

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/292831965>

Studi Literatur Tentang Perbandingan Metode Untuk Proses Analisis Sentimen di Twitter

Conference Paper · March 2016

CITATIONS

22

READS

15,194

3 authors:



Nurrin Muchammad Shiddieqy Hadna

Universitas Gadjah Mada

2 PUBLICATIONS 22 CITATIONS

SEE PROFILE



Paulus Insap Santosa

Universitas Gadjah Mada

247 PUBLICATIONS 2,067 CITATIONS

SEE PROFILE



Wing Winarno

STIE YKPN

52 PUBLICATIONS 320 CITATIONS

SEE PROFILE

STUDI LITERATUR TENTANG PERBANDINGAN METODE UNTUK PROSES ANALISIS SENTIMEN DI TWITTER

Nurrun Muchammad Shiddieqy Hadna¹, Paulus Insap Santosa², Wing Wahyu Winarno³

¹²³Departemen Teknik Elektro dan Teknologi Informasi, Fakultas Teknik, Universitas Gadjah Mada

Jl. Grafika No. 2 Kampus UGM, Yogyakarta, 55281

Telp (0274) 547506, 510983

E-mail: diky.cio13@mail.ugm.ac.id, insap@ugm.ac.id, wing@mti.ugm.ac.id

ABSTRAKS

The growth of websites and social provide us a huge-and-easy-accessed-textual-resources. Twitter is an example of a social media that has been widely used. People usually post their opinions, experiences or whatever concern them through this service. Thank's to technology, at the present of time we are able to analyze the sentiment polarity of the tweets. There are several methods that can be used to do sentiment analysis e.g. Naive Bayes Classifier, Support Vector Machine and Maximum Entropy. Before the sentiment analysis proces can be started, data must be preprocessed to make sure that the data is ready to analyze. In this preprocessing phase, the data should be labeled, tokenized and weighted, then the data can be analyzed. After being analyzed, the quality of the sentiment analysis process should be measured. There are several parameters that can be used to measure the quality of sentiment analysis e.g. accuracy, precision and recall.

Perkembangan berbagai macam portal web dan media sosial yang sangat pesat membuat ketersediaan sumber dokumen tekstual menjadi sangat besar dan mudah diakses. Salah satu contoh media sosial yang banyak digunakan adalah Twitter. Melalui media sosial ini, berbagai macam opini, pengalaman maupun hal lain yang menjadi perhatian banyak dituliskan setiap harinya oleh para penggunanya. Tulisan ini sering disebut dengan istilah tweet. Berkat perkembangan teknologi, tweet yang dikirimkan oleh pengguna Twitter sudah bisa dianalisis polaritas sentimennya. Ada beberapa macam metode untuk melakukan analisis sentimen yang bisa digunakan, antara lain Naive Bayes Classifier, Support Vector Machine dan Maximum Entropy. Sebelum data dapat dianalisis, terlebih dahulu dilakukan tahap preprocessing untuk memastikan data siap untuk dianalisis. Tahapan ini antara lain pelabelan, tokenisasi dan pembobotan kata, baru kemudian data memasuki tahap analisis sentimen. Setelah proses sentimen analisis dilakukan, masih terdapat proses lanjutan untuk menentukan kualitas pengujian dengan beberapa parameter, antara lain accuracy, precision dan recall.

Kata Kunci: Analisis sentimen, Naive Bayes Classifier, Support Vector Machine, Twitter

1. PENDAHULUAN

1.1 Latar Belakang

Sebuah survey yang dilakukan oleh APJII pada tahun 2014 (Asosiasi Pengusaha Jasa Internet Indonesia, 2014) menunjukkan bahwa jumlah pengguna internet di Indonesia mengalami kenaikan dalam beberapa tahun terakhir. Pada tahun 2005, tercatat sebanyak 16 juta orang pengguna internet di Indonesia dan angka ini bertambah menjadi 20 juta orang pengguna internet pada tahun 2006. Bahkan, pada tahun 2015 diprediksi bahwa terdapat pengguna internet sebanyak 139 juta orang di Indonesia. Salah satu kunci utama kesuksesan penetrasi internet di Indonesia adalah perkembangan infrastruktur yang sudah mencapai daerah pelosok di Indonesia (Asosiasi Pengusaha Jasa Internet Indonesia, 2012). Penelitian ini juga memberikan hasil bahwa penggunaan internet paling banyak adalah akses ke situs media sosial (87,4% dari total responden), terbanyak kedua adalah berselancar di situs-situs web (68,7% dari total responden) dan terbanyak ketiga adalah penggunaan internet untuk pesan instan (59,9% dari total responden).

Pertumbuhan teknologi informasi, terutama media sosial, banyak mengubah cara manusia untuk melakukan komunikasi dengan sesamanya. Penggunaan media sosial banyak dimanfaatkan oleh masyarakat umum. Masyarakat banyak menggunakan media sosial untuk mengekspresikan opini, pengalaman maupun hal lain yang menjadi perhatian mereka (Troussas et al., 2013). Hal-hal semacam ini acapkali disebut dengan sentimen.

