Kategorisasi Topik Tweet di Kota Jakarta, Bandung, dan Makassar dengan Metode Multinomial Naïve Bayes Classifier

Tweet Topic Categorization in Jakarta, Bandung, and Makassar with Multinomial Naïve Bayes Classifier

Muhammad Haerunnur Syahnur[1], Moch. Arif Bijaksana [2], Mohamad Syahrul Mubarok[3]
[1,2,3]Fakultas Informatika Universitas Telkom, Bandung
[1]haerunn@gmail.com, [2]arifbijaksana@gmail.com, [3]milo.mubarok@gmail.com

Abstrak
Twitter merupakan mikroblog yang sedang diminati oleh banyak orang di berbagai negara. Para penggunanya dapat memposting segala hal mengenai apa yang sedang terjadi di sekitar mereka. Baik itu *breaking news*, berita gosip selebriti, dan lain sebagainya.
Setiap harinya, tidak kurang dari 500 juta *tweet* dikirimkan oleh penggunanya dari seluruh penjuru dunia. Dari *tweet-tweet* tersebut, *Twitter* akan mendaftar topik mana saja yang sering dibicarakan secara *real-time* melalui fitur *Trending Topic.* Namun, pada kenyataannya tren topik tersebut mengacu pada banyak, dan seringnya suatu kata, atau frase dikicaukan oleh pengguna di berbagai lokasi mengenai suatu *event* besar yang sedang terjadi.
Penelitian ini bertujuan untuk mengklasifikasikan *tweet-tweet* yang dikirimkan dari kota Jakarta, Bandung, dan Makassar ke dalam beberapa topik, kemudian melihat tingkat kepopuleran topik tersebut di berbagai paruh waktu. Penelitian ini diawali dengan pengumpulan data *tweet* dari twitter.com, kemudian dilakukan praproses. Dilanjutkan ke tahap pembentukan *classifier* memanfaatkan data latih. Tahap terakhir adalah klasifikasi *tweet* ke dalam parameter topik tertentu menggunakan metode Multinomial Naïve Bayes Classifier. Penelitian ini menghasilkan rata-rata *f1-score* yang cukup tinggi, yaitu 76,81% jika dibandingkan dengan *classifier* lain seperti Random Forest, SVM, dan Nearest Neighbor (75,18%, 68,82%, dan 50,22%).

Kata kunci: Klasifikasi topik, Twitter, Jakarta, Bandung, Makassar, Multinomial Naïve Bayes Classifier

1. Pendahuluan

Twitter merupakan mikroblog yang sedang diminati oleh banyak orang di berbagai negara. Para penggunanya (*tweeps*) dapat memposting segala hal mengenai apa yang sedang terjadi di sekitar mereka. Baik itu *breaking news*, berita gosip selebriti, dan lain sebagainya.

Jumlah *tweet* rata-rata yang dikirim tiap harinya adalah sekitar 500 juta. Atau setara dengan 350 ribu *tweet* per menit, atau 6000 *tweet* per detik [1]. *Twitter* memiliki satu fitur yang disebut *Trending Topic*, yaitu suatu daftar *real-time* mengenai topik yang sedang populer/ dibicarakan banyak orang [2]. Daftar *Trending Topic* ini selalu berubah seiring waktu, dan cenderung mengikuti *event* besar yang sedang terjadi pada saat itu. Meskipun dinamakan "*Trending Topic*", sebenarnya fitur ini mendeteksi teks/ string yang sering muncul dalam *tweet*. Sedangkan pada kenyataannya, suatu topik yang sedang populer dibahas oleh para pengguna tidaklah harus selalu mengandung suatu kata/ string yang sama dengan *tweet* lain. Misalnya, untuk topik olahraga, satu *tweeps* bisa saja membahas tentang balap motor di Sepang, sedangkan *tweeps* lain membahas tentang sepak bola Liga Inggris. Sehingga, secara teori, *Trending Topic* tidak bisa selalu menjadi acuan mengenai topik yang sedang populer dibahas oleh para pengguna.

Tren topik di Twitter cenderung berubah dari waktu ke waktu menyesuaikan keseharian penggunanya. Dari sini penulis berasumsi bahwa terdapat kategori tertentu yang sering menjadi obrolan para pengguna *Twitter* pada tiap kurun waktu, meskipun tidak ada kejadian besar yang terjadi pada saat itu. Pada pagi hari, misalnya, kebanyakan *tweeps* akan memposting mengenai keadaan lalu lintas, sedangkan pada siang hari akan lebih sering membicarakan tentang makan siang.

