

Pengukuran Happiness Index Masyarakat Kota Bandung pada Media Sosial Twitter Menggunakan Pendekatan Ontologi Top-Down Hierarchy

Happiness Index Measurement of Bandung Citizen on Social Media Twitter Using Top-Down Hierarchy Ontology Approach

Ika Rahayu Ponilan¹, Anisa Herdiani, M.T.², Nungki Selviandro, M.Kom.³

^{1,2,3}Fakultas Informatika Universitas Telkom, Bandung

¹ikarahayuponilan@gmail.com, ²anisaherdiani@gmail.com, ³Selviandro@gmail.com

Abstrak

Paradigma pengukuran tingkat kemakmuran suatu negara dari pendapatan per kapita *Gross National Product* (GNP), sekarang telah digeser oleh paradigma *happiness index* atau *Gross National Happiness* (GNH). Esensi dari GNH adalah kedamaian dan kebahagiaan dari setiap orang, selain itu juga keamanan dan kedaulatan bangsa. Belakangan ini, Bandung menjadi kota pertama di Indonesia yang mengadopsi inovasi peningkatan *happiness index* masyarakat. Pemerintah kota Bandung berharap agar inovasi ini dapat dijadikan *review* untuk menjadikan kota Bandung nyaman ditinggali dan memperbaiki *value* kota Bandung. Penelitian ini akan menganalisis *tweet* dari Twitter secara berkala, ke dalam parameter *happiness* berdasarkan Badan Pusat Statistik (BPS) Bandung dengan menggunakan pendekatan ontologi.

Penelitian ini terbagi atas enam tahap. Tahap pertama yaitu *crawling* (pengambilan) data Twitter berdasarkan wilayah kota Bandung dan melabelkan data. Tahap kedua yaitu *preprocessing* data yang mencakup data *cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *lemmatization*. Tahap ketiga yaitu *sentiment analysis* untuk mengklasifikasikan *tweet* (opini) ke dalam opini positif dan negatif. Tahap keempat yaitu *ontology construction* berdasarkan parameter *happiness index* BPS Bandung. Tahap kelima yaitu pengelompokan *tweet* berdasarkan ontologi yang telah dibangun. Tahap keenam yaitu perhitungan *happiness index* per parameter yang telah ditentukan sebelumnya. Pada penelitian ini menghasilkan nilai *happiness index* masyarakat kota Bandung sebesar 55.50% dari data aktual dan 52.22% dari data prediksi oleh sistem. Nilai tersebut dapat dijadikan sebagai salah satu alat bantu dalam pengambilan kebijakan Pemerintahan Kota Bandung.

Kata kunci: Twitter, *ontology*, *sentiment analysis*, *happiness index*

1. Pendahuluan

Berdasarkan *survey* yang dilakukan oleh Badan Pusat Statistik (BPS) Bandung terhadap masyarakat Bandung, terdapat sepuluh parameter *happiness index* yaitu pendidikan, kesehatan, pekerjaan, pendapatan, keamanan, hubungan sosial, ketersediaan waktu luang, kondisi rumah, kondisi lingkungan dan keharmonisan keluarga [1]. ¹Pada tahun 2014, BPS Bandung mengukur *happiness index* masyarakat kota Bandung dengan cara *survey* yang melibatkan 1.080 koresponden rumah tangga yang tersebar di ratusan blok sensus kota Bandung, dengan memanfaatkan sepuluh parameter. Metode *survey* ini memiliki kekurangan dalam hal biaya yang besar, waktu yang lama dan umumnya dibutuhkan usaha keras dalam pelaksanaannya. Pengambilan opini tidak hanya dapat dilakukan melalui *survey* secara langsung, tapi dapat dilakukan melalui media sosial mengingat pertumbuhan media sosial sangat pesat di masyarakat. Berdasarkan Portal Statistic Online PeerReach per Oktober 2013, pengguna aktif media sosial di Indonesia khususnya Twitter berjumlah 6,5 persen dari seluruh dunia, sehingga Indonesia menempati posisi ketiga. Opini yang diberikan masyarakat dalam Twitter sangat beragam dan memungkinkan antar opini terdapat kesamaan istilah, struktur dan makna yang sebenarnya mengekspresikan domain pengetahuan yang sama. Oleh karena itu diperlukan pendekatan ontologi, untuk mengelompokkan dan menganalisis opini masyarakat Bandung di Twitter ke dalam parameter *happiness index* BPS Bandung.

Penelitian ini terbagi menjadi enam tahap. Tahap pertama yaitu *crawling* (pengambilan) data Twitter berdasarkan wilayah kota Bandung dan pelabelan data. Menggunakan media sosial Twitter, karena data Twitter dapat diambil berdasarkan wilayah. Tahap kedua yaitu *preprocessing* yang mencakup *data cleaning*, *case folding*, *tokenizing*, *stopword removal*, dan *lemmatization*. Tahap ketiga yaitu *sentiment analysis* menggunakan Support Vector Machine (SVM) untuk mengklasifikasikan data ke dalam opini positif atau negatif. SVM digunakan karena dapat membagi data menjadi bentuk klasifikasi yang paling optimal. Tahap keempat yaitu *ontology construction* menggunakan pendekatan ontologi *top-down hierarchy* berdasarkan parameter *happiness index* BPS Bandung. Pendekatan ontologi digunakan karena mampu mendukung pencarian data secara semantik dan *query expansion*, pemetaan dan penggabungan data, dan manajemen pengetahuan, sedangkan menggunakan *top-down hierarchy* dikarenakan kelas (parameter *happiness*) yang akan didefinisikan dimulai dari pendefinisian konsep umum di domain, dilanjutkan dengan pendefinisian konsep khusus. Tahap kelima yaitu pengelompokan opini berdasarkan ontologi yang telah dibangun. Tahap keenam yaitu perhitungan *happiness index* per parameter yang telah ditentukan sebelumnya.