Analisis sentimen adalah sebuah teknik untuk mendeteksi opini terhadap suatu subyek (misalnya individu, organisasi ataupun produk) dalam sebuah kumpulan data (Nasukawa & Yi, 2003). Menurut Cutlip dan Center dalam Resphati (Resphati, 2010), opini adalah pengekspresian suatu sikap mengenai persoalan yang mengandung suatu sikap mengenai persoalan yang mengandung pertentangan. Opini di sini bisa berupa *mention* di media sosial maupun artikel di situs berita dan blog pribadi.

Pada berbagai penelitian tentang analisis dokumen tekstual yang sudah pernah dilakukan (Taheri & Mammadov, 2013; Hall, 2006; Shahheidari et al., 2013), paling banyak menggunakan metode *Naive Bayes Classifier*. *Naive*

Bayes Classifier adalah salah satu algoritma yang populer digunakan untuk keperluan *data mining* karena kemudahan penggunaannya (Hall, 2006) serta waktu pemrosesan yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan tingkat efektifitas yang tinggi (Taheri & Mammadov, 2013)

Selain Naive Bayes Classifier, metode lain yang juga cukup populer dan banyak digunakan di berbagai penelitian adalah *Support Vector Machine*.

Di dalam paper ini, akan dibahas tahapan yang dilalui untuk melakukan proses analisis sentimen. Dimulai dari tahap *preprocessing* (pelabelan, tokenisasi dan pembobotan), kemudian tahap analisis sentimen dengan perbandingan dua metode yaitu *Naive Bayes Classifier* dan *Support Vector Machine* serta bagaimana mengukur kualitas hasil analisis menggunakan beberapa parameter seperti *accuracy*, *precision* dan *recall*.

1.2 Twitter

Twitter adalah sebuah media sosial dan layanan *microblogging* yang memungkinkan penggunanya untuk mengirimkan pesan *realtime*. Pesan ini populer dengan sebutan *tweet*. *Tweet* adalah sebuah pesan pendek dengan panjang karakter yang dibatasi hanya sampai 140 karakter. Dikarenakan keterbatasan karakter yang bisa dituliskan, sebuah *tweet* seringkali mengandung singkatan, bahasa *slang* maupun kesalahan pengejaan (Agarwal et al., 2014). Sejak awal, Twitter memang dibuat sebagai layanan berbasis *mobile* yang didesain sesuai dengan batasan karakter pada sebuah pesan teks (SMS), dan sampai hari ini, Twitter masih bisa digunakan pada sembarang telepon genggam yang memiliki kemampuan untuk mengirim dan menerima pesan teks (The Twitter Government and Election Team, 2014).

Twitter diciptakan untuk menjadi tempat saling berbagi pengalaman antar sesama penggunanya tanpa adanya sekat penghalang. Dengan menggunakannya, pengguna akan mudah untuk mengikuti tren, cerita, informasi dan berita dari seluruh penjuru dunia. Selain itu, Twitter juga membantu penggunanya untuk selalu terhubung dengan orang-orang terdekatnya. Ketika penggunanya mengirimkan *tweet*, pesan tersebut bersifat publik dan bisa diakses oleh siapapun, dimanapun dan kapanpun. Bahkan, bagi orang-orang yang mengikuti (*follow*) akun Twitter tersebut, *tweet* tersebut akan secara otomatis muncul di lini masa orang tersebut.

Berikut ini adalah beberapa istilah yang dikenal dalam Twitter:

1. *Mention*. *Mention* adalah menyebut atau memanggil pengguna Twitter lain dalam sebuah *tweet*. *Mention* dilakukan dengan menuliskan '@' diikuti dengan nama pengguna lain.
2. *Hashtag*. *Hashtag* digunakan untuk menandai sebuah topik pembicaraan di Twitter. Penulisan *hashtag* dimulai dengan tanda '#' diikuti dengan topik yang sedang dibahas. *Hashtag* biasa digunakan untuk meningkatkan visibilitas *tweet* pengguna.
3. *Emoticon*. *Emoticon* adalah ekspresi wajah yang direpresentasikan dengan kombinasi antara huruf, tanda baca dan angka. Pengguna biasa menggunakan emoticon untuk mengekspresikan mood yang sedang mereka rasakan.
4. *Trending topics*. Jika *hashtag* adalah cara untuk menandai sebuah topik pembicaraan di Twitter, maka *trending topics* adalah kumpulan dari topik pembicaraan yang sangat populer di Twitter.

Sebuah fakta yang cukup mengejutkan disampaikan oleh Giummole et al. (Giummole et al., 2013) bahwa *trending topics* di Twitter akan meningkatkan prediksi hasil pencarian di Google. Hal ini bisa terjadi karena dalam setiap lima menit, Twitter akan mengeluarkan daftar topik yang sangat populer (*trending topics*) dengan cara memonitor dan menganalisa pembicaraan, sedangkan Google juga akan mengeluarkan daftar pencarian populer yang dicari oleh penggunanya pada setiap jam.

