Pembentukan *Dataset* Topik Kata Bahasa Indonesia pada Twitter Menggunakan TF-IDF & *Cosine Similarity*

http://dx.doi.org/10.28932/jutisi.v4i3.862

Kristian Adi Nugraha[#1], Danny Sebastian[#2]

[#1,2]*Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana*

Jl. Dr. Wahidin Sudirohusodo 5-25 Yogyakarta 55224

[1]adinugraha@ti.ukdw.ac.id

[2]danny.sebastian@staff.ukdw.ac.id

Abstract — **Social media is evidently the most popular platform compared to other web applications. Indonesians spend an average of 3 hours and 15 minutes every day to access social media, resulting in a substantial amount of information flow. Even though research on information retrieval with social media data is common, only an inconsiderable amount concentrate using social media content in the Indonesian language. Our research aims to form an Indonesian language topic dataset using social media data from Twitter. The methods used in this research include TF-IDF for data formation and cosine similarity to group the Twitter data. Based on the test we conducted, our system is able to produce a fairly accurate result with 64% as its most optimal percentage for the process of every 200 Tweets.**

Keywords— ***dataset, cosine similarity, social media, TF-IDF, twitter***

I. PENDAHULUAN

Media sosial telah menjadi bagian hidup masyarakat Indonesia saat ini, di mana jumlah pengguna media sosial di Indonesia saat ini mencapai 50% dari jumlah keseluruhan total penduduk [1]. Saat ini hampir seluruh informasi, seperti pesan, iklan, promosi, dan pengumuman, disebarkan melalui media sosial dibandingkan dengan media yang lain. Hal ini dapat terjadi karena masyarakat Indonesia menghabiskan cukup banyak waktu untuk mengakses media sosial, yaitu rata-rata 3 jam 15 menit per hari, dengan demikian peluang sebuah informasi sampai ke tujuan menjadi lebih besar [2]. Selain itu, penyebaran informasi melalui media sosial tidak terikat oleh waktu, sehingga sebuah informasi dapat diakses setiap saat di manapun pengguna berada. Berbeda dengan media televisi atau radio, di mana sebuah informasi hanya disampaikan pada waktu tertentu saja, sehingga cukup besar kemungkinan seorang pengguna melewatkan informasi tersebut apabila tidak menonton televisi atau mendengarkan radio pada waktu yang bersamaan dengan penyiaran informasi tersebut. Selain itu media sosial juga memiliki kecepatan yang sangat tinggi dalam hal pembaruan (*update*) informasi. Apabila terdapat suatu kejadian seperti kecelakaan atau bencana alam yang baru saja terjadi, informasi tersebut dapat langsung tersebar di media sosial dengan cepat. Berbeda dengan media lain seperti televisi atau radio yang membutuhkan persiapan lebih banyak dalam penyiaran informasi, terlebih media cetak yang hanya dapat menyebarkan informasi paling cepat pada hari berikutnya. Di samping itu media sosial dibangun berbasis komunitas, artinya setiap orang dapat menjadi sumber informasi, tidak sekedar penerima saja. Hal ini menjadi salah satu faktor penyebab penyebaran informasi dapat berlangsung cukup cepat.

Secara umum, media sosial memiliki konten dalam bentuk teks yang terdiri dari beragam topik, meliputi promosi, pekerjaan, hobi, atau bahkan dunia politik. Selain itu, konten media sosial di tiap negara berbeda-beda karena setiap negara memiliki permasalahan dan kejadian yang berbeda-beda di dalamnya. Hal tersebut membuat media sosial menjadi topik yang menarik untuk diteliti, karena konten di dalamnya terus berkembang sesuai dengan pola yang terjadi dalam kehidupan masyarakat. Saat ini terdapat banyak penelitian seputar *text mining* terkait dengan konten-konten yang ada pada media sosial, khususnya di Indonesia. Namun kendala utama dalam penelitian-penelitian tersebut adalah proses ekstraksi fitur yang sulit dilakukan, karena minimnya pustaka dalam Bahasa Indonesia yang dapat digunakan untuk melakukan ekstraksi fitur pada konten media sosial. Ekstraksi fitur adalah proses mengubah sebuah *data* menjadi bentuk kuantitatif agar dapat diolah menggunakan proses matematis. Dalam hal ini *data* yang dimaksud adalah kata-kata dalam Bahasa Indonesia, yang hendak diubah menjadi bentuk angka agar dapat digunakan dalam berbagai keperluan riset.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk pembuatan *dataset* fitur berbahasa Indonesia. *Dataset* fitur Bahasa Indonesia akan dibangun dengan menggunakan metode *TF-IDF* dan *Cosine Similarity*.

TF-IDF akan digunakan untuk pembobotan kata dalam dokumen, sedangkan *Cosine Similarity* digunakan untuk mencari dokumen-dokumen yang memiliki kemiripan, sehingga didapatkan kelompok-kelompok dokumen sesuai topiknya masing-masing. Dari masing-masing kelompok dokumen, kata-kata kunci yang merepresentasikan kelompok dokumen tersebut disimpan dan diolah ke dalam *dataset* fitur lengkap beserta bobotnya. Dalam penelitian ini, media sosial Twitter dipilih sebagai obyek penelitian, namun tidak menutup kemungkinan hasil penelitian ini dapat digunakan pada media sosial yang lain apabila pihak pengembang memiliki akses terhadap konten-konten di dalam media sosial tersebut. Luaran dari penelitian ini adalah pustaka *dataset* fitur Bahasa Indonesia yang dapat digunakan untuk mendapatkan fitur dari sebuah kata. Harapan penulis, *dataset* ini dapat digunakan untuk mendukung penelitian-penelitian lain terkait dengan *text mining*, khususnya *text mining* dengan konten Bahasa Indonesia.

II. TINJAUAN PUSTAKA

Penelitian yang dilakukan oleh Zaanen dan Kansters [3] untuk mengklasifikasikan dokumen lirik lagu ke beberapa kelas *mood* yang tersedia dengan menggunakan metode pembobotan *TF-IDF*. Proses pengelompokan dilakukan berdasarkan kata-kata yang terdapat dalam lirik masing-masing lagu. Kata-kata tersebut diolah menggunakan *TF-IDF*, kemudian diklasifikasikan berdasarkan jenis kata yang muncul. Dengan demikian, diharapkan lagu-lagu yang memiliki jenis *mood* yang sama dapat diklasifikasikan dalam kelas yang sama karena memiliki kata-kata dalam lirik yang sejenis.