Dengan mengetahui perubahan tren topik tersebut, seseorang dapat memanfaatkannya untuk kepentingan pemasaran, atau bahkan untuk kampanye politik. Untuk itulah penulis akan melakukan suatu penelitian yang bertujuan untuk mengklasifikasi tren topik *Twitter* tersebut berdasarkan kurun waktu tertentu. Pemilihan ketiga kota tersebut didasarkan pada asumsi penulis yang beranggapan bahwa adanya potensi bisnis, atau aspek lainnya yang bisa didapatkan dengan melihat persebaran, dan tren topik pembicaraan di ketiga kota tersebut yang didasarkan pada fakta bahwa kota Jakarta merupakan pusat pemerintahan negara Indonesia, sehingga segala sesuatu yang terjadi pada kota Jakarta akan sangat memengaruhi berbagai aspek di Indonesia. Untuk kota Bandung, sejak dipimpin oleh Ridwan Kamil kota Bandung mulai berbenah, dan meraih banyak pencapaian [3]. Dengan mengetahui topik pembicaraan warga Bandung melalui Twitter, sedikit banyak dapat menjadi bahan pembelajaran bagi pemerintah kota Bandung, ataupun pemerintah daerah lainnya dalam meningkatkan kinerja, dan pelayanan masyarakat. Sedangkan kota Makassar, kota ini adalah kota terbesar di kawasan Indonesia Timur

[4], dan juga menjadi titik pusat perekonomian, dan perdagangan di Indonesia Timur sejak jaman kerajaan [5], sehingga dengan mengamati tren topik pembicaraan warga Makassar di Twitter dapat sedikit banyak bermanfaat bagi pemerintah kota dalam memperbaiki berbagai aspek kehidupan, baik itu ekonomi, maupun aspek lainnya. Merujuk pada penelitian terkait yang menetapkan 18 kategori untuk tren topik [6], penulis akan menyusutkan jumlah tersebut menjadi hanya 13 kategori, yaitu: *berita, dan peristiwa, bisnis, liburan, dan perjalanan, teknologi, kesehatan, televisi, pendidikan, politik, dan pemerintahan, keagamaan, olahraga, cinta, dan romansa, musik,* dan *kuliner*. Untuk kurun waktu, akan diklasifikasikan menjadi tren topik pada pagi hari (pukul 04:00-10:00), siang hari (pukul 10:00-14:00), sore hari (pukul 14:00-16:30), petang hari (pukul 16:00-18:30) dan malam hari (pukul 18:30-04:00) [7].

Penelitian ini diawali dengan pengumpulan data *tweet* berdasarkan lokasi pada kurun-kurun waktu yang telah ditetapkan. Kemudian dilanjutkan ke praproses, lalu pembuatan pola berdasarkan data yang telah dipelajari. Terakhir, pengklasifikasian berdasarkan pola tersebut menggunakan metode Multinomial Naïve Bayes. Berdasarkan penelitian serupa [6], penggunaan klasifikasi dengan metode ini menghasilkan akurasi yang cukup tinggi, yaitu 65% hingga 70%. Selain itu, berdasarkan Dan Jurafsky (Stanford University), metode Naïve Bayes Classifier merupakan metode yang unggul dalam hal *robustness,* yaitu dapat mengenali kata spesifik yang berkorelasi erat terhadap suatu kategori, karena menggunakan model probabilistik yang dapat memperlihatkan perbedaan antar kata dengan jelas, sehingga dapat mengklasifikasi suatu dokumen uji dengan tingkat akurasi cukup tinggi.

2. Dasar Teori

2.1. Kategorisasi Tweet

Beberapa penelitian mengenai kategorisasi *tweet* sudah dilakukan beberapa kali. Seperti yang dilakukan oleh Kathy Lee et al. yang melakukan kategorisasi *trending topic* Twitter yang diklasifikasikan ke dalam 18 kategori, yaitu: *arts & design*, *books*, *business*, *charity & deals*, *fashion*, *food & drink*, *health*, *holidays & dates*, *humor*, *music*, *politics*, *religion*, *science*, *sports*, *technology*, *tv & movies*, *other news,* and *other* menggunakan pendekatan *Bag of Words,* dan *Network Based Classification* [6].

Sriram et al. [8] mengklasifikasikan tweet ke dalam beberapa kelas umum, seperti *news*, *events*, *opinions*, *deals,* dan *private messages* berdasarkan *author information*, dan fitur *domain-specific* yang diekstrak dari *tweet*, seperti adanya pemendekan kata, dan *slang*, frase waktu-kejadian, penekanan pada kata-kata, mata uang, dan persentase tanda-tanda, "@username" pada awal tweet, dan "@username" dalam *tweet* [6].