Pengembangan Twitter sampai saat ini telah menawarkan kesempatan untuk mempelajari perilaku budaya manusia yang belum pernah dilakukan sebelumnya. Hal ini bisa dilakukan karena Twitter sudah menyiapkan *structured data* (misalnya XML dan daftar kategori) maupun *unstructured data* (misalnya pada bagian teks *tweet*) (Liu et al., 2012; Dey et al., 2013). Twitter memiliki model *structured* (Liu et al., 2012) sekaligus *unstructured text* (Liu et al., 2012; Dey et al., 2013). *Structured text* bisa dilihat dalam implementasi metadata di setiap *tweet*-nya, dimana selalu terdapat informasi *username*, *timestamp*, *text*, *retweet*, *favorite* dan informasi lainnya. *Unstructured text* bisa dilihat di bagian *text* atau lebih sering disebut dengan *tweet* karena konten pada bagian ini memang tidak memiliki struktur khusus. Satu-satunya aturan yang berlaku adalah jumlah maksimal karakter yang bisa dimasukkan adalah sebanyak 140 karakter (Hassan et al., 2013), (Doan et al., 2012).

1.3 Analisis Sentimen

Sentiment analysis atau analisis sentimen dalam Bahasa Indonesia adalah sebuah teknik atau cara yang digunakan untuk mengidentifikasi bagaimana sebuah sentimen diekspresikan menggunakan teks dan bagaimana sentimen tersebut bisa dikategorikan sebagai sentimen positif maupun sentimen negatif (Nasukawa & Yi, 2003). Pendapat yang hampir senada dikemukakan oleh (Cvijikj & Michahelles,

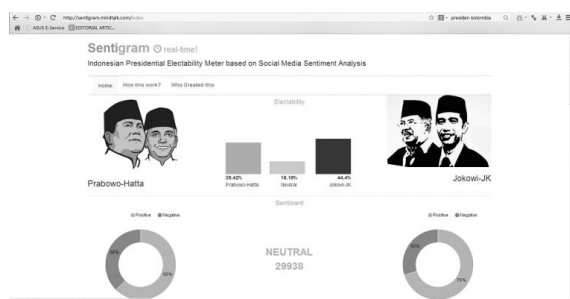
2011), di mana analisis sentimen digunakan untuk memahami komentar yang diciptakan oleh pengguna (internet) dan menjelaskan bagaimana sebuah produk maupun brand diterima oleh mereka.

Sedangkan menurut (Coletta et al., 2014), analisis sentimen adalah proses yang digunakan untuk menentukan opini, emosi dan sikap yang dicerminkan melalui teks, dan biasanya diklasifikasikan menjadi opini negatif dan positif.

Dari ketiga pendapat di atas, bisa diambil kesimpulan bahwa analisis sentimen adalah sebuah proses untuk menentukan sentimen atau opini dari seseorang yang diwujudkan dalam bentuk teks dan bisa dikategorikan sebagai sentimen positif atau negatif.

Sebagaimana yang sudah dituliskan sebelumnya bahwa pengguna internet banyak menuliskan pengalaman, opini dan segala hal yang menjadi perhatian mereka. Tulisan tentang apa yang mereka rasakan ini berupa perasaan positif, netral maupun negatif yang bisa diungkapkan dengan cara yang cukup kompleks (Troussas et al., 2013).

Salah satu aplikasi analisis sentimen yang sempat populer di Indonesia adalah Sentigram, sebuah aplikasi berbasis web yang dibuat oleh PT Ansvia. Perusahaan yang merupakan induk dari jejaring MindTalk ini membuat aplikasi analisis sentimen terhadap calon presiden pada masa kampanye Pemilihan Umum Presiden Indonesia tahun 2014 yang lalu. Halaman depan Sentigram bisa dilihat pada Gambar 1.



Gambar 1. Halaman Depan Sentigram

Walaupun saat ini situs Sentigram sudah tidak bisa lagi diakses, namun dari berbagai sumber informasi lain yang masih bisa diakses (Beritasatu.com n.d.; MetroTV News n.d.; Daily Social n.d.), Sentigram berhasil memprediksi kemenangan pasangan calon presiden Jokowi-JK yang kini sudah menjabat sebagai Presiden dan Wakil Presiden Republik Indonesia untuk periode 2014-2019. Hal ini menunjukkan bahwa analisis sentimen cukup *reliable* untuk melakukan prediksi kejadian di dunia nyata.

2. TAHAPAN ANALISIS SENTIMEN

Untuk melakukan analisis sentimen, terdapat beberapa tahapan yang perlu dilakukan untuk mendapatkan hasil pengujian yang terbaik.

Berikut ini adalah penjabaran tentang tahapan analisis sentimen.

2.1 Pengumpulan & Pelabelan Data

Tahap pertama untuk melakukan proses analisis sentimen adalah pengumpulan data. Dalam penelitian ini, data diambil dari Twitter. Pengambilan data dari Twitter cukup mudah dilakukan karena Twitter sudah menyediakan API (*Application Programming Interface*) (Twitter Inc., 2015) yang ditujukan kepada pengembang sistem untuk mempermudah pengambilan data dari Twitter.

Selain Twitter, data bisa dikumpulkan dari sumber lain, misalnya portal berita, forum daring, media sosial lain maupun situs blog pribadi. Hal yang membedakan dari Twitter adalah cara pengambilan data karena tidak semua situs menyediakan API yang bisa digunakan sembarang orang.

