVM	Yes	Yes	1	1	No	Trigram	Yes	0,44	40,678

Pada Tabel 5 dapat dilihat bahwa setelah data dilakukan normalisasi dari tidak baku menjadi baku, metode Naive Bayes rata-rata lebih unggul daripada SVM. Walaupun rata-rata akurasi yang tinggi adalah Naive Bayes, tetapi akurasi tertinggi adalah dengan menggunakan token N-

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

Gram pada teknik *preProcessing* C menggunakan metode SVM dengan akurasi sebesar 88,7006% (Saputra et al., 2015).

5.1.2 Data Sudah Dinormalisasi dan Stemming (Percobaan 2)

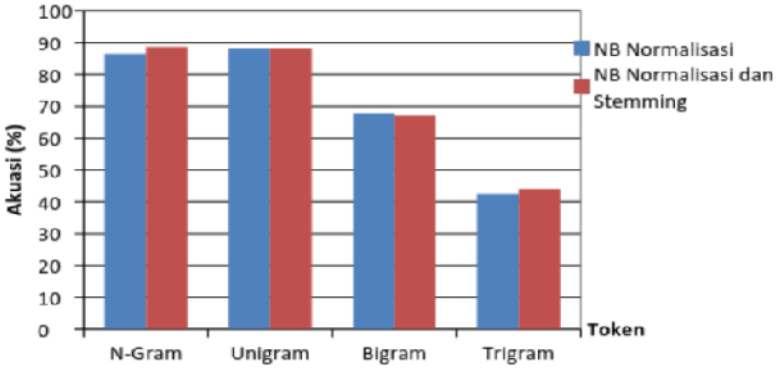
Percobaan 2 ini adalah data yang sudah dinormalisasi dan stemming, hasil akurasi pada jenis data ini dapat dilihat pada Tabel 6.

Tabel 6. Data sudah normal dan stemming oleh Saputra et al.(2015)

TPre	Met	TF-IDF	LC	MTF	N	SW	T	Emo	TTB(s)	Hasil (%)
A	NB	Yes	Yes	1	1	Yes	N-Gram	Yes	1,15	83,0508
A	SVM	Yes	Yes	1	1	Yes	N-Gram	Yes	0,61	81,9209
B	NB	Yes	Yes	1	1	Yes	N-Gram	No	0,92	81,9209
B	SVM	Yes	Yes	1	1	Yes	N-Gram	No	0,56	81,3559
C	NB	Yes	Yes	1	1	No	N-Gram	Yes	1,03	88,7006
C	SVM	Yes	Yes	1	1	No	N-Gram	Yes	0,59	88,7006
D	NB	Yes	Yes	1	1	No	Unigram	Yes	0,34	88,1356
D	**SVM**	**Yes**	**Yes**	**1**	**1**	**No**	**Unigram**	**Yes**	**0,36**	**89,2655**
E	NB	Yes	Yes	1	1	No	Bigram	Yes	0,66	67,2316
E	SVM	Yes	Yes	1	1	No	Bigram	Yes	0,27	66,1017
F	NB	Yes	Yes	1	1	No	Trigram	Yes	0,49	44,0678
F	SVM	Yes	Yes	1	1	No	Trigram	Yes	0,38	44,0678

Pada Tabel 6 terdapat peningkatan akurasi dibandingkan Tabel 5 pada beberapa teknik *preProcessing*. Selain itu, terjadi perubahan akurasi tertinggi dibanding Tabel 5, pada Tabel 6 akurasi tertinggi dengan menggunakan token Unigram dan metode SVM dengan hasil akurasi sebesar 89,2655% (Nurirwan et al, 2015).

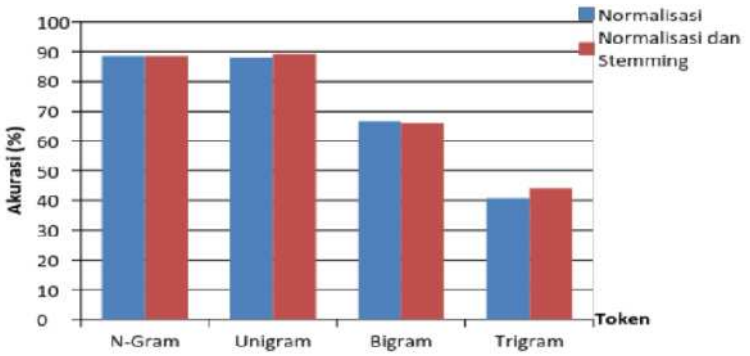
Grafik perbandingan token pada metode Naive Bayes berdasarkan percobaan 1 dan 2 dapat dilihat pada Gambar 5 (Nurirwan et al, 2015).