Genc et al. [9] memperkenalkan teknik klasifikasi berbasis Wikipedia. Penulis mengklasifikasikan *tweet* dengan melakukan pemetaan pesan ke halaman Wikipedia yang paling mirip dengan *tweet* tersebut, dan menghitung jarak semantik antar pesan berdasarkan jarak antara halaman-halaman Wikipedia yang terdekat [6].

2.2. Twitter

Twitter (/ twɪtər /) adalah layanan jejaring sosial online yang memungkinkan pengguna untuk mengirim dan membaca pesan dalam 140 karakter yang disebut "tweets".

Twitter dibuat pada bulan Maret 2006 oleh Jack Dorsey, Evan Williams, Biz Stone dan Nuh Kaca dan diluncurkan pada bulan Juli 2006. Layanan ini dengan cepat mendapatkan popularitas di seluruh dunia, dengan lebih dari 100 juta pengguna yang pada 2012 memposting 340 juta tweet per hari [10]. Pada 2013, Twitter adalah salah satu dari sepuluh situs yang paling sering dikunjungi, dan telah digambarkan sebagai "SMS internet." [11] [12] Pada Mei 2015, Twitter memiliki lebih dari 500 juta pengguna, yang terdiri dari 302 juta pengguna aktif [13].

Twitter memiliki satu fitur yang berfungsi untuk mendaftar topik-topik yang sedang hangat dibicarakan orang pada waktu tertentu yang disebut *Trending Topic*. Seringkali, topik-topik tersebut diawali dengan *hashtag sign* (#) [14], seperti #BoyRevaAnakJalanan #HateSpeech yang merupakan salah satu tren topik *Worldwide* pada hari Selasa, 3 November pukul 18:10 WIB.

Twitter sering digunakan untuk memposting keseharian oleh para penggunanya. Namun, bisa juga digunakan untuk kampanye politik seperti kampanye I Stand On the Right Side yang dilakukan para relawan Jokowi pada pemilu tahun 2014 lalu [15].

Hal yang sama telah dilakukan oleh Presiden Amerika Serikat, Barack Obama, pada pemilihan presiden Amerika tahun 2008 lalu. Melalui akun Twitter-nya, @BarackObama, ia melakukan kampanye politiknya [16].

Selain itu, Twitter juga sering digunakan untuk *marketing* atau pemasaran suatu produk. Hal ini bisa dilihat dari banyaknya akun-akun *branding* yang bermunculan di Twitter, seperti @triindonesia, @indosat, @telkomsel, yang sering memposting *tweet* mengenai produk mereka.

2.3. Praproses

Preprocessing atau praproses adalah strategi dan teknik yang saling berkaitan untuk membuat data lebih mudah atau cocok untuk digunakan dalam proses mining [17]. Praproses bertujuan untuk meningkatkan hasil analisis terkait masalah waktu, biaya dan kualitas. Terdapat tiga tahapan praproses data dalam penelitian ini yaitu:

1. *Tokenization* adalah pemotongan karakter, dan sebuah set dokumen yang diberikan menjadi potongan-potongan kata atau karakter yang sesuai dengan kebutuhan sistem. Potongan-potongan tersebut dikenal dengan istilah token [18].
2. *Stop Word Removal* adalah tahap penghapusan kata-kata yang kurang relevan dalam penentuan topik sebuah dokumen, misal "and", "or", "the" [19]. Atau contohnya dalam bahasa Indonesia adalah "dan", "dia", "sebuah".

Stemming adalah tahapan pengubahan suatu kata menjadi akar atau kata dasarnya dengan cara menghilangkan imbuhan awal atau akhir pada kata tersebut, misal "eating" menjadi "eat" setelah melalui proses *Stemming* [20].

2.4. Algoritma Nazief-Adriani

Konsep *stemming* dengan algoritma Nazief dan Adriani dijelaskan dalam sebuah laporan teknis yang tidak dipublikasikan dari Universitas Indonesia pada tahun 1996. Algoritma ini mengacu kepada aturan morfologi bahasa Indonesia yang mengelompokkan imbuhan, yaitu imbuhan yang diperbolehkan atau imbuhan yang tidak diperbolehkan. Pengelompokkan ini termasuk imbuhan di depan (awalan), imbuhan kata dibelakang (akhiran), imbuhan kata di tengah (sisipan), dan kombinasi imbuhan pada awal dan akhir kata (konfiks). Algoritma ini menggunakan kamus kata keterangan yang digunakan untuk mengetahui bahwa proses *stemming* telah mendapatkan kata dasar [21].