¹ <http://infobandung.co.id/survei-membuktikan-warga-kota-bandung-bahagia/> (diakses 20 Oktober 2015, jam 11:08)

2. Dasar Teori

2.1. Pengukuran Happiness Index

Seiring dengan perkembangan media sosial yang sangat pesat, terdapat pendekatan lain sebagai alternatif untuk pengukuran *happiness*, yaitu yang sudah dilakukan oleh beberapa penelitian diantaranya oleh Kramed, Adam 2010, dengan mengukur *behaviour model* masyarakat sebuah negara yang tergambar pada media sosial. Pendekatannya dilakukan dengan memodelkan sentimen positif dan negatif masyarakat di suatu negara dan menghitung *score*-nya. Apabila lebih besar *score* sentimen positif maka dapat disimpulkan masyarakat pada negara tersebut bahagia dan sebaliknya [2].

2.2. Ontologi

Gruber dalam Antoniou & van Hermelen (2003), mendefinisikan ontologi sebagai sebuah spesifikasi formal dan eksplisit dari sebuah konseptual. Makna konseptual merujuk pada model abstrak dari suatu hal. Eksplisit mengindikasikan bahwa elemen-elemen konseptual harus didefinisikan dengan jelas dan formal berarti bahwa spesifikasi tersebut harus dapat diproses oleh mesin. Ontologi merupakan representasi pengetahuan dari sebuah domain dengan sekumpulan objek dan relasi dideskripsikan oleh *vocabulary* [3]. Uschold dan Jasper dalam Breitman *et al* (2007) mengungkapkan bahwa ontologi mempunyai sebuah *vocabulary* dari *term-term*, spesifikasi dari masing-masing *term* dan sebuah indikasi bagaimana *term-term* tersebut saling berelasi. *Term* merujuk pada konsep-konsep pada sebuah domain [4].

2.3. Twitter

Twitter adalah layanan *microblogging* yang tidak seperti media sosial lainnya, seperti Facebook atau MySpace yang mana hubungan antara orang yang diikuti dan orang yang mengikuti tidak bisa saling merespon. Twitter memungkinkan pengguna mengikuti ataupun diikuti oleh orang lain. Satu pengguna dapat mengikuti pengguna lainnya dan pengguna yang diikuti tersebut tidak perlu untuk mengikuti kembali pengguna yang mengikutinya. Menjadi pengikut di Twitter berarti pengguna tersebut menerima semua pesan yang disebut *tweet* dari pengguna yang diikutinya. Istilah yang digunakan dalam Twitter antara lain, RT untuk *retweet*, '@' untuk mengidentifikasi nama pengguna dan '#' merepresentasikan sebuah *hashtag* atau topik. Twitter sendiri memiliki batas penulisan, yang mana dalam setiap *tweet* hanya bisa terdiri dari 140 karakter. Mekanisme *retweet* mendukung pengguna Twitter untuk menyebarkan informasi pilihan mereka [5].

2.4. API Twitter

API atau yang biasa disebut *Application Programming Interface* adalah suatu program atau aplikasi yang disediakan oleh pihak pengembang tertentu agar pihak pengembang aplikasi lainnya dapat lebih mudah mengakses aplikasi tersebut. Dengan kata lain, API berfungsi sebagai "jembatan" antara aplikasi satu dengan aplikasi yang lain [6].² Pengembang membutuhkan *consumer key* dan *consumer secret* untuk mengakses fasilitas yang disediakan Twitter. *Consumer key* dan *consumer secret* diperoleh setelah pengembang mendaftarkan aplikasinya di <https://apps.twitter.com/>. API Twitter yang digunakan pada penelitian ini adalah REST API Search, yaitu API yang mengizinkan pengembang untuk melakukan pencarian *tweet*.

2.5. Term Frequency

Term Frequency atau sering disebut juga TF, adalah salah satu metode pembobotan *term* yang paling sederhana. Pada metode ini, setiap *term* diasumsikan memiliki proporsi kepentingan sesuai dengan jumlah terjadinya (munculnya) *term* tersebut dalam dokumen. Dengan metode ini, nilai kontribusi (bobot) suatu *term* pada suatu dokumen adalah sama dengan jumlah munculnya *term* tersebut pada dokumen. Bobot *term* (t) pada dokumen (d) diberikan dengan persamaan [7]:

$$TF(d, t) = \frac{count(d, t)}{count(d)}$$
 (1)

TF (d, t) adalah frekuensi munculnya *term* t pada dokumen d.