Setelah data berhasil dikumpulkan menjadi sebuah dataset, tahap selanjutnya adalah pelabelan. Pelabelan di sini ditujukan untuk membagi data ke dalam beberapa kelas sentimen yang akan digunakan dalam penelitian. Jumlah kelas sentimen yang banyak digunakan adalah dua dan tiga kelas, yaitu negatif, netral dan positif. Tujuan dari proses pelabelan ini adalah membagi dataset menjadi 2 bagian, yaitu menjadi *training data* dan *testing data*. *Training data* adalah data yang digunakan untuk melatih sistem agar mampu mengenali pola yang sedang dicari, sedangkan *testing data* adalah data yang digunakan untuk menguji hasil pelatihan yang sudah dilakukan. Untuk melakukan pembagian, salah satu metode yang bisa digunakan adalah *n-fold cross validation* yang akan dijelaskan pada poin selanjutnya.

Ada baiknya, dalam proses pelabelan ini melibatkan ahli atau pakar bahasa sesuai bahasa yang digunakan. Tujuannya adalah ahli atau pakar bahasa tersebut lebih memahami pemaknaan dari setiap teks yang tertulis.

2.2 Preprocessing

Setelah dilakukan pelabelan data, tahap selanjutnya yang harus dilalui adalah *preprocessing*. Tahap ini adalah tahapan di mana data disiapkan agar menjadi data yang siap untuk dianalisis. Ada beberapa tahap dalam *preprocessing* ini, antara lain *tokenization*, *cleansing* dan *filtering*.

Berikut ini adalah penjelasan secara rinci tentang tahap *preprocessing*.

2.2.1 Tokenization

Tokenisasi adalah sebuah proses yang dilakukan untuk memotong atau memecah kalimat menjadi beberapa bagian atau kata (Manning et al.,

2009). Hasil dari pemotongan ini disebut dengan *token*. Pada beberapa kasus, proses tokenisasi juga dilakukan dengan membuang tanda baca yang tidak diperlukan.

Ada beberapa model *tokenization* yang bisa digunakan, yaitu *unigram*, *bigram*, *trigram* dan *n-gram*.

2.2.2 Cleansing

Cleansing adalah suatu tahap di mana karakter maupun tanda baca yang tidak diperlukan dibuang dari teks. Contoh karakter yang dibuang adalah tanda seru, tanda tanya, koma dan titik.

2.2.3 Filtering

Filtering adalah tahap menghilangkan kata-kata yang muncul dalam jumlah besar, namun dianggap tidak memiliki makna (*stopwords*). Pada dasarnya, *stop words list* adalah sekumpulan kata-kata yang banyak digunakan dalam berbagai bahasa. Alasan penghapusan *stop words* dalam banyak program aplikasi yang berkaitan dengan *text mining* adalah karena penggunaannya yang terlalu umum, sehingga pengguna dapat berfokus pada kata-kata lain yang jauh lebih penting (Ganesan, 2015).

2.3 Pembobotan Kata

Pembobotan kata adalah suatu mekanisme untuk memberikan skor terhadap frekuensi kemunculan sebuah kata dalam dokumen teks. Salah satu metode populer untuk melakukan pembobotan kata adalah TF-IDF (*Term Frequency-Inverse Document Frequency*). *Term Frequency-Inverse Document Frequency* adalah sebuah metode pembobotan yang menggabungkan dua konsep, yaitu *Term Frequency* dan *Document Frequency* (Asrofi, 2015). *Term Frequency* adalah konsep pembobotan dengan mencari seberapa sering (frekuensi) munculnya sebuah term dalam satu dokumen (Huang & Wu, 2013). Dikarenakan setiap dokumen memiliki panjang yang berbeda-beda, bisa saja terjadi sebuah kata muncul lebih banyak di dokumen yang panjang dibandingkan dengan dokumen-dokumen yang pendek. Dengan demikian, *term frequency* sering dibagi dengan panjangnya dokumen (total kata yang ada di dokumen tersebut).

Sedangkan *Document Frequency* adalah banyaknya jumlah dokumen di mana sebuah term itu muncul (Huang & Wu, 2013). Semakin kecil frekuensi kemunculannya, maka semakin kecil pula nilai bobotnya. Ketika proses perhitungan *term frequency*, semua kata di dalamnya dianggap sama pentingnya. Namun, terdapat kata yang sebenarnya kurang penting dan tidak perlu diperhitungkan seperti “di-”, “ke-”, “dan” dan lain sebagainya. Oleh sebab itu, kata-kata yang kurang penting tersebut perlu dikurangi bobotnya dan menambah bobot kata penting lainnya. Inilah ide dasar mengapa diperlukan *stopword*.

Skor TF-IDF dapat diperoleh menggunakan Persamaan (1).

$$tf - idf_{t,d} = tf_{t,d} * idf_t \quad (1)$$

2.4 Analisis Sentimen

2.4.1 Naive Bayes Classifier

Naive Bayes Classifier adalah salah satu algoritma yang populer digunakan untuk keperluan *data mining* karena kemudahan penggunaannya (Hall, 2006) serta waktu pemrosesan yang cepat, mudah diimplementasikan dengan strukturnya yang cukup sederhana dan tingkat efektifitas yang tinggi (TaHERi & Mammadov, 2013).