Gambar 5. Grafik Perbandingan Token pada Metode Naive Bayes oleh Saputra et al.(2015)

Dan grafik perbandingan token pada metode SVM berdasarkan percobaan 1 dan 2 dapat dilihat pada Gambar 6 (Nurirwan Saputra, 2015).

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016



Gambar 6. Grafik Perbandingan Token pada Metode SVM oleh Saputra et al.(2015)

Pada Grafik yang terlihat pada Gambar 5 dan Gambar 6 memperlihatkan bahwa metode N-Gram dan Unigram memperlihatkan akurasi yang lebih baik dibanding Bigram dan Trigram, kemudian Bigram lebih baik dibandingkan Trigram. Unigram dan N-Gram lebih tinggi dikarenakan masing-masing membagi kalimat menjadi satu kata, sehingga kemungkinan mendapatkan kata yang sama sangat besar sedangkan pada Bigram membagi kalimat per dua kata, sehingga dapat diartikan dua kata tersebut menjadi satu arti dan lebih kecil kemungkinan bertemu dua kata yang sama tersebut di kalimat yang lain, terlebih lagi trigram yang membagi kalimat per tiga kata, lebih kecil lagi kemungkinan bertemu kembali kepada tiga kata yang sama yang sudah dibagi di tiap kalimat (Saputra et al., 2015).

Berdasarkan penelitian yang dilakukan, dapat ditarik kesimpulan sebagai berikut (Saputra et al., 2015).

1) Berdasarkan penelitian yang dilakukan, akurasi yang dihasilkan ketika data dilakukan stemming, terdapat peningkatan rata-rata sebesar 0,85% untuk metode Naive Bayes dan 0,85% untuk metode SVM.

2) Akurasi yang dihasilkan metode SVM tidak selalu unggul dibandingkan metode Naive Bayes, begitu pula sebaliknya. Untuk metode yang paling tinggi di masing-masing percobaan adalah metode SVM pada percobaan 1 mendapatkan akurasi 88,7006% untuk teknik *preProcessing* C, percobaan 2 mendapat akurasi 89,2655% untuk teknik *preProcessing* D.

3) Dengan melakukan normalisasi dan stemming, hasil yang didapat lebih tinggi dibandingkan dengan data yang hanya dilakukan normalisasi saja, dikarenakan dengan adanya normalisasi, kata-kata yang tidak sesuai dengan KBBA dapat disesuaikan dengan standar yang ada, ditambah lagi dengan adanya stemming, kata-kata yang dianggap berbeda karena terdapat imbuhan, dapat diubah terlebih dahulu berdasarkan kata dasarnya, sehingga variasi kata semakin sedikit, semakin sering muncul, dan dapat memberikan bobot positif, negatif ataupun netral terhadap kata tersebut semakin baik.

Saran yang bisa diberikan berdasarkan penelitian yang sudah dilakukan di antaranya sebagai berikut (Saputra et al., 2015).

1) Perlu adanya stopwords yang dikhususkan untuk analisis sentimen yang mampu meningkatkan akurasi pada analisis sentimen.

2) Penelitian mengenai analisis sentimen menggunakan teknik sinonim dan akronim.

ISBN: 978-979-3649-96-2 *Unisbank Semarang, 28 Juli 2016*

Adapun sebuah penelitian dari Haddi(2013) membahas tentang peranan preprocessing dalam analisis sentimen. Tabel 7 membandingkan kinerja classifier yang dihasilkan dari klasifikasi pada kedua data baik yang tidak diberlakukan pre-processing maupun yang diberlakukan pre-processing untuk masing-masing matriks fitur (TF-IDF, FF, FP).