2.5. Bag of Words

Bag of Words adalah representasi sederhana yang digunakan dalam *Natural Language Processing* dan *Information Retrieval* (IR). Dalam model ini, sebuah teks (seperti kalimat, atau dokumen) direpresentasikan sebagai kumpulan (multiset) dari kata-katanya, mengabaikan tata bahasa, dan bahkan urutan kata tetapi menjaga keanekaragaman [22].

2.6. API Twitter

API atau yang biasa disebut *Application Programming Interface* adalah suatu program/ 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 [23].

Untuk mengakses fasilitas yang disediakan Twitter, pengembang memerlukan *Consumer Key* dan *Consumer Secret* yang 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.

2.7. Multinomial Naïve Bayes Classifier

Multinomial Naïve Bayes Classifier merupakan metode klasifikasi dengan pembelajaran *supervised* yang menggunakan model probabilistik. Probabilitas suatu dokumen berada di suatu kelas dapat dirumuskan sebagai berikut [18]:

$$P(C|D) \propto P(C) \prod_{1 \le k \le n} P(t_k|C) \quad (1)\ [18]$$

Persamaan di atas dilakukan sebanyak jumlah kategori/ kelas untuk mendapatkan *posterior probability* masing-masing dokumen untuk suatu kelas. Kemudian, semua *posterior probability* kategori tersebut akan dibandingkan satu sama lain untuk mendapatkan *posterior probability* tertinggi yang akan menjadi kategori/ kelas dari dokumen yang sedang diuji dengan menggunakan persamaan sebagai berikut:

$$C_{MAP} = \arg\max_{c \in C} P(C|D) = \arg\max_{c \in C} P(C) \prod_{1 \le k \le n} P(t_k|C) \quad (2)\ [18]$$

$P(C)$ pada persamaan 2.1 merupakan probabilitas suatu kategori untuk terjadi dibandingkan dengan keseluruhan kategori. Berikut ini adalah rumus *prior probability:*

$$P(C) = \frac{N_c}{N} \quad (3)\ [18]$$

Keterangan:
P(c) : probabilitas kategori c
N_c : jumlah kategori c
N : jumlah seluruh kategori

Sedangkan $P(t|c)$ merupakan probabilitas kondisional untuk *term* t yang terjadi pada dokumen *class c.* Persamaan *likelihood probability* adalah sebagai berikut:

$$P(t|c) = \frac{count(t,c) + 1}{count(c) + |V|} \quad (4)$$

Keterangan:

$P(t|c)$: probabilitas *term* t pada kategori c
$count(t,c)$: jumlah *term* t pada kategori c
$count(c)$: jumlah *term* pada kategori c
$|V|$: jumlah seluruh kosa kata

2.8. Evaluasi Sistem

Pada tahap evaluasi sistem, akan dilakukan pengujian dan analisis performansi sistem dengan melakukan perhitungan menggunakan tabel kontingensi seperti Tabel 2.1:

Tabel 2.1 Tabel Kontingensi Evaluasi Sistem

#	***Relevant***	***Not Relevant***
Retrieved	*True Positive(TP)*	*False Positive(FP)*
Not Retrieved	*False Negative(FN)*	*True Negative(TN)*

Keterangan:
True Positive (TP) : jumlah kelas positif (kelas yang menjadi fokus klasifikasi) dan terklasifikasi dengan benar sebagai kelas positif oleh sistem.
True Negative (TN) : jumlah kelas negatif (kelas selain kelas yang menjadi fokus klasifikasi) dan terklasifikasi dengan benar sebagai kelas negatif oleh sistem
False Positif (FP) : jumlah kelas negatif yang terklasifikasi sebagai kelas positif oleh sistem.
False Negative (FN) : jumlah kelas positif yang terklasifikasi sebagai kelas negatif oleh sistem.

Pengukuran performansi sistem akan dilakukan dengan menghitung akurasi *precision, recall* dan *f-measure* [24].

Precision adalah rasio jumlah dokumen relevan dengan total jumlah dokumen yang ditemukan oleh *classifier.*

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Recall adalah rasio jumlah dokumen dokumen yang ditemukan kembali oleh *classifier* dengan total jumlah dokumen yang relevan.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F-Measure adalah kombinasi rata-rata harmonik dari *precision* dan *recall* yang berbanding lurus dengan nilai keduanya.