Document similarity adalah metode utuk menghitung tingkat kemiripan antar dokumen. Salah satu rumus perhitungan yang dapat digunakan untuk menghitung kemiripan antar dokumen adalah *cosine similarity*. Penelitian yang dilakukan oleh Kompan dan Biolikova [4] pada tahun 2010, dilakukan menggunakan metode *TF-IDF* dan *cosine similarity*, dengan obyek yang diteliti adalah artikel berita. Penelitian tersebut menghasilkan sebuah solusi yang dapat mencari rekomendasi artikel yang memiliki karakteristik sejenis dengan artikel yang sering dibaca oleh seorang pengguna. Mula-mula dilakukan proses pembobotan terhadap isi dari artikel yang dibaca oleh pengguna dengan menggunakan *TF-IDF*, kemudian dari bobot yang telah diperoleh dilakukan proses perhitungan *cosine similarty* antara artikel tersebut dengan seluruh artikel yang lain. Dari nilai *cosine similarity* tersebut, akan diperoleh artikel-artikel yang mirip dengan artikel yang sering dibaca oleh pengguna tersebut. Penelitian lain yang menggunakan metode *TF-IDF* dan *cosine similarity* juga dilakukan oleh Lahitani, Permanasari, dan Setiawan [5]. Mereka melakukan penelitian mengenai bagaimana melakukan penilaian otomatis terhadap jawaban dari soal *essay* (*Automated Essay Scoring*). Proses pemberian nilai mula-mula dilakukan dengan menggunakan *TF-IDF* untuk pembobotan, kemudian berdasarkan nilai bobot tersebut dilakukan perhitungan *cosine similarity* antara jawaban yang didapat dengan jawaban yang diharapkan. Nilai *cosine similarity* tersebut nantinya digunakan untuk menentukan nilai dari jawaban soal *essay* tersebut.

Berdasarkan hasil penelitian sebelumnya, maka diperoleh kesimpulan bahwa metode *TF-IDF* dan *cosine similarity* dapat digunakan untuk mencari tingkat kemiripan antar dokumen dengan hasil yang cukup baik. Oleh karena itu, penelitian aktual yang dilakukan penulis menggunakan metode pembobotan *TF-IDF* dan *cosine similarity* untuk menghasilkan *dataset* kata bahasa Indonesia menggunakan obyek twitter feeds. Perbedaan penelitian aktual dengan penelitian terdahulu dapat dilihat pada TABEL I.

TABEL I
PERBEDAAN PENELITIAN AKTUAL DENGAN PENELITIAN TERDAHULU

Judul	*Automatic Mood Classification Using TF*IDF Based on Lyrics*	*Content-based News Recommendation*	*Cosine similarity to determine similarity measure: Study case in online essay assessment*	Pembentukan *Dataset* Topik Kata Bahasa Indonesia pada Twitter Menggunakan *TF-IDF* dan *Cosine Similarity*
Oleh	Menno Van Zaanen dan Pieter Kanters	Michal Kompan dan Maria Bielikova	Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, dan Noor Akhmad Setiawan	Kristian Adi Nugraha dan Danny Sebastian
Tahun	2010	2010	2016	2018
Metode	*TF-IDF*	*TF-IDF* dan *cosine-similarity*	*TF-IDF* dan *cosine-similarity*	*TF-IDF* dan *cosine similarity*
Obyek	Lirik lagu	Konten berita	Ujian Essay	Twitter *feeds*
Bahasa	Inggris	Inggris	Indonesia	Indonesia

III. LANDASAN TEORI

A. Media Sosial

Media sosial atau *social media* adalah aplikasi berbasis internet yang dibangun berdasarkan ideologi dan teknologi dari *Web 2.0*, dan memungkinkan pengguna membuat konten [3]. Pengguna media sosial adalah orang dari seluruh dunia dengan karakter yang beraneka ragam, mereka saling bertukar informasi, berkolaborasi, dan berbagi konten antar pengguna [4]. *Web 2.0* adalah istilah yang pertama kali digunakan pada tahun 2004, yang mengubah paradigma konten website harus dibangun oleh pemilik *website* menjadi konten *website* dibangun oleh seluruh pengguna aplikasi *website* [5] [6]. Menurut Tim O'Reilly, Web 2.0 dapat didefinisikan sebagai revolusi bisnis di industri komputer yang disebabkan oleh penggunaan internet sebagai *platform*, dan merupakan suatu percobaan untuk memahami berbagai aturan untuk mencapai keberhasilan pada *platform* baru tersebut. Salah satu aturan utamanya adalah membangun aplikasi yang mengeksploitasi efek jaringan untuk mendapatkan lebih banyak pengguna. Terdapat 6 prinsip utama dalam *web 2.0* [6], yaitu:

1. *Website as a Platform*
   Aplikasi *website* berkembang menjadi aplikasi yang dapat digunakan oleh perangkat lain. Salah satu tujuan dari *web 2.0* adalah *website* sebagai *development platform* bagi aplikasi lain.

2. *Data is the Intel Inside*
   Data menjadi penting, semakin banyak dan menarik *data* yang ada, aplikasi *website* menjadi lebih menarik.

3. *End of Software Release Cycle*
   Aplikasi *website* diletakkan di *server*, sehingga untuk melakukan pembaruan versi perangkat cukup dilakukan melalui *server*.

4. *Lightweight programming models*
   Bahasa pemrograman *website* relatif ringan dijalankan pada semua perangkat. Hal ini dikarenakan pengolahan *data* dilakukan di server.

5. *Software above a single device*
   Aplikasi *website* dapat digunakan di semua perangkat melalui *browser*. Standarisasi dilakukan pada proses interpretasi yang dilakukan oleh *browser*.

6. *Rich User Experiences*
   Aplikasi *website* menjadi mempertimbangkan kenyamanan pengguna dalam menggunakan aplikasi *website*. Implementasi untuk meningkatkan kenyamanan pengguna adalah menggunakan *javascript*.

User Generated Content atau *UGC* merupakan konsep dimana pengguna media sosial dapat membuat konten berupa teks yang disebarkan dan digunakan/dibaca melalui *social media* [7] [8]. *UGC* disebut juga *electronic world-of-mouth* (*eWOM*) yang bekerja sama seperti *world-of-mouth* (*WOM*) konvensional [9]. Yang membedakan *eWOM* dan *WOM* konvensional adalah sarana persebaran yang dilakukan secara *online* [10]. Pengguna yang mengisikan konten ke media sosial, umumnya dilakukan secara sukarela atau tanpa dibayar [11].

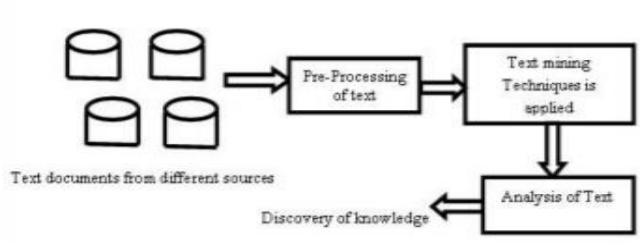
Dalam beberapa penelitian, media sosial digunakan oleh peneliti untuk melihat *trend* yang terjadi di masyarakat [7][12]. Pada *social media*, konten yang dituliskan oleh pengguna dapat dianalisis menggunakan beberapa kriteria untuk menghasilkan sebuah informasi atau *trend* [13]. Salah satu metode yang digunakan untuk melakukan analisis media sosial adalah dengan *text mining* [14].