Tabel 7. Akurasi Klasifikasi dalam Prosentase dari Dat-1400, kolom not pre-proc mengacu pada hasil yang dilaporkan oleh (B. Pang et al., 2002) dalam Haddi(2013)

	TF-IDF		*FF*			*FP*		
	no pre-proc	pre-proc	no pre-proc1	no pre-proc2	pre-proc	no pre-proc1	no pre-proc2	pre-proc
Accuracy	78.33	81.5	72.7	76.33	83	82.7	82.33	83
Precision	76.66	83	NA	77.33	80	NA	80	82
Recall	79.31	80.58	NA	76.31	85.86	NA	83.9	83.67
F-Measure	77.96	81.77	NA	76.82	82.83	NA	81.9	82.82

Tabel 7 menunjukkan bahwa untuk data yang tidak tunduk pada pre-processing, peningkatan yang baik terjadi pada akurasi dari matriks FF, dari 72,8% yang dilaporkan dalam (B. Pang et al., 2002) untuk 76,33%, sementara mereka akurasi dari matriks FP sedikit yang berbeda, kita mencapai 82,33% sedangkan (B. Pang et al., 2002) melaporkan 82,7%. Selain itu kami memperoleh 78,33% akurasi di TF-IDF matrix mana (B. Pang et al., 2002) tidak menggunakan TF-IDF. Dengan menyelidiki lebih lanjut dalam hasil kita melihat peningkatan akurasi ketika menerapkan classifier pada data pre-proc setelah transformasi data, dengan akurasi tertinggi 83% untuk kedua matriks FF dan FP.(Haddi, 2013)

Tabel 7 menunjukkan bahwa meskipun akurasi dicapai dalam matriks FP dekat satu dicapai sebelumnya dan dalam (B. Pang et al., 2002), ada perubahan besar dalam kinerja classifier pada TF-IDF dan FF matriks, dan ini menunjukkan pentingnya berasal dan menghapus stopwords dalam mencapai akurasi yang lebih tinggi dalam klasifikasi sentimen. Kami menekankan bahwa untuk dapat menggunakan classifier SVM pada seluruh dokumen, kita harus merancang dan menggunakan kernel untuk itu masalah tertentu (B. Scholkopf, 1997).(Haddi, 2013)

6. KESIMPULAN

6.1 Kesimpulan

Berdasarkan pembahasan di atas dapat diketahui bahwa:

1) *Preprocessing* memiliki peran dalam menjaga sistem analisis sentimen dari gangguan *tweet spam* atau *retweet*. Walaupun demikian, proses ini tidak dapat menjamin sepenuhnya pada ketahanan terhadap gangguan terutama dari BOT.
2) Terdapat beberapa variasi subproses dari *preprocessing*, sebagian penelitian memasukkan pembobotan atau pemilihan fitur ke dalamnya, namun sebagian yang lain hanya memasukkan proses yang cukup minim seperti *case folding* dan *cleansing*.

6.2 Saran

1) Perlu dilakukan penelitian lebih lanjut terkait preprocessing khususnya dikaitkan dengan klasifikasi. Hal ini dapat memberikan wawasan terkait perbedaan peran dari kedua proses tersebut pada sistem analisis sentimen dalam hal mengatasi gangguan seperti BOT.

ISBN: 978-979-3649-96-2 Unisbank Semarang, 28 Juli 2016

2) Perlu dikaji lebih dalam terkait variasi dari preprocessing beserta subprosesnya, mungkin seperti analisis komparatif. Hal ini dapat memberikan wawasan terkait perbedaan serta background dari masing-masing versi.
3) Perlu kajian khusus terkait analisis sentimen publik yang dikhususkan pada opini publik terkait bakal calon legsilatif ataupun gubernur. Hal ini juga dapat terkait dengan perancangan sistem yang memang disesuaikan dengan objek tersebut.

DAFTAR PUSTAKA

Ahsan, Muhammad Rafiqudin. (2015). *Pelaksanaan Kegiatan Promosi JToku pada Web Series Tahun 2012-2014*: Ilmu Komunikasi. Universitas Muhammadiyah Yogyakarta. Yogyakarta: Indonesia.

Nur Muhamad Yusuf & Diaz D. Santika. (2011). *Analisis Sentimen Pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine*. Universitas Bina Nusantara, Jakarta.

J. Han, M. Kamber, & J. Pei. *Data Mining: Concepts and Techniques, Second Edition, 2nd ed*. Morgan Kaufmann, 2006.

Hidayat, Andi Nurul. (2015). *Analisis Sentimen Terhadap Wacana Politik Pada Media Masa Online Menggunakan Algoritma Support Vector Machine Dan Naive Bayes*. Jurnal Elektronik Sistem Informasi dan Komputer (Jesik) Vol.1 No.1 Januari-Juni 2015. Palu: STMIK Bina Mulia.

Sheela, L.Jaba. (2016). *A Review of Sentiment Analysis in Twitter Data Using Hadoop*. International Journal of Database Theory and Application Vol.9, No.1 (2016), pp.77-86. Available at http://dx.doi.org/10.14257/ijdta.2016.9.1.07. Panimalar Engineering College, Chennai, India.