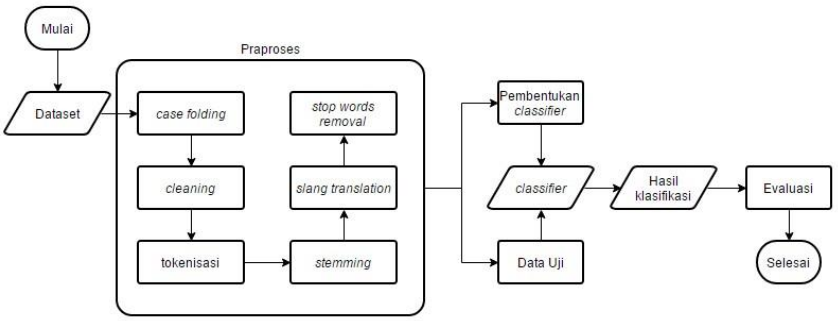
$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (7)$$

3. Perancangan Sistem

Data dalam penelitian ini merupakan *tweet* berbahasa Indonesia yang dikirimkan dari tiga wilayah penelitian (Jakarta, Bandung, dan Makassar) dalam kurun waktu tertentu. Data tersebut akan dibagi menjadi dua, menjadi data latih dan data uji. Sistem yang akan dibuat terdiri dari beberapa proses, yaitu sebagai berikut:

1. Praproses data dengan cara *case folding, cleaning, tokenization, stop word removal, slang translations* dan *stemming.*
2. Pembentukan *classifier* menggunakan data latih.
3. Pengklasifikasian data uji.
4. Pengujian/ evaluasi sistem.

Gambar 3.1 di bawah ini adalah alur sistem secara umum:



Gambar 3.1 Alur Sistem secara Umum

Penelitian ini dimulai dengan pengumpulan data yang memanfaatkan API Twitter. Beberapa data yang menjadi data latih dicari secara manual berdasarkan kata yang berkorelasi dengan kategori tersebut. Misalnya, kata "KPK", "UU", dan "KORUPTOR" pada *tweet* "Wahai rakyat Indonesia, bangunlah..sadarlah...KPK akan dilemahkan...Ayo Tolak Revisi UU KPK demi keutuhan NKRI dari penjahat2 KORUPTOR..." memiliki korelasi yang erat dengan kategori "politik_pemerintahan".

Data yang telah terkumpul akan dibersihkan/ mengalami praproses terlebih dahulu. Berikut ini adalah tahap praproses data:

a) *Case folding*, yaitu mengubah semua karakter menjadi huruf kecil.
b) *Cleaning*, yaitu pembersihan simbol, dan karakter bukan huruf dari *tweet*, termasuk *hashtag*, dan *mentions.*
c) *Tokenization*, yaitu proses memotong kalimat pada tiap dokumen berdasarkan menjadi potongan kata, atau karaker.
d) *Stemming*, yaitu proses mengembalikan kata pada dokumen ke bentuk awalnya. Proses ini menggunakan algoritma Nazief-Adriani. Beberapa *tweet* harus melalui proses *stemming* manual, karena tidak dapat ditangani oleh algoritma tersebut. Kata-kata yang harus di-*stemming* manual adalah kata-kata seperti: "jalan2", atau "penjahat2".
e) *Slang translation*, yaitu proses mengubah kata-kata ragam cakap ke dalam bentuk baku. Proses ini menggunakan daftar kata yang dibuat dari hasil pembelajaran data di awal penelitian. Kata ragam cakap yang dimaksud adalah kata-kata seperti, "gatau" yang berarti "tidak tahu", atau "elo" yang berarti "kamu". Beberapa kata yang masih belum terkoreksi dari proses ini akan melalui proses translasi manual.
f) *Stop words removal*, yaitu proses membersihkan data dari kata-kata yang tidak relevan. Proses ini menggunakan daftar *stop words* Tala [25].

Semua data yang telah mengalami praproses akan diberi label secara manual (*hand labeled*), kemudian disimpan di basis data.

Data latih akan dipelajari oleh sistem dengan cara menghitung *prior probability* tiap kategori, dan *likelihood probability* tiap kata tiap kategori. Karena data latih yang seimbang di tiap kategori (50 data), maka *prior probability* untuk setiap kategori adalah sama, yaitu: 0.076923, yang merupakan hasil dari:

$$\text{[illegible]} = \frac{50}{650} = 0.0769 \qquad (8)$$

Untuk mendapatkan informasi mengenai *likelihood probability*, sistem terlebih dahulu akan menghitung jumlah kosa kata (*vocabulary*) pada keseluruhan data yang terkumpul, juga jumlah kosa kata pada masing-masing kategori. Informasi-informasi tersebut selanjutnya akan disimpan di basis data.