B. Text Mining

Text mining atau *text data mining*, adalah bidang pengetahuan yang mencakup area *information retrieval* [15], *text analysis*, *information extraction* [16], *clustering* [17], *categorization* [18] [19], *visualization*, *database technology*, *machine learning*, dan *data mining* [20]. Dua komponen dari *text mining frameworks* adalah *text refining* dan *knowledge distillation*. *Text refining* melakukan proses transformasi dari *unstructured document* atau dokumen yang tidak terstruktur ke *intermediate form*. Sedangkan *knowledge distillation* memproses *intermediate form* menjadi *pattern* atau *knowledge*.

Dalam beberapa tahun terakhir, obyek dari *text mining* yang banyak diteliti adalah *website* atau *world wide web* [21] [22]. *World wide web* memiliki banyak konten dokumen teks yang dapat diolah lebih lanjut menggunakan *text mining*, antara lain berita [23], media sosial [3] [24] [25], *e-commerce* [26] , dan lainnya.

Tahap pertama dari *text mining* adalah *text preprocessing*, dimana *text preprocessing* mengolah *data* tidak terstruktur menjadi *data* terstruktur. Keluaran dari *text preprocessing* digunakan untuk masukan ke *algoritma text mining*. Keluaran dari *algoritma text mining* dianalisis untuk menjadi sebuah pengetahuan. Tahapan dalam *text mining* dapat dilihat pada Gambar 1.



Gambar 1. Proses *text mining*
Dikutip dari: Vijayarani, S., Ms J. Ilamathi, and Ms Nithya, 2015, *Preprocessing Techniques for Text Mining – An Overview* [21]

C. Text Preprocessing

Tahapan dalam *text mining* dimulai dengan *text preprocessing*. *Text preprocessing* menyiapkan *data* teks menjadi kata/token yang siap diolah lebih lanjut. *Text preprocessing* berpengaruh terhadap keberhasilan *algoritma text mining* yang digunakan [27]. Proses yang dilakukan dalam *text preprocessing* adalah:

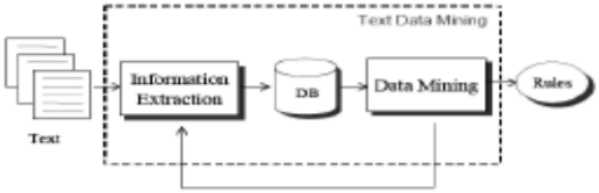
1. Tokenisasi
   Dalam proses tokenisasi, dokumen teks akan dipecah menjadi sebuah *token* atau sebuah kata [28]. Dalam proses tokenisasi, dilakukan penghapusan karakter spesial dan tanda baca, dan menyesuaikan tipe kapitalisasi teks.

2. Menghilangkan *stop word*
   Setiap *token* yang dihasilkan dari proses tokenisasi, akan dibersihkan dari *stop word*, atau disebut juga dengan proses *filtering* [27]. *Stop word* merupakan kata yang dianggap tidak mencerminkan keyword dari dokumen. Menghilangkan *stop word* dapat mengurangi dimensi dari jumlah kata yang diolah, sehingga mempercepat proses analisis [18]. *Stop word* disesuaikan dengan dengan bahasa dari dokumen yang diproses.

3. *Stemming & Lematisasi*
   Stemming adalah metode yang digunakan untuk menghasilan *stem/root*/kata dasar dari sebuah *token* [18]. Tujuan dari proses *stemming* adalah menghilangkan imbuhan, sehingga mengurangi jumlah dari kata yang diproses dalam *text mining*, menghemat waktu, dan menghemat memori. Lematisasi adalah proses mengubah sebuah kata menjadi bentuk yang sesuai (*lemma*), sehingga dapat dikelompokkan dengan kata lain yang sama [27]. Tujuan dari lematisasi adalah mengubah *infinite tense* dan *noun* menjadi sebuah kata dalam Bahasa Inggris yang sama. Pada penelitian ini lematisasi tidak diperlukan karena kata-kata dalam Bahasa Indonesia tidak memiliki bentuk-bentuk khusus (*infinite tense*, *noun*) seperti kata dalam Bahasa Inggris.

D. Information Extraction

Information extraction merupakan identifikasi kata kunci dari sebuah dokumen teks dan relasi antar dokumen teks secara otomatis [18] [27]. *Information extraction* memproses *data* yang tidak terstruktur menjadi *data* terstruktur yang siap digunakan algoritma *text mining*. *Information extraction* melihat pola sekuensial dari kata kunci yang sudah ditentukan, ini disebut dengan *pattern matching* [29]. *Information extraction* bertujuan menghasilkan informasi penting dari kelompok dokumen dengan jumlah yang banyak. Kerangka kerja dari *Information Extraction* dalam *text mining* dapat dilihat pada Gambar 2.



Gambar 2. *Overview of Information Extraction-based text mining framework*
Dikutip dari: V. Gupta and G. S. Lehal, 2009, *A Survey of Text Mining Techniques and Applications* [32]

E. Term Frequency – Inverse Document Frequency

Terms Frequency & *Inverse Document Frequency* (*TF-IDF)* merupakan metode pembobotan secara statistik yang menunjukkan seberapa pentingnya sebuah kata pada suatu dokumen, dimana dokumen terletak pada sebuah kelompok dokumen [18] [30]. Metode pembobotan *TF-IDF* biasanya digunakan dalam *text mining*. *Term frequency* adalah jumlah sebuah kata pada dokumen. Rumus *TF* dapat dilihat pada rumus 1.

$$tf(t,d) = .5 + \frac{0.5 \times f(t,d)}{Maximum\ occurences\ of\ words} \quad [1]$$

Dengan:
$tf(t,d)$: *term frequency* kata t pada dokumen d
$f(t,d)$: jumlah frekuensi kata t pada dokumen d

Inverse document frequency atau *IDF* adalah nilai yang digunakan untuk mengukur seberapa penting sebuah kata pada koleksi dokumen. Nilai dari *IDF* akan semakin kecil apabila suatu kata muncul di banyak dokumen. Sedangkan nilai dari *IDF* akan semakin besar apabila suatu kata hanya muncul di sedikit dokumen. Rumus *IDF* dapat dilihat pada rumus 2.

$$idf(t,d) = \log \frac{|D|}{no\ of\ documents\ term\ t\ appears} \quad [2]$$

Dengan:
$idf(t,d)$: *inverse document frequency* kata t dalam dokumen d
$|D|$: jumlah dokumen

Setelah mendapatkan nilai *TF* dan nilai *IDF*, langkah selanjutnya adalah menghitung nilai *TF-IDF*. Nilai *TF-IDF* dihitung menggunakan rumus 3 untuk setiap kata dalam koleksi dokumen.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,d) \quad [3]$$