Penelitian ini menggunakan TF normalisasi dalam proses pengklasifikasian sentimen. TF normalisasi digunakan untuk membandingkan frekuensi sebuah kata dengan jumlah keseluruhan kata pada dokumen [8]. TF normalisasi dihitung dengan persamaan sebagai berikut:

$$TF(d, t) = 0.5 + (0.5 * \frac{count(d, t)}{\max(count(d, t))})$$
 (2)

²http://jagocoding.com/tutorial/427/Pengenalan_Apa_Itu_Twitter_API_dan_Pembuatan_C%20onsumer_Key_da_n_Consumer_Secret.%20[Accessed%2020%20Mei%202016]. (diakses 21 Januari 2016, jam 22:15)

2.6. TF·RF

Term Frequency.Relevance Frequency atau TF·RF adalah metode penggabungan antara metode TF dan RF dengan tujuan untuk mendapatkan performansi yang lebih baik [9]. Pada TF·RF, bobot dari suatu *term* dihitung dengan menggunakan persamaan:

$$TF \cdot RF = TF * \log \left(2 + \frac{a}{\max(1, c)} \right) \tag{3}$$

Keterangan:

a adalah jumlah dokumen pada *positive category* c_j yang mengandung *term* t_i .

c adalah jumlah dokumen pada *negative category* yang mengandung *term* t_i .

2.7. Sentiment Analysis

Pertumbuhan media sosial yang meningkat seperti forum diskusi, *blog*, *micro-blog*, Twitter, komentar dan *posting* dalam situs jaringan sosial di internet, menyebabkan organisasi dan perorangan sudah menggunakan fitur ini sebagai pendukung pembuat keputusan. Indikator yang paling penting dalam opini adalah kata yang disebut *opinion words*. Sebagai contoh kata-kata yang biasanya digunakan untuk menunjukkan opini positif atau negatif adalah “indah”, “luar biasa” dan “hebat” adalah kata opini positif, sedangkan “menakutkan”, “sebal” dan “jengkel” adalah kata opini yang menunjukkan opini negatif. Disamping itu, ada beberapa *frase* atau istilah yang mungkin mengartikan opini positif atau negatif, contohnya “Bagus sekali suaramu, lebih bagus lagi jika kamu tutup mulut.” Kata-kata dan *frase* dalam opini sangat berarti dalam analisis sentimen [10].

2.8. Support Vector Machine

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane-hyperplane* terbaik yang berfungsi sebagai pemisah dua buah *class* pada *input space* [11]. *Hyperplane* atau bidang pemisah berfungsi untuk meminimalkan rata-rata *error* pada data latih dan memiliki generalisasi yang baik. Generalisasi adalah kemampuan sebuah hipotesis untuk mengklasifikasikan data yang tidak terdapat dalam data pelatihan yang benar. SVM bekerja berdasarkan prinsip Structural Risk Minimization (SRM) untuk menjamin generalisasi ini [12].

2.9. Performance Evaluation

Parameter performansi yang biasa diterapkan untuk mengukur performansi suatu *classifier*, dapat juga diterapkan untuk membandingkan performansi metode-metode pembobotan kata. *Precision*, *recall*, dan *f-measure* adalah beberapa contoh dari parameter performansi yang bisa juga dimanfaatkan untuk mencari metode pembobotan yang terbaik dari beberapa metode pembobotan yang diujikan [13]. Gambaran matriks *contingency* untuk mempermudah pemahaman mengenai *precision*, *recall*, dan *f-measure* dapat dilihat pada tabel 2.1.

Tabel 2.1 Matriks Kontingensi untuk Kelas Prediksi dan Aktual

		Predicted Class	
		Class = Yes	Class = No
Actual Class	Class = Yes	TP	FN
	Class = No	FP	TN

Keterangan:

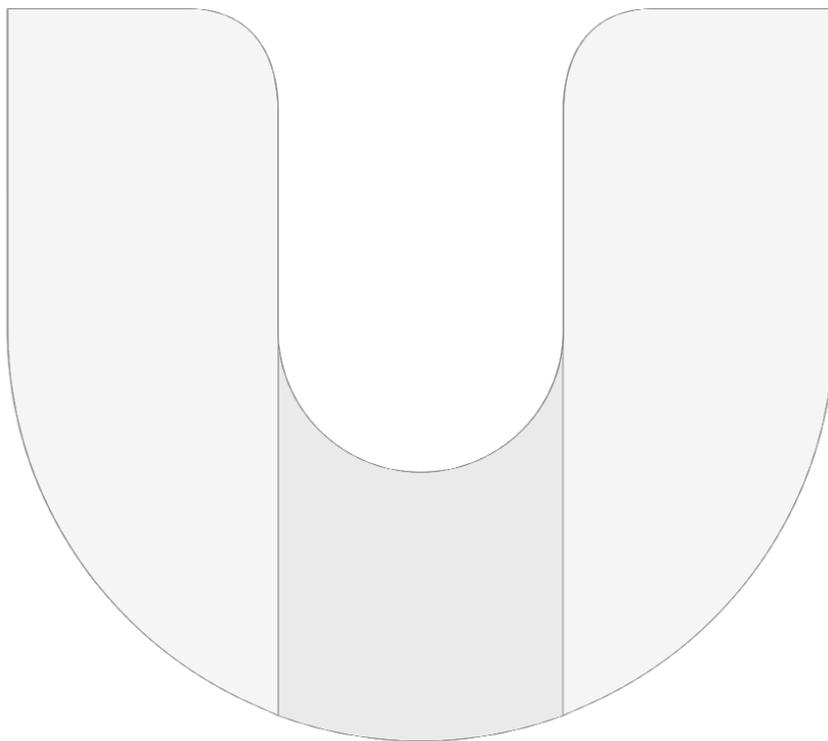
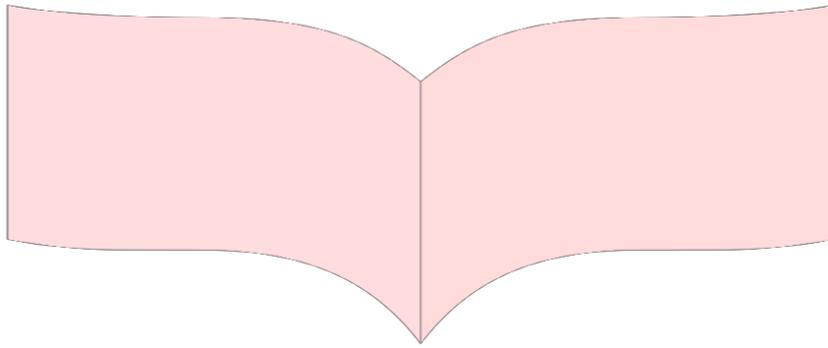
- TP (*True Positive*), adalah kelas yang diprediksi *yes*, dan ternyata faktanya *yes*.
- TN (*True Negative*), adalah kelas yang diprediksi *no*, dan ternyata faktanya *no*.
- FP (*False Positive*), adalah kelas yang diprediksi *yes*, tetapi faktanya *no*.
- FN (*False Negative*), adalah kelas yang diprediksi *no*, tetapi faktanya *yes*.

Precision dapat diartikan sebagai rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total prediksi yang diklasifikasikan ke dalam kelas tersebut [14]. Bila dinyatakan dengan rumus, maka dapat dituliskan sebagai berikut:

$$Precision = \frac{TP}{(TP + FP)} \tag{4}$$

Recall dapat diartikan sebagai rasio dari jumlah ketepatan prediksi suatu kelas terhadap jumlah total fakta yang diklasifikasikan ke dalam kelas tersebut [14]. Bila dinyatakan dengan rumus, maka dapat dituliskan sebagai berikut:

$$R_{\text{eff}} = \frac{TP}{(T_0 + F_0)} \quad (5)$$

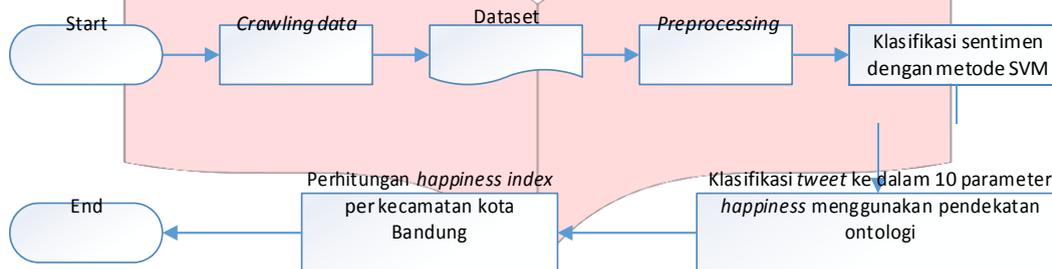


Terkadang perhitungan antara *precision* dan *recall* memiliki perbedaan yang cukup tinggi, oleh karena itu dilakukan penyetaraan nilai *precision* dan *recall* menggunakan *F-Measure*. *F-Measure* merupakan pengukuran performansi yang menggabungkan perhitungan *precision* dan *recall* [14], dan dirumuskan sebagai berikut:

$$F\text{-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

3. Perancangan Sistem

Gambaran umum sistem yang akan dibuat dalam penelitian ini mencakup *crawling* data, *preprocessing* terhadap data hasil *crawling* yang telah dilabel, *sentiment analysis* untuk mengklasifikasikan data ke dalam opini positif dan negatif, mengklasifikasikan data yang telah di-sentimen menggunakan *term matching* ke dalam *bag of words ontology* yang sudah di-construct sebelumnya, perhitungan *happiness index* per kecamatan di kota Bandung. Gambaran umum sistem dapat dilihat pada gambar 3.1 dan alur dari gambaran umum sistem dapat dilihat pada tabel 3.1.