Pada penelitian ini terdapat 2543 kosa kata, sedangkan untuk jumlah kata pada tiap kategori dapat dilihat pada Tabel 3.1 di bawah ini:

Tabel 3.1 Jumlah Kata per Kategori

Kategori	Jumlah Kata
cinta_romansa	235
keagamaan	308
kesehatan	326
olahraga	331
musik	338
politik_pemerintahan	349
kuliner	349

teknologi	364
pendidikan	371
liburan_perjalanan	378
bisnis	385
berita_peristiwa	391
televisi	398

Untuk menentukan kategori prediksi, tiap kata pada tiap data uji akan dicari *likelihood probability*-nya. *Dot product* dari probabilitas tiap kata tersebut kemudian akan dikalikan dengan *prior probability* kategori yang sedang ditinjau untuk mendapatkan probabilitas dokumen terhadap kategori (*posterior probability*). Langkah ini akan diulang sebanyak 13 kali; sejumlah kategori. *Posterior probability* untuk semua kategori akan saling dibandingkan, dan probabilitas tertinggi yang akan menjadi kategori data uji tersebut.

Tabel 3.2 adalah contoh proses klasifikasi kategori *tweet* oleh sistem yang telah dibangun:

Tabel 3.2 Contoh Klasifikasi Sistem

No.	Deskripsi	Hasil
1.	*Tweet* asli	Wahai rakyat Indonesia, bangunlah..sadarlah...KPK akan dilemahkan...Ayo Tolak Revisi UU KPK demi keutuhan NKRI dari penjahat2 KORUPTOR...
2.	*Tweet* telah melalui praproses	rakyat indonesia bangun sadar kpk lemah ayo tolak revisi uu kpk utuh nkri jahat jahat koruptor
3.	Pembobotan masing-masing kata tiap kategori (contoh kategori: politik_pemerintahan)	Array ([rakyat] => 0.0017289073305671 [indonesia] => 0.00069156293222683 [bangun] => 0.00034578146611342 [sadar] => 0.00034578146611342 [kpk] => 0.011756569847856 [lemah] => 0.00069156293222683 [ayo] => 0.00034578146611342 [tolak] => 0.0027662517289073 [revisi] => 0.0031120331950207 [uu] => 0.0020746887966805 [nkri] => 0.00034578146611342 [jahat] => 0.00069156293222683 [koruptor] => 0.00069156293222683)
4.	Penghitungan total bobot tiap kategori (contoh kategori: politik_pemerintahan)	P (dokumen | politik_pemerintahan) = 0.076923076923077 * 1.1870660893134E-39 = 9.1312776101028E-41

4. Pembahasan

Tabel 4.1 adalah hasil pengujian dari sistem yang telah dibuat, dan Tabel 4.2 merupakan hasil Pengujian Pengenalan Kata Kunci:

Tabel 4.1 Hasil Klasifikasi Sistem

Kategori	*Recall*	*Precision*	*f1-measure*
keagamaan	93%	76%	83.64%
bisnis	77%	87%	81.70%
politik_pemerintahan	82%	79%	80.47%
cinta_romansa	77%	84%	80.35%
kuliner	77%	83%	79.89%
berita_peristiwa	76%	84%	79.80%
teknologi	79%	79%	79.00%
televisi	70%	90%	78.75%
olahraga	75%	81%	77.88%
pendidikan	74%	70%	71.94%
musik	80%	64%	71.11%
kesehatan	72%	69%	70.47%
liburan_perjalanan	67%	50%	57.26%
Rata-rata	**77%**	**76.62%**	
f1-measure	**76.73%**		

Tabel 4.2 Pengujian Pengenalan Kata Kunci

Dokumen Uji (setelah praproses)	Kata kunci	Kategori Prediksi	
		Kata Kunci Tidak Dihapus	Kata Kunci Dihapus
silakan pesan pembuatan atribut	pesan	bisnis	unindentified
kalau rindu ini api mungkin aku sudah lama jadi abu	rindu	cinta_romansa	unindentified
jadikan sabar salat sebagai penolongmu	sabar, salat	keagamaan	berita_peristiwa
di luar hujan deras dan induk semang datang pakai payung bawakan jagung rebus masih panas panas nikmat anak kos mana lagi yang kamu dustakan	jagung, rebus, nikmat	kuliner	berita_peristiwa
pemerintah gandeng google guna meningkatkan ekonomi digital indonesia	google	teknologi	pendidikan
asik nonton syuting di pasar minggu jakarta selatan lihat kak voke deh yey	nonton	televisi	liburan_perjalanan
mantap itu di gunung mana itu kalau boleh tahu	gunung	liburan_perjalanan	unindentified
exit tol rawamangun padat imbas volume lalin arteri ts	tol, lalin, padat	berita_peristiwa	unindentified