Dengan:
$tf(t,d)$: *term frequency* kata t pada dokumen d
$idf(t,d)$: *inverse document frequency* kata t dalam dokumen d

F. *Cosine Similarity*

Cosine similarity merupakan metode pengukuran yang banyak digunakan di *pattern recognition* dan *text classification* [31]. *Cosine similarity* mengukur kemiripan dua buah vektor dalam sebuah *product space* dengan mengukur cosine dari sudut kedua vektor [32]. Rumus perhitungan *cosine similarity* dapat dilihat pada rumus nomor 4.

$$Cosine(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{|\vec{x}||\vec{y}|} \quad [4]$$

Dengan:
$\vec{x}$: representasi dokumen kedalam bentuk vektor
$\vec{y}$: representasi dokumen kedalam bentuk vektor

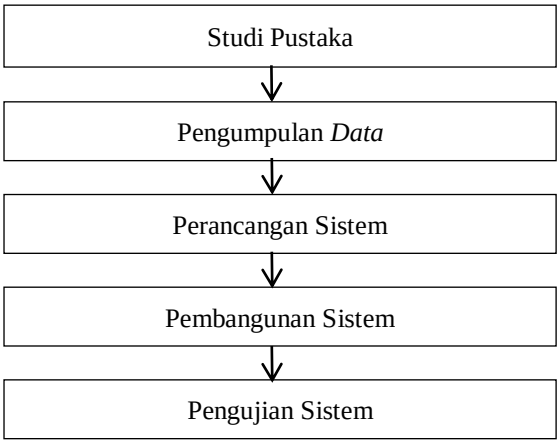
Berbeda dengan perhitungan *similarity* berbasis jarak, cosine similarity menghitung nilai kemiripan dua buah titik dengan cara menghitung kedekatan nilai sudut yang dibentuk terhadap koordinat (0,0). Semakin dekat sudut yang dibentuk dari kedua buah titik, maka semakin mirip kedua buah titik tersebut. Dalam penelitian ini, titik-titik tersebut merupakan Tweet yang hendak diolah oleh sistem.

IV. METODE PENELITIAN

Penelitian yang akan dilakukan oleh penulis terdiri dari enam tahap seperti yang ditunjukkan pada Gambar 3. Adapun penjelasan masing-masing tahapan adalah sebagai berikut:

A. *Studi Pustaka*

Studi mengenai metode *TF-IDF* dan *Cosine Similarity* dilakukan agar dapat diimplementasikan ke dalam sistem yang dibangun. Penulis mempelajari metode tersebut berdasarkan buku teks mengenai *text mining*. Selain itu penulis juga mempelajari tentang struktur *data* dari Tweet yang akan diteliti, meliputi karakteristik konten di dalamnya serta *meta-data* apa saja yang perlu untuk diolah.



Gambar 3. Langkah penelitian

B. *Pengumpulan Data*

Data Tweet dikumpulkan dalam periode waktu empat bulan terakhir (April, Mei, Juni, Juli) bersumber dari akun Twitter yang telah ditentukan. *Data-data* tersebut akan digunakan untuk bahan percobaan selama pengembangan sistem. Sedangkan *data* yang akan digunakan untuk pengujian akhir adalah *data-data* paling terbaru, terhitung sebelum tanggal 31 Juli 2018. Akun-akun Twitter yang dijadikan obyek penelitian adalah akun-akun media elektronik nasional terverifikasi Kompas dan Detik. Terverifikasinya akun pada Twitter ditandai dengan adanya tanda centang dengan latar belakang biru di sebelah nama akun Twitter yang bersangkutan seperti yang ditunjukkan pada Gambar 4.

Gambar 4. Contoh akun terverifikasi

Kategori akun media elektronik nasional bereputasi dipilih dengan pertimbangan bahasa yang digunakan oleh akun-akun tersebut memiliki struktur yang baku dan berbentuk formal, sehingga diharapkan dapat lebih mudah untuk diproses pada sistem nantinya. Selain itu, konten yang ada di dalamnya sangat beragam dan menyesuaikan segala aktivitas dan kejadian yang terjadi di negara Indonesia saat ini. Dengan alasan tersebut, maka dapat dikatakan bahwa seluruh konten yang dimiliki oleh akun-akun media nasional

tersebut merupakan cerminan dari Negara Indonesia itu sendiri.

C. Perancangan Sistem

Setelah mengumpulkan *data*, penulis melakukan perancangan sistem. Perancangan sistem meliputi perancangan basis *data* serta perancangan program yang akan dibangun. Beberapa parameter yang perlu diperhatikan dalam perancangan basis *data* adalah berkaitan dengan bentuk data Tweet yang berhasil dikumpulkan, meliputi isi, *user id*, dan tanggal pembuatan.

D. Pembangunan Sistem

Sistem dibangun dengan menggunakan bahasa pemrograman PHP dengan *framework* Laravel. *Platform web* dipilih untuk pembangunan sistem karena *platform website* secara otomatis terhubung dengan jaringan *internet*, sehingga proses komunikasi untuk pengambilan data ke *server* Twitter dapat diimplementasikan lebih mudah.

Dalam penelitian ini, penulis tidak menggunakan metode lematisasi. Hal ini dikarenakan lematisasi digunakan untuk mengelompokkan kata yang sama, tetapi memiliki beda bentuk, seperti perbedaan *tense* dalam Bahasa inggris.

E. Pengujian Sistem

Pengujian sistem dilakukan dengan menggunakan *data* Tweet yang berhasil dikumpulkan sebelum tanggal 31 Juli 2018. Penulis mengambil *data* sebanyak 400 Tweet ke belakang untuk setiap akun media elektronik yang telah ditentukan sebelumnya. Pengujian tersebut dibagi dalam tiga tahap yaitu mengolah 100 dari 400 Tweet, sehingga dibutuhkan 4 kali proses untuk menguji seluruh Tweet. Kemudian mengolah 200 dari 400 Tweet, sehingga dibutuhkan 2 kali proses untuk menguji seluruh Tweet. Tahap terakhir adalah menguji 400 Tweet sekaligus. Dari setiap tahap pengujian yang dilakukan oleh penulis, maka akan didapatkan daftar kelompok-kelompok yang anggotanya merupakan gabungan dari dua atau lebih Tweet dengan nilai kemiripan paling tinggi. Nilai kemiripan tersebut dihitung dengan menggunakan persamaan *cosine similarity* seperti yang ditunjukkan pada rumus 3.