Gambar 3.1 Gambaran Umum Sistem

Tabel 3.1 Gambaran Umum Sistem

No	Tahap	Input	Process	Output
1.	<i>Crawling</i> data	<i>Geocode</i> , <i>token</i> API	<i>Crawling</i> data pada media sosial Twitter berdasarkan wilayah per kecamatan kota Bandung, menggunakan <i>geocode</i> dan <i>token</i> API Twitter	<i>Dataset</i>
2.	<i>Preprocessing</i>	<i>Dataset</i>	Membersihkan <i>dataset</i> dari beberapa karakter yang tidak dibutuhkan dalam penelitian	<i>Bag of words tweet (terms tweet)</i>
3.	Klasifikasi sentimen	<i>Bag of words tweet</i>	<i>Dataset</i> yang telah di- <i>preprocessing</i> akan dibobotkan menggunakan TR.RF. Tujuan pembobotan ini adalah untuk memperoleh pola data sehingga mempermudah <i>machine learning</i> dalam mempelajari data. Setelah itu dilanjutkan ke proses klasifikasi <i>tweet</i> positif atau negatif menggunakan metode SVM dengan memanfaatkan Weka	Hasil klasifikasi sentimen
4.	Klasifikasi <i>tweet</i> menggunakan pendekatan ontologi	<i>Bag of words ontology</i> , <i>bag of words tweet (terms tweet)</i>	Setiap <i>term</i> dari <i>tweet</i> akan di- <i>match</i> dengan <i>terms</i> dari ontologi, proses ini dinamakan <i>term matching</i> . Namun sebelumnya, kedua jenis <i>terms</i> tersebut harus dilakukan proses <i>lemmatization</i> terlebih dahulu, tujuannya untuk mengembalikan <i>term</i> ke bentuk dasar agar seragam. Selanjutnya dilakukan proses pembobotan jumlah kemunculan <i>terms tweet</i> di dalam <i>terms</i> ontologi menggunakan metode TF. Tahap terakhir adalah penentuan kelas di ontologi sesuai dengan jumlah kemunculan <i>terms</i> terbanyak di suatu kelas	Hasil klasifikasi <i>tweet</i> di ontologi
5.	Perhitungan <i>happiness</i> per kecamatan	Data hasil klasifikasi sentimen dan hasil klasifikasi <i>tweet</i> di ontologi	Tahap pertama dari proses ini adalah, menghitung <i>tweet</i> positif dari keseluruhan kota Bandung, dilanjutkan dengan perhitungan <i>tweet</i> positif per kecamatan berdasarkan 10 parameter <i>happiness</i> yang direpresentasikan di kelas ontologi.	<i>Happiness index</i> per kecamatan

4. Implementasi Sistem

Dataset yang digunakan dalam penelitian ini merupakan hasil pencarian otomatis di sistem *crawling* yang telah dibangun dengan memanfaatkan API Twitter. Data ini berasal dari *tweet* 30 kecamatan di kota Bandung. Total *tweet* yang didapatkan dari tanggal 1 Februari hingga 31 Maret 2016 sebanyak 10.767. Tabel 4.1 adalah informasi mengenai jumlah *tweet* yang didapatkan pada 2 bulan (Februari dan Maret 2016).

Tabel 4.1 Detail Jumlah Data Hasil *Crawling*

Bulan	Jumlah <i>tweet</i> yang didapatkan	Berklasifikasi sesuai kebutuhan (%)
Februari	4185	950 (22.70%)
Maret	6582	990 (15.04%)

Adapun informasi mengenai jumlah *tweet* yang diklasifikasikan ke *tweet* positif dan negatif, serta *tweet* yang diklasifikasikan ke 10 parameter *happiness* dapat dilihat pada tabel 4.2

Tabel 4.2 Detail Jumlah Data per Parameter *Happiness* dan Sentimen

Kategori	Sentimen		Total
	positif	negatif	
hubungan sosial	359	253	612
ketersediaan waktu luang	201	82	283
keadaan lingkungan	112	205	317
keharmonisan keluarga	72	15	87
kesehatan	112	129	241
pendapatan rumah tangga	21	20	41
pekerjaan	73	26	99
kondisi keamanan	17	47	64
pendidikan	100	77	177
kondisi rumah dan aset	11	8	19
			1940

Terdapat dua jenis data hasil perhitungan *happiness* yaitu menggunakan data aktual hasil *hand-labeled* dan data hasil prediksi sistem yang dibangun. Penelitian ini menghasilkan perhitungan *happiness* seluruh kota Bandung sebesar 55.5% yang diperoleh dari data aktual. Nilai tersebut didapatkan dari parameter ketersediaan waktu luang sebanyak 10.36%, kesehatan 5.77%, pendidikan 5.15%, pekerjaan 3.76%, pendapatan 1.08%, hubungan sosial 18.5%, keharmonisan keluarga 3.71%, kondisi dan rumah 0.6%, kondisi keamanan 0.9%, dan parameter keadaan lingkungan sebanyak 5.77%. Kecamatan yang paling tinggi nilai *happiness*-nya adalah kecamatan Bandung Kulon. Sedangkan, kecamatan yang paling rendah nilai *happiness*-nya adalah kecamatan Sumur Bandung, dengan nilai *happiness* sebesar 35.2%. Perhitungan *happiness* dari data prediksi oleh sistem didapatkan nilai *happiness* kota Bandung sebesar 55.22%. Nilai tersebut didapatkan dari parameter ketersediaan waktu luang sebanyak 6.28%, kesehatan 5.36%, pendidikan 3.25%, pekerjaan 2.94%, pendapatan 1.65%, hubungan sosial 8.2%, keharmonisan keluarga 1.55%, kondisi dan rumah 0.4%, kondisi keamanan 0.8%, dan parameter keadaan lingkungan sebanyak 6.6%. Perhitungan *happiness* menggunakan data prediksi dari sistem menghasilkan nilai *happiness* tertinggi di kecamatan Regol, sedangkan kecamatan yang paling rendah nilai *happiness*-nya adalah kecamatan Astana Anyar dengan nilai *happiness* sebesar 32.07%.