Dari proses pengujian sistem yang menggunakan 601 data dapat dianalisis beberapa hal sebagai berikut:

1. Dengan melihat grafik tren topik populer secara umum berdasarkan paruh waktu, dapat dianalisis bahwa saat pihak tertentu ingin melakukan sesuatu yang berkaitan dengan topik "cinta_romansa", misalnya memasarkan novel *genre* cinta, maka sebaiknya dilakukan pada malam hari dibandingkan pada pagi hari di mana fokus para pengguna akan tertuju ke topik tersebut, sedangkan di pagi hari fokus para pengguna Twitter akan terbagi ke dalam tiga topik populer selain "cinta_romansa", seperti "berita_peristiwa", dan "olahraga".
2. Rendahnya *precision* pada suatu kategori dapat disebabkan karena menonjolnya *likelihood probability* suatu kata pada suatu data uji yang bukan merupakan kategori data uji tersebut, sehingga terjadi kesalahan klasifikasi/ misklasifikasi. Seperti pada data uji "bandung mulai gelap" yang berkategori "berita_peristiwa", namun diklasifikasikan sebagai "liburan_perjalanan" oleh sistem, karena *likelihood probability* kata "bandung" untuk kategori "liburan_perjalanan" lebih tinggi dibandingkan pada kategori "berita_peristiwa" (0.00205:0.00068).
3. Keseimbangan *prior probability* suatu kategori juga akan memengaruhi hasil klasifikasi suatu dokumen. Semakin tinggi *prior probability* suatu kategori, maka kesempatan suatu dokumen untuk diklasifikasikan ke kategori tersebut makin tinggi. Namun, di penelitian ini hal tersebut tidak terlihat, karena *prior probability* yang seimbang.
4. Model probabilistik metode Multinomial Naïve Bayes Classifier memungkinkan sistem untuk dapat mengenali "kata kunci" pada suatu kategori. "Kata kunci" pada tiap kategori tersebut adalah kata-kata yang memiliki *likelihood probability* yang menonjol dibandingkan dengan suatu kata terhadap suatu kategori lainnya, yang berarti jika kata tersebut tidak ada pada suatu kategori, maka kemungkinan untuk misklasifikasi akan lebih tinggi.

5. Kesimpulan

Dari penelitian yang telah dilakukan, dapat ditarik kesimpulan sebagai berikut:

1. Dengan metode Multinomial Naïve Bayes Classifier, dan fitur *Bag of Words* dapat mencapai *f1-measure* rata-rata yang cukup tinggi, yaitu 77% (dibandingkan dengan Random Forest, SVM, dan Nearest Neighbor Classifier).
2. Menonjolnya *likelihood probability* suatu kata dalam suatu kategori sangat mempengaruhi klasifikasi sistem. Hal ini dapat dilihat pada *tweet* "Bandung mulai gelap" yang semestinya berkategori "berita_peristiwa", tapi diprediksi sebagai "liburan_perjalanan", karena kata "Bandung" pada kategori

"liburan_perjalanan" memiliki *likelihood probability* lebih tinggi dibandingkan pada kategori "berita_peristiwa" (0.00205:0.00068).

3. Dilihat dari persamaan Naïve Bayes, maka *prior probability* pun akan memengaruhi klasifikasi sistem. Suatu data uji akan cenderung diklasifikasikan ke kategori dengan *prior probability* yang lebih tinggi.
4. "Kata kunci" pada tiap kategori merupakan kata-kata yang memiliki *likelihood probability* tertinggi terhadap kategori tersebut. Saat "kata kunci" tersebut tidak ditemukan pada suatu dokumen uji, maka kemungkinan misklasifikasi kategori oleh sistem akan meningkat.
5. 14 dari 15 grafik distribusi tren topik keluaran sistem menghasilkan topik populer yang sama dengan topik populer pada data sebenarnya, sehingga grafik distribusi tren topik tersebut dapat diandalkan untuk mengamati topik populer tiap kota tiap paruh waktu.

6. Daftar Pustaka

[1] "Twitter Usage Statistics," [Online]. Available: http://www.internetlivestats.com/twitter-statistics/.

[2] Twitter, "FAQs about trends on Twitter," [Online]. Available: https://support.twitter.com/articles/101125#.