Dengan bahasa yang lebih sederhana, *Naive Bayes Classifier* mengasumsikan bahwa keberadaan maupun ketidakberadaan sebuah fitur dalam sebuah kelas tidak memiliki keterkaitan dengan keberadaan maupun ketidakberadaan fitur lainnya. Sebagai contoh, sesuatu yang berwarna merah, bulat dan memiliki diameter sekitar 10 cm bisa dikategorikan sebagai buah apel. Walaupun fitur ini bergantung antara satu dengan yang lainnya, *Naive Bayes Classifier* akan tetap menganggap bahwa fitur-fitur tersebut independen dan tidak memiliki pengaruh satu sama lainnya (Rocha, 2006).

Bergantung pada model probabilitasnya, *Naive Bayes Classifier* dapat dilatih untuk melakukan *supervised learning* dengan sangat efektif. Dalam berbagai macam penerapannya, estimasi parameter untuk model *Naive Bayes* menggunakan metode *maximum likelihood*; yang artinya pengguna dapat bekerja menggunakan model *Naive Bayes* tanpa perlu mempercayai probabilitas *Bayesian* atau tanpa menggunakan metode *Bayesian*.

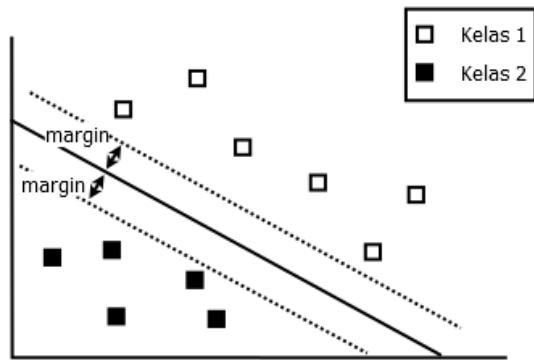
Naive Bayes Classifier dapat diformulasikan menjadi Persamaan (2).

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)} \quad (2)$$

2.4.2 Support Vector Machine

Support Vector Machine adalah seperangkat metode pembelajaran terbimbing (*supervised learning*) yang menganalisis data dan mengenali pola, digunakan untuk klasifikasi dan analisis regresi (Saraswati, 2011). Klasifikasi ini dilakukan dengan mencari *hyperplane* atau garis pembatas (*decision boundary*) yang memisahkan antara satu kelas dengan kelas yang lain (Novantirani et al., 2015). Dalam konsep ini, *Support Vector Machine* berusaha untuk mencari *hyperplane* terbaik diantara fungsi yang tidak terbatas jumlahnya (Widiastuti, 2007). Fungsi yang tidak terbatas dalam pencarian *hyperplane* di metode *Support Vector Machine* merupakan sebuah keuntungan, dimana pemrosesan pasti akan selalu bisa dilakukan bagaimanapun data

yang dimilikinya. *Hyperplane* dapat dilihat pada Gambar 2.



Gambar 2. Hyperplane memisahkan dua kelas

Data yang digunakan pada metode Support Vector Machine dengan $x_i \in \mathbb{R}^d$ dan label dinotasikan dengan $y_i \in \{1, 2\}$ untuk $i = 1, 2, \dots, l$ yang mana l adalah banyaknya data.

Support Vector Machine dapat diformulasikan ke dalam Persamaan (3) untuk $y_i = +1$ dan Persamaan (4) untuk $y_i = -1$.

$$x_i * w + b \geq +1 \quad (3)$$

$$x_i * w + b \leq -1 \quad (4)$$

2.5 Pengukuran Kualitas Hasil Uji

Setelah proses pengujian ini selesai dilakukan, diperlukan satu tahap lagi untuk menentukan kualitas dari proses yang sudah dilakukan, yaitu evaluasi hasil. Pada tahapan ini, performa dari perhitungan yang sudah dilakukan akan diuji dengan dua parameter, yaitu *accuracy*, *precision* dan *recall* (Manning et al., 2009), (Feldman & Sanger, 2007).

Precision (P) adalah seberapa banyak hasil pemrosesan yang relevan terhadap informasi yang ingin dicari atau dengan kata lain *precision* adalah klasifikasi positif yang benar (*true positive*) dan keseluruhan data yang diprediksikan sebagai kelas positif. *Precision* dapat diperoleh dengan menggunakan Persamaan (4).

$$P = \frac{tp}{tp + fp} \quad (4)$$

Recall (R) adalah seberapa banyak dokumen relevan dalam *collection* yang dihasilkan oleh sistem, atau dengan kata lain *recall* adalah jumlah dokumen yang memiliki klasifikasi positif yang benar (*true positive*) dari semua dokumen yang benar-benar positif (termasuk di dalamnya *false*

negative). *Recall* dapat diperoleh dengan menggunakan Persamaan (5).

$$R = \frac{tp}{tp + fn} \quad (5)$$

Sedangkan *Accuracy* (A) adalah jumlah dokumen yang diklasifikasikan dengan benar, baik itu *true positive* maupun *true negative*. *Accuracy* dapat diperoleh dengan menggunakan Persamaan (6).

$$A = \frac{(tp + tn)}{(tp + fp + tn + fn)} \quad (6)$$

Variabel seperti TN, TP, FN dan FP berasal dari *confusion matrix*. TN adalah singkatan dari *True Negative*, data negatif yang diklasifikasikan sebagai negatif. TP adalah singkatan dari *True Positive*, data positif yang diklasifikasikan sebagai positif. FN adalah singkatan dari *False Negative*, data positif yang diklasifikasikan sebagai negatif. FP adalah singkatan dari *False Positive*, data negatif yang diklasifikasikan sebagai positif. Untuk penjelasan lebih jelas, perhatikan Tabel 1.