5. Pembahasan

Sistem yang dibangun pada penelitian ini pada akhirnya akan menghasilkan perhitungan *happiness index* dari seluruh kota Bandung dan 30 kecamatan di kota Bandung. Pengujian dibagi ke dalam 2 modul, yaitu pengujian *tweet* terhadap klasifikasi sentimen dan klasifikasi parameter *happiness* menggunakan *bag of words ontology*. Pengujian pada penelitian ini menggunakan tingkat *precision*, *recall* dan *f-measure* di dalam sistem.

5.1. Pengujian klasifikasi sentimen

Penelitian ini memanfaatkan *classifier SVM* di Weka untuk mengklasifikasikan sentimen data. Proses pengujian menggunakan *cross validation* dengan *10-fold*. Hasil pengujian modul 1 dapat dilihat pada tabel 5.1.

Tabel 5.1 Nilai *Recall*, *Precision*, *F-measure* untuk beberapa Data

	<i>Precision</i>	<i>Recall</i>	<i>F-measure</i>
Pengujian 1 (500 data)	77.6%	77.8%	77.4%
Pengujian 2 (1000 data)	80.3%	80.4%	80.3%
Pengujian 3 (1500 data)	82.1%	82.1%	82.1%
Pengujian 4 (1940 data)	84.0%	84.0%	84.0%

Pada tabel 5.1 dapat dilihat bahwa kenaikan nilai *precision*, *recall* dan *f-measure* berbanding lurus dengan jumlah data yang digunakan dalam pengujian. Hal ini disebabkan, semakin banyak jumlah data maka semakin beragam pula pola data yang dipelajari (di-*learning*) oleh *classifier*, sehingga data tersebut tepat diprediksi sesuai dengan data aktual. Semakin tinggi nilai *precision*, *recall* dan *f-measure*, semakin akurat hasil klasifikasi sistem terhadap data aktual.

5.2. Pengujian klasifikasi parameter happiness menggunakan bag of words ontology

Pengujian ini menggunakan keseluruhan data yang telah dilabelkan sebanyak 1940 data. Berikut adalah hasil pengujian pada modul 2.

Tabel 5.2 Nilai *Precision*, *Recall* dan *F-measure* Klasifikasi Parameter *Happiness*

Parameter	<i>Precision</i>	<i>Recall</i>
ketersediaan waktu luang	0.770732	0.79798
kesehatan	0.868132	0.835979
pendidikan	0.848739	0.848739
pekerjaan	0.523364	0.682927
pendapatan rumah tangga	0.2	0.518519
hubungan sosial	0.914754	0.69403
keharmonisan keluarga	0.615385	0.740741
kondisi rumah dan aset	0.142857	0.4
kondisi keamanan	0.764706	0.634146
keadaan lingkungan	0,804	0.858974
Hasilnya	0.645267	0.701203
<i>F-measure</i>	0,672073315	

Pada gambar 5.2, terlihat bahwa nilai *precision*-nya adalah 64.52%, *recall* 70.12% dan *f-measure* 67.20%. Nilai tersebut tidak mencapai 100%, hal ini dikarenakan kelas data (*tweet*) yang diprediksi sistem tidak sesuai dengan kelas data aktual. Ketidaksiuaian ini dipengaruhi oleh beberapa faktor, diantaranya adalah:

1. Bobot *term* sama di beberapa kelas

Tweet yang jumlah kemunculan *term*-nya sama di beberapa kelas, maka *tweet* tersebut tidak diklasifikasikan di kelas manapun dan dideninisikan sebagai *tweet* “*undefined*”.

Tabel 5.3 *Tweet Undefined* Faktor Bobot Sama

Tanggal	Tweet	Kecamatan	Kelas Aktual	Kelas Prediksi
Mon Feb 1 11:15:47	jadi pengasuh anak sementara hehe	Kiara_Condong	pekerjaan	undefined

Tweet di atas setelah melalui tahap *preprocessing* terdapat *term* “*asuh*” dan “*anak*”. Pada ontologi sendiri, kelas (parameter) pekerjaan terdapat *term* “*pengasuh*” sebagai jenis pekerjaan dan kelas keharmonisan keluarga terdapat *term* “*anak*” sebagai anggota keluarga, kedua *term* di ontologi tersebut setelah di-*lemmatization* akan menjadi “*asuh*” dan “*anak*”. Setelah itu dilakukan *term matching* antara *term* dari *tweet* dan *term* dari ontologi. *Term* dari *tweet* tersebut masuk di kedua kelas dan memiliki bobot kemunculan yang sama yaitu 1 (kelas pekerjaan: 1, kelas keharmonisan keluarga: 1), maka dari itu kelas prediksi pada *tweet* di atas adalah “*undefined*”.

2. Keterbatasan *term* di ontologi

Selain faktor bobot *term* yang sama dikedua kelas, yang menyebabkan suatu *tweet* tidak diklasifikasikan di kelas manapun (“*undefined*”) adalah, *tweet* tersebut tidak mengandung *term* yang ada di ontologi atau bobotnya 0 di semua kelas. Adapun *tweet* yang tidak diklasifikasikan ke kelas manapun karena faktor ini dapat dilihat pada tabel 5.4.