F. Analisis Hasil Pengujian

Setelah didapatkan kelompok-kelompok Tweet pada tahap sebelumnya, penulis melakukan analisis terhadap kelompok-kelompok tersebut. Pertama-tama analisis dilakukan dengan cara memberi label untuk masing-masing kelompok yang ada. Pemberian label dilakukan secara *manual* berdasarkan topik umum yang mewakili anggota kelompok di dalamnya. Setelah penulis melakukan pelabelan, penulis melakukan pencarian kata kunci yang mewakili label-label tersebut kemudian memasukkannya ke dalam *dataset*. *Dataset* yang merupakan luaran akhir penelitian ini nantinya dapat digunakan untuk mengetahui bobot dari sebuah kata yang dimasukkan ke dalamnya.

V. ANALISIS & PEMBAHASAN

Penulis melakukan pengujian terhadap *data-data* Tweet yang telah berhasil dikumpulkan sebelumnya. Data-data tersebut berasal dari sembilan akun Twitter milik media nasional yang telah terverifikasi, yaitu:

1. BBC Indonesia (@bbcindonesia)
2. CNN Indonesia (@CNNIndonesia)
3. Detik (@detikcom)
4. JPNN (@jpnncom)
5. Kompas (@kompascom)
6. Kontan (@kontannews)
7. Koran Tempo (@korantempo)
8. Media Indonesia (@mediaindonesia)
9. Tempo (@tempodotco)

Tweet yang berhasil dikumpulkan diolah ke dalam sistem untuk melihat kesamaan (*similarity*) antara satu Tweet dengan Tweet yang lain. Pengecekan kesamaan Tweet dilakukan dengan menggunakan rumus *cosine similarity* seperti yang ditunjukkan pada rumus nomor 3. Sebuah Tweet akan dibandingkan dengan seluruh Tweet yang lain dengan menghitung nilai kesamaan menggunaan rumus *cosine similarity*. Setelah nilai kesamaan didapatkan, maka Tweet tersebut akan membentuk kelompok bersama dengan Tweet yang nilai kesamaannya paling tinggi diantara yang lain. Proses ini akan diulang terus-menerus hingga sebuah Tweet telah membentuk kelompok bersama dengan Tweet-Tweet yang nilai kesamaannya paling tinggi. Apabila telah terbentuk kelompok-kelompok Tweet yang sejenis, maka penulis memberi label terhadap kelompok tersebut sesuai dengan isi Tweet dari kelompok tersebut. Setelah itu, penulis mengambil kata-kata kunci yang mencerminkan kelompok tersebut untuk dimasukkan ke dalam *dataset*. Setelah seluruh proses selesai dilakukan, maka terbentuklah sebuah *dataset* yang berisi kata-kata dalam Bahasa Indonesia lengkap beserta bobot dari masing-masing kata. Pengguna *dataset* tersebut nantinya dapat memasukkan kata ke dalamnya untuk mengecek bobot kategori dari kata yang dimasukkan.

Untuk masing-masing akun Twitter, penulis mengambil 400 Tweet terakhir sebelum tanggal 31 Juli 2018 ke belakang. Dari 400 Tweet yang berhasil didapatkan, penulis menguji Tweet tersebut menggunakan tiga cara, yaitu:

1. Menguji tiap 100 Tweet (4 kali pengujian)
2. Menguji tiap 200 Tweet (2 kali pengujian)
3. Menguji 400 Tweet sekaligus

Proses pengujian dilakukan dengan tiga cara untuk melihat konsistensi kemunculan kelompok-kelompok yang sama antara jenis pengujian satu dengan yang lain. Misalnya kelompok-kelompok yang muncul di pengujian 100 Tweet seharusnya juga muncul di pengujian 200 Tweet dan seterusnya. Setelah seluruh pengujian berhasil dilakukan,

penulis merangkumnya ke dalam tabel seperti yang ditunjukkan pada TABEL II.

TABEL II
JUMLAH KELOMPOK YANG DIHASILKAN

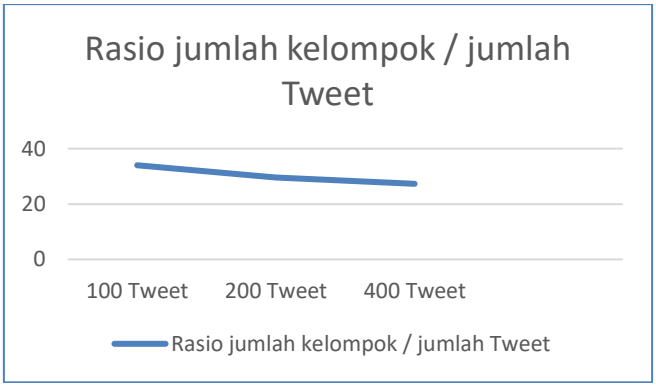
Akun	100			200			400
	MIN	MAX	RATA	MIN	MAX	RATA	RATA
bbcindonesia	28	30	29	50	54	52	105
cnnindonesia	23	36	32.3	50	57	53.5	99
detikcom	20	32	26.8	48	54	51	98
jpnncom	34	40	36.8	57	62	59.5	87
kompascom	30	35	32.3	54	56	55	94
kontannews	28	35	30.8	51	58	54.5	109
korantempo	34	41	37	62	65	63.5	118
mediaindonesia	42	47	45	76	79	77.5	150
tempodotco	33	37	35.5	59	63	61	120
MIN	20	30	26.8	48	54	51	87
MAX	42	47	45	76	79	77.5	150
RATA-RATA	30.2	37	33.9	56.3	60.9	58.6	109

Pada TABEL II terlihat bahwa rata-rata jumlah kelompok yang terbentuk adalah 34 kelompok untuk 100 Tweet, 59 kelompok untuk 200 Tweet, dan 109 kelompok untuk 400 Tweet. Jumlah kelompok paling sedikit untuk 100 dan 200 Tweet didapat dari akun Detik, yaitu masing-masing 20 kelompok (100 Tweet) dan 48 kelompok (200 Tweet). Sedangkan kelompok paling banyak untuk 100 dan 200 Tweet didapat dari akun Media Indonesia, yaitu masing-masing 47 kelompok (100 Tweet) dan 79 kelompok (200 Tweet). Sedangkan untuk 400 Tweet, jumlah kelompok paling sedikit didapat dari akun JPNN yaitu 87 kelompok, sedangkan jumlah kelompok paling banyak didapatkan dari akun Media Indonesia yaitu 150 kelompok. Dengan demikian, maka dapat disimpulkan bahwa secara keseluruhan rentang jumlah kelompok masing-masing tahapan adalah 20 sampai 47 kelompok untuk 100 Tweet, 48 sampai 79 kelompok untuk 200 Tweet, 87 sampai 150 kelompok untuk 400 Tweet. Apabila dihitung rasio antara rata-rata jumlah kelompok yang berhasil terbentuk dengan jumlah Tweet yang diolah, maka didapatkan hasil 34% untuk 100 Tweet, 29.5% untuk 200 Tweet, dan 27.3% untuk 400 Tweet. Grafik perhitungan rasio tersebut ditunjukkan pada

Gambar 5.