Tabel 1. Confusion Matrix (Prabowo & Thelwall, 2009)

	Mesin Katakan Ya	Mesin Katakan Tidak
Manusia Katakan Ya	TP	FN
Manusia Katakan Tidak	FP	TN

Setelah data berhasil dikumpulkan, maka data tersebut akan dibagi menjadi dua bagian, yaitu *training data* dan *testing data*. *Training data* adalah data yang digunakan untuk melatih sistem agar mampu mengenali pola yang sedang dicari, sedangkan *testing data* adalah data yang digunakan untuk menguji hasil pelatihan yang sudah dilakukan. Pembagian data tersebut akan dilakukan dengan menggunakan metode *n-fold cross validation*. *N-fold cross validation* adalah membagi keseluruhan dokumen menjadi n bagian sama besar, kemudian proses *training* dan *testing* dilakukan sejumlah n kali, dan setiap kali pemrosesan menggunakan bagian yang berbeda sebagai *testing data* (Feldman & Sanger, 2007). *N-fold cross validation* (dalam hal ini $n = 10$) dapat divisualisasikan pada Tabel 2.

Tabel 2. Visualisasi Metode N-Fold Cross Validation (n=10)

Uji ke-	Pembagian Data Ke-									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Dari Tabel 2, bagian kotak berwarna putih adalah *training data*, sedangkan bagian kotak berwarna hitam adalah *testing data*.

2.6 Perbandingan Metode

Penelitian yang dilakukan oleh (Padmaja et al., 2014) membandingkan dua metode analisis sentimen, yaitu *Naive Bayes Classifier* dan *Support Vector Machine*. Penelitian ini dilakukan dengan mengumpulkan data berupa berita sebanyak 689 artikel terhadap tiga partai yang bersaing pada Pemilu India 2009 yaitu United Progressive Alliance (UPA), Telugu Desam Party (TDP) dan Telangana Rasthra Samithi (TRS).

Data yang digunakan dalam penelitian ini dikumpulkan dari surat kabar harian yang beredar di India dan menggunakan Bahasa Inggris seperti *The Hindu* dan *Times of India and Economic Times*. Hasil penelitian ini menunjukkan bahwa secara umum *Support Vector Machine* memiliki skor parameter yang lebih baik dibandingkan dengan *Naive Bayes Classifier*, walaupun dalam beberapa perhitungan *Naive Bayes Classifier* tampak lebih unggul.

Hasil perbandingan metode yang dilakukan oleh Padmaja et al. dapat dilihat pada Tabel 3.

Tabel 3. Hasil Perbandingan Metode Analisis oleh Padmaja et al.

Me- tri- c	UPA			TDP			TRS		
	SV M	NB	S W N	SV M	NB	S W N	SV M	NB	S W N
Re- cal- l	0.7 19	0.6 52	0.5 8	0.7 00	0.7 25	0.5 0	0.6 67	0.6 36	0.5 4
Pre- cisi- on	0.6 90	0.7 25	0.8 6	0.6 89	0.7 21	0.5 4	0.6 58	0.6 23	0.6 3

F- Me- asu- re	0.6 88	0.6 35	0.5 65	0.6 87	0.7 23	0.5 1	0.6 57	0.6 2	0.4 8
Ac- cur- acy	0.7 12	0.6 52	0.7 42	0.7 00	0.7 25	0.6 01	0.6 66	0.6 36	0.6 01

Penelitian lain tentang perbandingan metode *Naive Bayes Classifier* dan *Support Vector Machine* juga dilakukan oleh (Chandani et al., 2014). Penelitian ini dilakukan terhadap 1000 data *review* film yang diambil dari situs IMDB. Hasil dari penelitian ini menunjukkan bahwa metode *Support Vector Machine* lebih baik jika dibandingkan dengan metode *Naive Bayes Classifier*, walaupun selisih skor parameter yang dihasilkan tidak terpaut jauh.

Hasil perbandingan metode yang dilakukan oleh Chandani et al. dapat dilihat pada Tabel 4.

Tabel 4. Hasil Perbandingan Metode Analisis oleh Chandani et al.