Tabel 5.4 *Tweet Undefined* Faktor Tidak ada Bobot

Tanggal	Tweet	Kecamatan	Kelas Aktual	Kelas Prediksi
Mon Feb 1 22:03:08	aku titipkan semua yang aku tinggalkan kamu harus bisa menjaga apa yang harus kau jaga kau permataku aku percaya padamu	Cidadap	hubungan sosial	undefined

Setelah melalui tahapan *preprocessing*, *tweet* di atas memiliki *term* “titip”, “tinggal”, “jaga”, “permata”, “percaya”. *Terms* tersebut tidak ada di *bag of words ontology* (keterbatasan *term* di ontologi), sehingga jumlah kemunculan *term*-nya adalah 0 di semua kelas pada ontologi. Hal inilah yang menyebabkan *tweet* tidak dapat diklasifikasikan ke dalam kelas di ontologi (“*undefined*”).

3. Proses *lemmatization* yang tidak sempurna

Tweet dan *bag of words ontology* dilakukan *lemmatization* agar semua *term* kembali ke kata dasar. Tujuan *lemmatization* untuk mencocokkan (*match*) antara *term* di *tweet* dan *term* di ontologi, selain itu *lemmatization* berfungsi untuk mengurangi dimensi kata yang sebenarnya memiliki arti yang sama. Jika *term* pada *tweet* atau ontologi tidak sempurna di-*lemmatization*, maka saat proses *term matching* tidak akan cocok, sehingga mempengaruhi bobot suatu *term*. Berikut adalah *tweet* yang diprediksi tidak sama dengan kelas aktual.

Tabel 5.5 Klasifikasi *Tweet* Faktor *Lemmatization*

Tanggal	Tweet	Kecamatan	Kelas Aktual	Kelas Prediksi
Tue Feb 2 00:18:51	orang lagi berduka atas musibah dan sedang merasakan sakit dia memuji yang bukan muhromnya kamu tidak berpikir ya	Cidadap	hubungan sosial	undefined

Tweet di atas setelah melalui proses *lemmatization* menghasilkan *terms* “orang”, “duka”, “atas”, “musibah”, “rasa”, “sakit”, “puji”, “muhromnya” dan “pikir”. *Terms* tersebut akan dicocokkan ke dalam *bag of words ontology*. *Terms* yang cocok adalah “orang” muncul 1 kali di kelas hubungan sosial dan 1 kali di kelas keharmonisan keluarga, *term* “sakit” muncul 2 kali di kelas kesehatan karena kelas kesehatan mengandung *term* “sakit” dan “penyakit”, jika di-*lemmatization* keduanya menjadi “sakit”, sehingga jumlah kemunculan *term* “sakit” di kelas kesehatan adalah 2, dan *term* “atas” muncul 1 kali di kelas pendapatan rumah tangga sebagai *instance* “menengah atas”. *Term* “rasa” seharusnya muncul di kelas hubungan sosial dalam ontologi, karena di kelas tersebut terdapat *term* “perasaan”, akan tetapi setelah dilakukan *lemmatization*, *term* “perasaan” berubah menjadi “asa”, sehingga tidak cocok dengan *term* “rasa” di *tweet*. Jika dijumlahkan semua, maka bobot nilai per masing-masing kelas adalah, kelas hubungan sosial mempunyai bobot 2, kelas kesehatan mempunyai bobot 2, kelas pendapatan rumah tangga mempunyai bobot 1, dan kelas keharmonisan keluarga mempunyai bobot 1. Dikarenakan bobot kelas hubungan sosial dan kelas kesehatan sama, yaitu 2 maka *tweet* tersebut tidak dapat diklasifikasikan ke dalam kelas di ontologi (“*undefined*”).

4. Tidak dapat mengenali frasa

Perhitungan *tweet* berbeda dengan perhitungan frasa. Jika suatu frasa dipotong per kata (*term*), menghasilkan makna yang berbeda. Selain itu pula, pemotongan frasa menjadi *term*, menyebabkan sifat *term* ini menjadi umum (dapat digunakan pada semua konteks kalimat atau *tweet*). Adapun pembuktian dari faktor ini adalah sebagai berikut.

Tabel 5.6 Klasifikasi *Tweet* Faktor Perhitungan *Term*

Tanggal	Tweet	Kecamatan	Kelas Aktual	Kelas Prediksi
Mon Feb 1 20:09:46	mudah mudah ada rezekinya buat anak	Kiara_Condong	pendapatan rumah tangga	kondisi keamanan

Tweet di atas mengandung *term* “mudah”, “ada”, “rezeki”, “anak” setelah dilakukan *preprocessing*. *Term* “mudah” terdapat di kelas kondisi keamanan dalam ontologi, *term* ini berasal dari nama *instance* “kemudahan akses” dari *subclass* pelaporan. Frasa “kemudahan akses” jika diletakkan dalam *bag of words ontology* akan menjadi “kemudahan” dan “akses”, lalu jika dilakukan *lemmatization* pada *bag of words ontology* tersebut, masing-masing *term* akan menjadi “mudah” dan “akses”. *Term* “mudah” dari *tweet* di atas muncul 2 kali pada kelas kondisi keamanan, *term* “ada” muncul di kelas keadaan lingkungan, *term* tersebut berasal dari *term* “keadaan” yang di-*lemmatization*. Sedangkan *term* “rezeki” muncul 1 kali di kelas pendapatan rumah tangga, dan *term* “anak” muncul 1 kali di kelas keharmonisan keluarga. Frasa “mudah mudahan” dalam *tweet* di atas memiliki arti sebenarnya “semoga”, jika frasa tersebut dipotong menjadi *term* “mudah”, maka akan memiliki makna lain dan sifat *term* “mudah” bisa digunakan dalam banyak konteks. Apabila jumlah kemunculan *term* per kelas ditotal, maka *tweet* di atas akan diklasifikasikan ke dalam kelas kondisi keamanan karena memiliki bobot paling besar.