Pada

Gambar 5 dapat terlihat bahwa semakin banyak jumlah Tweet yang diolah, maka rasio jumlah kelompok terhadap jumlah Tweet yang diolah semakin kecil. Hal ini disebabkan karena kelompok-kelompok kecil yang terbentuk pada Tweet dengan jumlah sedikit (100 Tweet) akan bergabung menjadi satu kelompok ketika diolah menggunakan Tweet dengan jumlah yang lebih besar (400 Tweet).



Gambar 5. Grafik rasio jumlah kelompok terhadap jumlah Tweet

Setelah penulis mendapatkan kelompok-kelompok Tweet yang berhasil terbentuk pada tahap sebelumnya, penulis melakukan pelabelan terhadap kelompok-kelompok tersebut. Penulis memberi label terhadap kelompok yang memiliki anggota minimal 10 Tweet di dalamnya. Label tersebut ditentukan secara *manual* berdasarkan topik konten secara umum dari anggota kelompok tersebut. Pada TABEL III dapat terlihat jumlah kelompok yang telah diberi label untuk masing-masing akun. Rata-rata jumlah kelompok yang memiliki label untuk masing-masing tahapan adalah 1 kelompok untuk 100 Tweet, 2 kelompok untuk 200 Tweet, dan 3 kelompok untuk 400 Tweet. Untuk rentang jumlah kelompok yang memiliki label masing-masing adalah 0 sampai 2 kelompok untuk 100 Tweet, 0 sampai 3 kelompok untuk 200 Tweet, dan 1 sampai 4 kelompok untuk 400 Tweet.

TABEL III
JUMLAH KELOMPOK DENGAN LABEL

Akun	100			200			400
	MIN	MAX	RATA	MIN	MAX	RATA	RATA
bbcindonesia	1	2	1.25	1	1	1	4
cnnindonesia	0	1	0.25	3	3	3	3
detikcom	1	2	1.25	2	3	2.5	2
jpnncom	0	1	0.25	2	3	2.5	4
kompascom	0	2	0.75	0	2	1	4
kontannews	0	2	0.75	1	2	1.5	4
korantempo	0	0	0	0	1	0.5	1
mediaindonesia	0	0	0	0	1	0.5	3
tempodotco	0	1	0.25	1	1	1	2
MIN	0	0	0	0	1	0.5	1
MAX	1	2	1.25	3	3	3	4
RATA-RATA	0.22	1.22	0.53	1.11	1.89	1.5	3

Berdasarkan hasil analisis penulis terhadap data kelompok yang telah diberi label, terdapat sebagian kelompok yang anggota-anggota Tweet di dalamnya memiliki topik yang cukup berbeda. Hal ini dapat terjadi karena antara Tweet-Tweet tersebut memiliki kata kunci yang sama meskipun topik yang diangkat berbeda, sehingga ketika dihitung menggunakan perhitungan *cosine similarity*

Tweet-Tweet tersebut dapat berada di kelompok yang sama. Seluruh kelompok yang anggotanya berjumlah 10 atau lebih namun tidak dapat didefinisikan labelnya, karena topik dari masing-masing anggota label tersebut sangat berbeda, maka penulis menganggap bahwa kelompok tersebut tidak valid sehingga diberi nilai salah (false). Sedangkan seluruh kelompok yang dapat diberi label yang sesuai diberi nilai benar (true). Berdasarkan nilai benar dan salah yang berhasil dikumpulkan, penulis menghitung akurasi tersebut dengan cara membagi nilai seluruh kelompok berlabel yang bernilai benar dengan jumlah seluruh kelompok berlabel yang ada.

TABEL IV
AKURASI PELABELAN KELOMPOK PER-100 TWEET

Akun	100		
	T	F	AKURASI
bbcindonesia	3	2	60%
cnnindonesia	0	1	0%
detikcom	2	3	40%
jpnncom	0	1	0%
kompascom	2	1	67%
kontannews	3	0	100%
korantempo	0	0	0%
mediaindonesia	0	0	0%
tempodotco	1	0	100%
MIN	0	0	0%
MAX	3	3	100%
RATA-RATA	1.22	0.89	41%

TABEL V
AKURASI PELABELAN KELOMPOK PER-200 TWEET

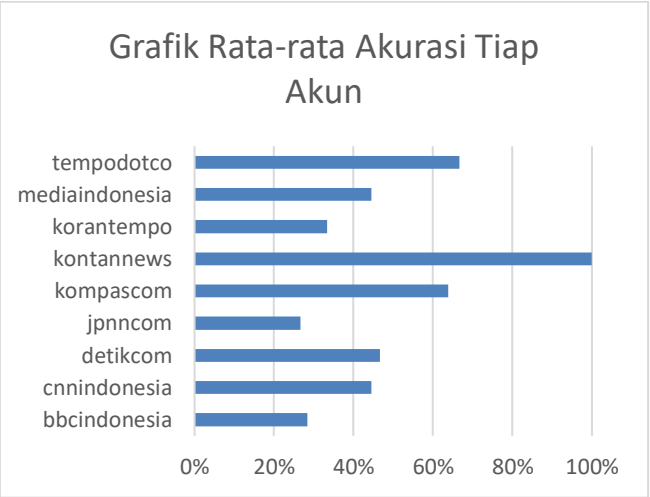
Akun	200		
	T	F	AKURASI
bbcindonesia	0	2	0%
cnnindonesia	6	0	100%
detikcom	5	0	100%
jpnncom	4	1	80%
kompascom	1	1	50%
kontannews	3	0	100%
korantempo	0	0	0%
mediaindonesia	10	0	100%
tempodotco	1	1	50%
MIN	0	0	0%
MAX	10	2	100%
RATA-RATA	3.33	0.63	64%

TABEL VI
AKURASI PELABELAN KELOMPOK PER-400 TWEET

Akun	400		
	T	F	AKURASI
bbcindonesia	1	3	25%
cnnindonesia	1	2	33%
detikcom	0	2	0%
jpnncom	0	4	0%
kompascom	3	1	75%
kontannews	4	0	100%
korantempo	1	0	100%
mediaindonesia	1	2	33%
tempodotco	1	1	50%
MIN	0	0	0%
MAX	4	4	100%
RATA-RATA	1.33	1.67	46%

Berdasarkan hasil yang ditunjukkan pada TABEL IV, TABEL V, dan TABEL VI dapat terlihat nilai rata-rata akurasi untuk masing-masing tahapan pengujian adalah 41% untuk 100 Tweet, 64% untuk 200 Tweet, 46% untuk 400 Tweet, dan rata-rata ketiganya adalah 50%. Secara keseluruhan, rentang akurasi terkecil hingga terbesar untuk seluruh tahapan adalah 0% sampai dengan 100%. Akun Twitter yang memiliki akurasi keseluruhan paling tinggi adalah akun milik Kontan dengan tingkat akurasi 100%. Hal ini dapat terjadi karena media Kontan memiliki isi Twitter yang cukup spesifik yaitu mengenai keuangan, ekonomi, atau perbankan. Sementara akun-akun Twitter yang lain memiliki konten yang lebih beragam karena membahas berita secara umum, tidak hanya dari dari satu topik saja. Grafik rata-rata akurasi keseluruhan dapat dilihat pada

Gambar 6. Apabila dihitung rata-rata secara keseluruhan dari seluruh data akurasi yang ada, maka akan didapatkan nilai persentase rata-rata sebesar 50%.