	Informatio n Gain		Chi Square		Forward Selection		Backward Eliminatio n	
	Top K (K=200)		Top K (K=100)					
	Acc ura cy	AU C	Acc ura cy	AU C	Acc ura cy	AU C	Acc ura cy	AU C
AN N	91.4 0%	0.91 4	79.6 0%	0.90 0	75.5 0%	0.78 1	70.2 0%	0.72 4
SV M	81.5 0%	0.92 9	80.8 0%	0.85 3	67.6 7%	0.68 9	79.2 5%	0.84 4
NB	80.8 0%	0.85 3	80.3 0%	0.86 7	79.0 0%	0.80 7	71.2 5%	0.68 9
Ave rage	84.5 7%	0.89 9	80.2 3%	0.87 3	74.0 6%	0.76 2	73.5 7%	0.75 2

Penelitian lain yang dilakukan oleh (Taheri & Mammadov, 2013) tentang model optimasi klasifikasi juga membandingkan beberapa metode klasifikasi yang berbeda seperti *Naive Bayes*, *Augmented Naive Bayes*, SVM, C4.5 dan NN.

Data set yang digunakan dalam penelitian ini adalah *public data sets* untuk *machine learning* yang diambil dari UCI seperti data set tentang *breast cancer*, *congress vote*, *credit approval*, *diabetes*, *German.numer*, *haberman*, *heart disease*, *hepatitis*, *ionosphere*, *liver disorders*, *sonar*, *spambase*, *svmguide1* dan *svmguide3*.

Simpulan akurasi dari hasil uji perbandingan metode ini dapat dilihat pada Tabel 5.

Tabel 5. Simpulan Hasil Uji Perbandingan Metode oleh Taheri & Mammadov

Data	NB	SVM
Breast Cancer	96.37	95.32
Congress Vote	91.43	96.76
Credit Approval	84.92	85.51
Diabetes	75.93	76.72
German.numer	75.32	76.15
Haberman	75.22	73.54

Tabel 5. Lanjutan

Data	NB	SVM
Heart Disease	81.67	80.54
Hepatitis	83.76	83.43
Ionosphere	82.95	85.98
Liver Disorders	61.84	60.05
Sonar	75.16	76.99
Spambase	90.31	90.47
Svmguide1	92.72	93.31
Svmguide3	81.34	80.11
Average	82.06	82.49

3. KESIMPULAN

Paper ini berisi informasi mengenai perbandingan dua metode yang bisa digunakan untuk melakukan proses analisis sentimen disertai tahapan yang akan dilalui. Metode yang dibandingkan adalah *Naive Bayes Classifier* dan *Support Vector Machine*. Dari beberapa penelitian sebelumnya, bisa dilihat bahwa hasil terbaik dapat diraih menggunakan metode *Support Vector Machine*. Namun, untuk tetap mendapatkan hasil klasifikasi terbaik, maka lebih baik dilakukan pengujian yang bersifat komparatif dengan beberapa metode klasifikasi pada penelitian yang sedang dilakukan.

PUSTAKA

- Agarwal, A. et al., 2014. Sentiment Analysis of Twitter Data. *Department of Computer Science Columbia University*. Available at: <http://www.cs.columbia.edu/~julia/papers/Agarwaleta11.pdf>.
- Asosiasi Pengusaha Jasa Internet Indonesia, 2012. Ini Potret Pengguna Internet Indonesia. Available at: <http://www.apjii.or.id/v2/read/article/apjii-at-media/129.html>.
- Asosiasi Pengusaha Jasa Internet Indonesia, 2014. *Statistik Pengguna Internet Indonesia 2014*, Available at: <http://www.apjii.or.id/v2/read/page/halaman-data/9/statistik.html>.
- Asrofi, G., 2015. *Analisis Sentimen Calon Presiden Indonesia 2014 dengan Lima Class Attribute*. Universitas Gadjah Mada. Available at: http://etd.repository.ugm.ac.id/index.php?mod=penelitian_detail&sub=PenelitianDetail&act=view&typ=html&buku_id=80122&obyek_id=4 [Accessed October 6, 2015].
- Beritasatu.com, Sentigram: Unggul di Lima Situs, Elektabilitas Jokowi-JK 44,22%, | Politik | Beritasatu.com. Available at: <http://www.beritasatu.com/politik/195463-sentigram-unggul-di-lima-situs-elektabilitas-jokowijk-4422.html> [Accessed November 12, 2015].
- Chandani, V., Wahono, R.S. & Purwanto, ., 2014. Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis

Sentimen Review Film. *Journal of Intelligent Systems*, 1(1), pp.55–59. Available at: <http://journal.ilmukomputer.org/index.php/jis/article/view/10> [Accessed April 29, 2015].