6. Kesimpulan

Berdasarkan penelitian yang dilakukan dapat disimpulkan bahwa:

1. *Crawling* data Twitter berdasarkan wilayah 30 kecamatan kota Bandung dapat memanfaatkan API Twitter menggunakan atribut *geocode* yang terdiri dari *longitude*, *latitude* dan *radius*. Namun, sistem ini masih terdapat kelemahan dalam proses penentuan *radius* suatu wilayah. Jika proses tersebut tidak dilakukan dengan teliti, maka penentuan *radius* dapat beririsan dengan wilayah lain, mengingat bentuk *radius* suatu wilayah adalah *circle*, sedangkan bentuk suatu wilayah tak beraturan.
2. Kata-kata *slang* yang beragam dalam data dan sedikitnya kata dalam kamus *slang translation*, menyebabkan beberapa data tidak terkoreksi oleh sistem sehingga membutuhkan pemeriksaan secara manual.
3. Parameter *happiness* berdasarkan BPS Bandung dapat dibangun menggunakan pendekatan ontologi *top-down hierarchy*, dengan menyusun dan mengumpulkan *terms* yang berkaitan dengan parameter *happiness* tersebut.
4. Pendekatan ontologi dapat digunakan untuk mengklasifikasikan *tweet*, meskipun terdapat beberapa kekurangan seperti dalam analisis pengujian.
5. Hasil klasifikasi sentimen dan klasifikasi parameter *happiness* sangat mempengaruhi perhitungan *happiness*, hal ini dikarenakan hasil klasifikasi sentimen digunakan untuk menghitung tingkat kebahagiaan (*happiness*), sedangkan hasil klasifikasi parameter *happiness* digunakan untuk menghitung tingkat *happiness* per parameter. Jika tingkat akurasi (*precision*, *recall* dan *f-measure*) dalam sistem rendah, maka akurasi untuk perhitungan *happiness* terhadap data aktual juga rendah.
6. Penelitian ini menghasilkan perhitungan *happiness* seluruh kota Bandung yang berasal dari klasifikasi di dalam sistem sebesar 55.22%. Kecamatan yang paling bahagia adalah kecamatan Regol dengan nilai *happiness* 81.48%, dan kecamatan yang paling tidak bahagia adalah kecamatan Astanaanyar dengan nilai *happiness* sebesar 32.07%.

7. Daftar Pustaka

- [1] Ilhamnoor, "Survei Membuktikan, Warga Kota Bandung Bahagia," 14 Januari 2015. [Online]. Available: <http://infobandung.co.id/survei-membuktikan-warga-kota-bandung-bahagia/>. [Accessed 20 Oktober 2015].
- [2] A. D. Kramer, "An Unobtrusive Behavioral Model of "Gross National Happiness," *Gross National Happiness*, pp. 287-290, 2010.
- [3] G. Antoniou and F. v. Harmelen, *A Semantic Web Primer*, London: The MIT Press Cambridge, 2003.

- [4] M. Casanova, K. Breitman and W. Truszkowski, "Semantic Web: Concepts, Technologies and Applications," no. 3, pp. 155-173, 2007.
- [5] C. L. H. P. S. M. Haewoon Kwak, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010.
- [6] B. F. S, "[Pengenalan] Apa Itu Twitter API dan Pembuatan Consumer Key dan Consumer Secret ?," 2014. [Online]. Available: http://jagocoding.com/tutorial/427/Pengenalan_Apa_Itu_Twitter_API_dan_Pembuatan_C%20onsumer_Key_dan_Consumer_Secret.%20. [Accessed 2016 Januari 21].
- [7] T. I. M. Tokunaga, "Text categorization based on weighted inverse document frequency," *Special Interest Groups and Information-Process Society of Japan (SIG-IPSJ)*, 1994.
- [8] C. D. P. R. H. S. Manning, "Introduction to information retrieval," Cambridge, Cambridge university press, 2008.
- [9] M. e. a. Lan, "Supervised and traditional term weighting methods for automatic text categorization," *Pattern Analysis and Machine Intelligence*, pp. 721-735, 2009.
- [10] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, 2012.
- [11] Y. Yukai , L. Yang, Y. Yongging, X. Hong, L. Weiming, L. Zhao and C. Xiaoyun, *K-SVM : An Effective SVM Algorithm Based on K-Means Clustering*, Academy Publisher, 2013.
- [12] N. J. S.-T. Cristianini, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.
- [13] P. Kristina, "Klasifikasi Dokumen Tumbuhan Obat Menggunakan Algoritma KNN Fuzzy," Institut Pertanian Bogor, Bogor, 2011.
- [14] F. e. a. Wang, "A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing," *Journal of Shanghai Jiaotong University (Science)* 20, pp. 44-50, 2015.