Gambar 6. Grafik rata-rata akurasi tiap akun

Berdasarkan hasil kelompok-kelompok yang telah diberi label, penulis merangkum label-label tersebut seperti yang ditunjukkan pada TABEL VII. Jumlah total kelompok yang memiliki label adalah 81 kelompok, terdiri dari 53 kelompok yang memiliki label *valid* dan 28 kelompok yang tidak terdefinisi. Tidak terdefinisi artinya Tweet yang terdapat pada label tersebut memiliki topik yang sangat berbeda antara satu dengan yang lain, sehingga penulis kesulitan untuk memberi label yang dapat mewakili keseluruhan isi dari kelompok tersebut. Pada tabel tersebut dapat terlihat bahwa topik yang paling sering muncul adalah

topik mengenai politik (12 label). Hal ini disebabkan karena pada saat ini media massa sedang banyak membahas mengenai pemilihan presiden yang diadakan pada tahun 2019, sehingga konten media massa saat ini dipenuhi dengan topik-topik seputar pemilihan presiden. Urutan berikutnya adalah topik mengenai ekonomi sebanyak sembilan label. Topik ini juga cukup banyak diperbincangkan terutama berkaitan dengan melemahnya kurs Rupiah terhadap Dollar Amerika. Urutan berikutnya terdapat topik mengenai sepak bola (piala dunia), gempa, dan hiburan, ketiganya berjumlah masing-masing sebanyak empat label.

TABEL VII
DAFTAR LABEL

Label	Jml	Sample Fitur Kata
Politik	12	pilkada, partai
Ekonomi	9	jual, kurs, rupiah
Sepak Bola	4	skor, rusia, juara
Gempa	4	gempa, skala
Hiburan	4	musik, nyanyi
Asian Games	3	asian, games, atlet
Penyelamatan Anak	2	thailand, anak
Perbankan	2	bank, uang, rupiah
Kriminal	2	penjara, hukum
Penculik	1	culik, korban
Berita Korea	1	korea, seoul
Hukum	1	undang, atur
Olah Raga	1	lomba, juara
KPK	1	kpk, korupsi
Lalu Lintas	1	celaka, macet
OSO	1	hanura, politik
Kekerasan	1	keras, aniaya
Pemerintah	1	presiden, perintah
Internasional	1	amerika, jepang
Youtube	1	youtube, video
Tidak Terdefinisi	28	-
Jumlah Label	81	

Berdasarkan analisis data hasil pengujian yang telah dilakukan penulis, terdapat beberapa kekurangan pada luaran akhir yang dihasilkan. Kekurangan pertama adalah pada beberapa media tidak ada standar dalam penulisan konten Tweet. Hal tersebut disebabkan oleh akun milik sebuah media dikelola oleh banyak *administrator*, sehingga tata bahasa dari konten yang dituliskan sangat bergantung pada orang yang menuliskan. Selain itu terdapat konten-konten yang tidak seragam dan seringkali tercampur satu sama lain pada sebuah akun. Salah satu contohnya seperti yang ditunjukkan pada TABEL VII terdapat satu label yang tidak mencerminkan konten Tweet yaitu label Youtube. Label Youtube dapat terbentuk karena terdapat konten-konten yang diawali dengan kalimat "I added a video to a @YouTube playlist..." karena akun tersebut baru saja mengungah video melalui akun media Youtube. Karena kalimat awalan yang mirip, sehingga seluruh Tweet dengan kalimat awalan tersebut akan dianggap sebagai Tweet yang mirip dan dijadikan satu kelompok tersendiri, meskipun konten video Youtube yang diunggah terdiri dari topik yang bermacam-macam. Kekurangan berikutnya adalah banyak akun-akun yang di dalamnya masih terdapat campuran bahasa asing seperti Inggris, kata serapan, atau bahasa tidak resmi (slang), meskipun akun tersebut merupakan akun media milik di Indonesia. Hal ini membuat kata-kata tersebut tidak dapat dikenali dalam proses *stemming* dan pembersihan *stopword*, sementara terdapat kemungkinan bahwa konten pada Tweet tersebut sejenis dengan konten Tweet lain sehingga seharusnya dapat berada pada kelompok yang sama. Kendala yang lain adalah terbatasnya jumlah karakter pada Twitter yaitu sebanyak 280 karakter, hal ini mengakibatkan terpotongnya kata atau kalimat pada konten Tweet yang ada, sehingga kalimat tersebut tidak utuh dan menjadi sulit untuk diproses.

Luaran akhir dari penelitian ini adalah *dataset* fitur kata dalam Bahasa Indonesia. Seluruh label valid beserta kata kunci yang telah berhasil dikumpulkan digunakan untuk membentuk *dataset*. *Dataset* ini dapat digunakan untuk mengetahui fitur dari sebuah kata yang dimasukkan. TABEL VIII berisi beberapa contoh tentang penggunaan *dataset* yang dihasilkan tersebut. *Dataset* dapat menerima masukan berupa kalimat, kemudian *dataset* akan menghasilkan luaran berupa persentaset topik-topik yang terkait dengan kalimat tersebut.

TABEL VIII
CONTOH PENGGUNAAN *DATASET*

Contoh Kalimat Masukan
pilkada membuat penjualan kendaraan meningkat
Stemming
pilkada buat jual kendara tingkat
Stopword
pilkada jual kendara
***Dataset* Fitur**
pilkada: politik jual: ekonomi kendara: lalu lintas
Persentase Luaran Akhir
33% politik 33% ekonomi 33% lalu lintas

Pada penelitian berikutnya, *dataset* ini nantinya dapat diteliti lebih lanjut serta dikembangkan dengan cara menambah jumlah data Tweet yang diproses menjadi lebih banyak, sehingga fitur kata di dalamnya menjadi semakin

lengkap dan topik-topik untuk fitur yang ada menjadi semakin lengkap. Selain itu sumber data dapat diperluas agar data yang diproses semakin banyak, tidak hanya terbatas pada akun-akun Twitter tertentu atau bahkan dapat mengambil sumber lain di luar Twitter. Namun tentunya terdapat beberapa proses tambahan yang perlu dipertimbangkan, seperti normalisasi kata agar segala jenis gaya bahasa atau penulisan, termasuk bahasa tidak formal dan tidak baku, dapat diproses lebih lanjut layaknya Bahasa Indonesia formal.