[3] D. Ramdhani, "Kompas.com," [Online]. Available: http://regional.kompas.com/read/2015/09/07/13102011/2.Tahun.Pimpin.Bandung.Ridwan.Kamil.Pamerkan.Prestasi. [Accessed 21 June 2016].

[4] A. Rasyid, "20 KOTA TERPENTING DI INDONESIA," 23 Janury 2016. [Online]. Available: https://ahsanrasyid.wordpress.com/2016/01/23/20-kota-terpenting-di-indonesia/. [Accessed 21 June 2016].

[5] P. Vicka, "Infrastruktur Memadai, Sulsel Kini jadi Pintu Gerbang Indonesia Timur," 23 October 2014. [Online]. Available: http://ekonomi.metrotvnews.com/read/2014/10/23/309171/infrastruktur-memadai-sulsel-kini-jadi-pintu-gerbang-indonesia-timur. [Accessed 21 June 2016].

[6] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal and A. Choudhary, "Twitter trending topic classification," in *2011 IEEE 11th International Conference*, 2011.

[7] I. Lanin, "Kapan menyebut selamat pagi, siang, sore, dan malam?," 16 Agustus 2014. [Online]. Available: http://tanja.portalbahasa.com/kapan-menyebut-selamat-pagi-siang-sore-dan-malam/. [Accessed 26 Februari 2016].

[8] B. Sriram, D. Fuhry, E. Demir and H. Ferhatosmanoglu, "Short text classification in twitter to improve information filtering," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

[9] Y. S. Yegin Genc and J. V. Nickerson, "Discovering context: Classifying tweets through a semantic transform based on wikipedia," in *Proceedings of HCI International*, 2011.

[10] Twitter, "Twitter turns six," 21 March 2012. [Online]. Available: https://blog.twitter.com/2012/twitter-turns-six.

[11] Alexa, "The top 500 sites on the web," [Online]. Available: http://www.alexa.com/topsites.

[12] L. D'Monte, "Swine flu's tweet tweet causes online flutter," 29 April 2009. [Online]. Available: http://www.business-standard.com/article/technology/swine-flu-s-tweet-tweet-causes-online-flutter-109042900097_1.html.

[13] P. Quintaro, "Twitter MAU Were 302M For Q1, Up 18% YoY," 28 April 2015. [Online]. Available: http://www.benzinga.com/news/earnings/15/04/5452400/twitter-mau-were-302m-for-q1-up-18-yoy.

[14] V. Doctor, "What Do Twitter Trends Mean?," 15 June 2012. [Online]. Available: https://www.hashtags.org/platforms/twitter/what-do-twitter-trends-mean/.

[15] Jatafest, "[AYO IKUT] Dukung Jokowi-Ernest Prakarsa & Joko Anwar Pasang Ava #StandOnTheRightSide," 6 June 2014. [Online]. Available: http://www.kaskus.co.id/thread/53918c148907e7b83e8b4651/ayo-ikut-dukung-jokowi-ernest-prakarsa-amp-joko-anwar-pasang-ava-standontherightside/.

[16] C. McGuinness, "The Social Seer," 30 May 2012. [Online]. Available: http://www.socialseer.com/2012/05/30/twitter-secrets-of-the-obama-campaign-introduction/. [Accessed 5 November 2015].

[17] S. Garcia, J. Luengo and F. Herrera, Data Preprocessing in Data Mining, Cham: Springer, 2015.

[18] C. D. Manning, P. Raghavan and H. Schutze, Introduction to Information Retrieval, Online Edition ed., London: Cambridge University Press, 2009.

[19] M. Bramer, Principles of Data Mining, London: Springer, 2007.

[20] N. Hardeniya, NLTK Essentials, Birmingham: Packt Publishing , 2015.

[21] B. A. Nazief and M. Adriani, "Confix Stripping: Approach to Stemming Algorithm for Bahasa Indonesia," *Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta,* 1996.

[22] Wikipedia, "Bag-of-words model," [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model. [Accessed 20 Mei 2016].

[23] S. F. Bramanda, "[Pengenalan] Apa Itu Twitter API dan Pembuatan Consumer Key dan Consumer Secret ?," 30 Juni 2014. [Online]. Available: http://jagocoding.com/tutorial/427/Pengenalan_Apa_Itu_Twitter_API_dan_Pembuatan_Consumer_Key_dan_Consumer_Secret. [Accessed 20 Mei 2016].

[24] F. Guillet and H. J. Hamilton, Quality Measures in Data Mining, Berlin: Springer, 2007.

[25] F. Z. Tala, A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands, 2003.