- Coletta, L.F.S. et al., 2014. Combining Classification and Clustering for Tweet Sentiment Analysis. In *2014 Brazilian Conference on Intelligent Systems*. IEEE, pp. 210–215. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6984832> [Accessed May 11, 2015].
- Cvijikj, I.P. & Michahelles, F., 2011. Understanding Social Media Marketing: A Case Study on Topics, Categories and Sentiment on a Facebook Brand Page.
- Daily Social, Mengikuti Tren Sentimen Terhadap Capres di Media Sosial dengan Sentigram - Dailysocial. Available at: <http://dailysocial.id/post/mengikuti-tren-sentimen-terhadap-capres-dengan-sentigram?number=2014061838926> [Accessed November 12, 2015].
- Dey, L. et al., 2013. A Framework to Integrate Unstructured and Structured Data for Enterprise Analytics. *Information Fusion (FUSION)*, 2013 16th International Conference on, pp.1988–1995.
- Doan, S., Ohno-Machado, L. & Collier, N., 2012. Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses. In *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*. IEEE, pp. 62–71. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6366191> [Accessed June 7, 2014].
- Feldman, R. & Sanger, J., 2007. *The Text Mining Handbook*.
- Ganesan, K., 2015. A Brief Note on Stop Words for Text Mining and Retrieval. Available at: <http://www.text-analytics101.com/2014/10/all-about-stop-words-for-text-mining.html>.
- Giummole, F., Orlando, S. & Tolomei, G., 2013. Trending Topics on Twitter Improve the Prediction of Google Hot Queries. In *2013 International Conference on Social Computing*. IEEE, pp. 39–44. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6693309> [Accessed June 7, 2014].
- Hall, M., 2006. A Decision Tree-Based Attribute Weighting Filter for Naive Bayes. *Knowledge-Based Systems*, pp.120–126. Available at: <http://www.cs.waikato.ac.nz/pubs/wp/2006/uow-cs-wp-2006-05.pdf>.
- Hassan, A., Abbasi, A. & Zeng, D., 2013. Twitter Sentiment Analysis: A Bootstrap Ensemble Framework. In *2013 International Conference*

- on *Social Computing*. IEEE, pp. 357–364. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6693353> [Accessed June 4, 2014].
- Huang, X. & Wu, Q., 2013. Micro-blog commercial word extraction based on improved TF-IDF algorithm. In *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*. IEEE, pp. 1–5. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6718884> [Accessed October 6, 2015].
- Liu, X. et al., 2012. SocialCube: A Text Cube Framework for Analyzing Social Media Data. In *2012 International Conference on Social Informatics*. IEEE, pp. 252–259. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6542448> [Accessed October 26, 2015].
- Manning, C.D., Raghavan, P. & Schütze, H., 2009. *An Introduction to Information Retrieval*, MetroTV News, Sentigram Prediksi Kemenangan Jokowi-Jusuf Kalla. Available at: <http://teknologi.metrotvnews.com/read/2014/07/09/263502/sentigram-prediksi-kemenangan-jokowi-jusuf-kalla> [Accessed November 12, 2015].
- Nasukawa, T. & Yi, J., 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*. pp. 70–77.
- Novantirani, A., Sabariah, M.K. & Effendy, V., 2015. *Analisis Sentimen pada Twitter untuk Mengenai Penggunaan Transportasi Umum Darat Dalam Kota dengan Metode Support Vector Machine*.
- Nuruddinsyah, M.F., 2015. *Media Sosial dalam Komunikasi dan Kampanye Politik*, Available at: https://www.academia.edu/7525397/Media_Sosial_dalam_Komunikasi_dan_Kampanye_Politik [Accessed November 2, 2015].
- Padmaja, S. et al., 2014. Comparing and evaluating the sentiment on newspaper articles: A preliminary experiment. In *2014 Science and Information Conference*. IEEE, pp. 789–792. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6918276> [Accessed October 12, 2015].
- Prabowo, R. & Thelwall, M., 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), pp.143–157.
- Resphati, A., 2010. *Opini Masyarakat Tentang Pemberitaan Demo 100 Hari Pemerintahan SBY-Boediono di Surat Kabar Jawa Pos Edisi 27-29 Januari 2010*. Universitas Pembangunan Nasional “Veteran” Jawa Timur.
- Rocha, A. de R., 2006. Naive Bayes Classifier Teaching Material. , pp.1–9. Available at: <http://www.ic.unicamp.br/~rocha/teaching/2011s2/mc906/aulas/naive-bayes-classifier.pdf>.
- Saraswati, N.W.S., 2011. *Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis*. Universitas Udayana.
- Shahheidari, S., Dong, H. & Daud, M.N.R. Bin, 2013. Twitter Sentiment Mining: A Multi Domain Analysis. In *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*. IEEE, pp. 144–149. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6603880> [Accessed June 4, 2014].
- Taheri, S. & Mammadov, M., 2013. Learning The Naive Bayes Classifier With Optimization Models. In *International Journal of Applied Mathematics and Computer Science*. pp. 787–795.
- The Twitter Government and Election Team, 2014. *The Twitter Government and Election Handbook*, San Francisco: Twitter Inc. Available at: <https://g.twimg.com/elections/files/2014/09/16/TwitterGovElectionsHandbook.pdf>.
- Troussas, C. et al., 2013. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In *IISA 2013*. IEEE, pp. 1–6. Available at: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6623713> [Accessed May 7, 2015].
- Twitter Inc., 2015. Twitter API v1.1. Available at: <https://dev.twitter.com/rest/public> [Accessed December 25, 2015].
- Widiastuti, D., 2007. Analisa Perbandingan Algoritma SVM, Naïve Bayes, dan Decision Tree dalam Mengklasifikasikan Serangan (Attack) pada Sistem Pendeteksi Intrusi. *Jurusan Sistem Informasi Universitas Gunadarma*, pp.1–8.