Nur, Muhamad Yusuf & Santika, Diaz D.. (2011). *Analisis Sentimen Pada Dokumen Berbahasa Indonesia dengan Pendekatan Support Vector Machine*. Konferensi Nasional Sistem dan Informatika 2011; Bali, November 12, 2011. KNS&I11-002. Jakarta: Universitas Bina Nusantara.

Liu, Bing. (2010). *Sentiment Analysis and Subjectivity, in Handbook of Natural Language Processing.*

Novantirani, Anita & Sabariah, Mira Kania, S.T., M.T & Effendy, Veronikha, ST.,M.T. (2014). *Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat dalam Kota dengan Metode Support Vector Machine*. Bandung: Universitas Telkom.

A. Go, R. Bhayani dan L. Huang. (2009). *Twitter Sentiment Classification using Distant Supervision*. California.

I. Sunni dan D. H. Widyantoro. (2012). *Analisis Sentimen dan Ekstraksi Topik Penentu Sentimen pada Opini terhadap Tokoh Publik.* vol. 1, no. 2.

F. Tala. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Institute for Logic, Language and Computation, Universiteit van Amsterdam.

Saputra, Nurirwan & Adji, Teguh Bharata & Permanasari, Adhistya Erna. (2015). *Analisis Sentimen Data Presiden Jokowi Dengan preProcessing Normalisasi dan Stemming*

Menggunakan Metode Naive Bayes dan SVM. Jurnal Dinamika Informatika,Volume 5, Nomor 1, November 2015.

Syahfitri Kartika Lidya & Opim Salim Sitompul & Syahril Efendi. (2015). *Sentiment Analysis pada Teks Bahasa Indonesia Menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN).* Seminar Nasional Teknologi Informasi dan Komunikasi 2015 (SENTIKA 2015) Yogyakarta, 28 Maret 2015, ISSN: 2089-9815 Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara Medan

Djaelani, Aunu Rofiq. (2013). *Teknik Pengumpulan Data Dalam Penelitian Kualitatif*. Majalah Ilmiah Pawiyatan Vol : XX, NO : 1, MARET 2013, hal. 88 .Semarang: FPTK IKIP Veteran.

Dewi, Ni Nengah, Cholissodin dan Hayyuning. (2014). *Analisis Sentimen Pencitraan Elite Politik Berdasarkan Opini Melalui Media Sosial Twitter Menggunakan Metode Additive Kernel SVM*. Program Studi Informatika/Ilmu Komputer, Universitas Brawijaya, Malang.

L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu. (2011). *Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.* HPL-2011-89, vol. 89, Jun. 2011.

B. Liu. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher.

Marsland, Stepen. (2009). "Machine Learning : An Algoritma Perspective" [PowerPoint slide]. Diambil dari www.cis.temple.edu/~latecki/Courses/AI-Fall10/Lectures/ch5SVM.ppt

B. Pang, L. Lee, S. Vaithyanathan. (2002). *Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

B. Scholkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, V. Vapnik. *Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers.* Signal Processing. IEEE Transactions on 45 (11) (1997) 2758 2765.

Seth Redmore, Jan 2013, CMO,Lexalytics:10 years in networking, 10 in text analytics; 3 for 4 on startups https://www.quora.com/Can-we-do-sentiment-analysis-on-Facebook-data-like-we-do-on-Twitter-using-API-in-Python-How-If-not-is-there-any-other-way-to-do-it. Diakses pada tanggal 09 Apr 2016 10:35:58 WIB

http://www.republika.co.id/berita/pemilu/hot-politic/14/07/04/n86tax-melihat-media-sosial-jokowijk-diperkirakan-menang-pemilu Jumat, 04 Juli 2014. Diakses tanggal 25 April 2016

Alia, Siti Sarifah & Amal Nur Ngazis. (2014). http://teknologi.news.viva.co.id/news/read/519990-sentigram-mindtalk--sentimen-positif-prabowo-65-. Diakses pada tanggal 09 Apr 2016 06:03:51.

Monarizqa, Nuvirta, Nugroho, Lukito Edi, Hantono, Bimo Sunarfri. (2014). *Penerapan Analisis Sentimen pada Twitter Berbahasa Indonesia sebagai Pemberi Rating*. Artikel Reguler JTETI UGM, Volume 1 Nomor 3, Oktober 2014.

ISBN: 978-979-3649-96-2 *Unisbank Semarang, 28 Juli 2016*