VI. KESIMPULAN & SARAN

Tingkat akurasi pengujian secara keseluruhan dapat dikatakan cukup baik, dengan persentase paling optimal adalah 64% untuk pemrosesan tiap 200 Tweet. Sistem telah mampu membentuk *dataset* dengan menggunakan konten yang diperoleh dari media sosial Twitter. *Dataset* yang berhasil terbentuk sangat bergantung dengan berita-berita yang sedang populer saat ini. Dengan demikian, *dataset* akan semakin lengkap apabila dibentuk menggunakan konten-konten dengan rentang waktu yang lebih lama.

Kekurangan yang masih terdapat pada penelitian ini berkaitan denngan beberapa konten masih belum memiliki standar yang sama sehingga sulit untuk diproses. Beberapa saran yang dapat dilakukan untuk meningkatkan performa dari sistem adalah dengan menambahkan proses normalisasi kata, terutama untuk kata-kata dalam bahasa asing, kata-kata yang disingkat, atau kata-kata tidak resmi (*slang*). Dengan adanya tambahan proses normalisasi, maka segala jenis konten baik yang standar maupun tidak dapat diolah lebih lanjut, sehingga proses pengujian dapat dilakukan dengan cakupan jenis akun yang lebih luas tidak terbatas pada akun media berita elektronik saja, termasuk data dari akun-akun personal milik perseorangan yang dianggap cukup berpengaruh.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kementrian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia (Ristekdikti) dan Fakultas Teknologi Infomasi Universitas Kristen Duta Wacana yang telah mendukung kegiatan penelitian ini sehingga dapat terlaksana dengan baik. Selain itu penulis juga mengucapkan terima kasih kepada saudari Vievin Efendy sebagai asisten peneliti yang telah banyak membantu penulis selama proses penelitian berlangsung.

DAFTAR PUSTAKA

[1] N. C. Laksana, "Ini Jumlah Total Pengguna Media Sosial di Indonesia," Okezone, 13 Maret 2018. [Online]. Available: https://techno.okezone.com/read/2018/03/13/207/1872093/ini-jumlah-total-pengguna-media-sosial-di-indonesia. [Accessed 26 Juli 2018].

[2] B. Agung, "Pengguna Internet di Indonesia Akses Medsos 3 Jam Per Hari," CNN Indonesia, 2017 Desember 2017. [Online]. Available: https://www.cnnindonesia.com/teknologi/20171218192500-192-263281/pengguna-internet-di-indonesia-akses-medsos-3-jam-per-hari. [Accessed 26 Juli 2018].

[3] M. V. Zaanen and P. Kanters, "Automatic Mood Classification Using TF*IDF Based on Lyrics," in *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 2010.

[4] M. Kompan and M. Bielikova, "Content-based news recommendation," in *International conference on electronic commerce and web technologies*, 2010.

[5] A. R. Lahitani, A. E. Permanasari and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International Conference on Cyber and IT Service Management*, Bandung, 2016.

[6] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons*, vol. 53, no. 1, pp. 59-68, 2010.

[7] T. L. Tuten, Advertising 2.0: social media marketing in a web 2.0 world: social media marketing in a web 2.0 world, ABC-Clio, 2008.

[8] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, pp. 59-68, 2010.

[9] T. O'reilly, *What is web 2.0*, 2005.

[10] A. J. Kim and K. K. Johnson, "Power of consumers using social media: Examining the influences of brand-related user-generated content on Facebook," *Computer in Human Behavior*, vol. 58, pp. 98-108, 2016.

[11] T. Daugherty, M. S. Eastin and L. Bright, "Exploring consumer motivations for creating user-generated content," *Journal of interactive advertising*, vol. 2, no. 2, pp. 16-25, 2008.

[12] K. A. Manap and N. Adzharudin, "The role of user generated content (UGC) in social media for tourism sector," in *The 2013 WEI International Academic Conference Proceedings*, 2013.

[13] A. Z. Bahtar and M. Muda, "The Impact of User--Generated Content (UGC) on Product Reviews towards Online Purchasing-A Conceptual Framework," in *Procedia Economics and Finance*, 2016.

[14] K. Crowston and I. Fagnot, "Stages of motivation for contributing user-generated content: A theory and empirical test," *International Journal of Human-Computer Studies*, vol. 109, pp. 89-101, 2018.

[15] J. Chae, D. Thom, H. Bosch, Y. Jang and R. Maciejewski, "Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition," in *Visual Analytics Science and Technology (VAST)*, 2012.

[16] M. Mathioudakis and N. Koudas, "Twittermonitor: trend detection over the twitter stream," in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, 2010.

[17] W. He, S. Zha and L. Li, "Social media competitive analysis and text mining: A case study in pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464-472, 2013.

[18] S. Inzalkar and J. Sharma, "A survey on text mining-techniques and application," *International Journal of Research In Science & Engineering*, vol. 24, pp. 1-14, 2015.

[19] S. Ahmad and R. Varma, "Information extraction from text messages using data mining techniques," *Malaya Journal of Matematik*, vol. S, no. 1, pp. 26-29, 2018.

[20] D. Agnihotri, K. Verma and P. Tripathi, "Pattern and cluster mining on text data," in *Fourth International Conference on Communication Systems and Network Technologies*, 2014.

[21] S. Vijayarani, J. Ilamathi and Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7-16, 2015.

[22] D. Sebastian, "Rancang Bangun Website Klasifikasi Untuk Pencarian

Produk Pasar Online Menggunakan Algoritma K-Nearest Neighbor," *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 3, no. 3, 2017.

[23] A.-H. Tan, "Text Mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, 1999.

[24] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: Information and pattern discovery on the world wide web," in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. IEEE,* 1997.

[25] R. Kosala and H. Blockeel, "Web mining research: A survey," *ACM Sigkdd Explorations Newsletter,* vol. 2, no. 1, pp. 1-15, 2000.

[26] J. A. Iglesias, A. Tiemblo, A. Ledezma and A. Sanchis, "Web news mining in an evolving framework," *Information Fusion,* vol. 28, pp. 90-98, 2016.

[27] A. R. Chrismanto and Y. Lukito, "Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes," *Jurnal Informatika dan Sistem Informasi,* vol. 2, no. 2, pp. 26-34, 2016.

[28] X. Chen, M. Vorvoreanu and K. Madhavan, "Mining Social Media Data for Understanding Student's Learning Experiences," *IEEE Transactions on Learning Technologies,* vol. 7, no. 3, pp. 246-259, 2014.

[29] R. Kohavi, "Mining E-Commerce Data: The good, the bad, and the ugly," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining,* 2001.

[30] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919,* 2017.

[31] S. A. Salloum, M. Al-Emran, A. A. Monem and K. Shaalan, "Using text mining techniques for extracting information from research articles," *Intelligent Natural Language Processing: Trends and Applications,* pp. 373-397, 2018.

[32] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence,* vol. 1, no. 1, pp. 60-76, 2009.

[33] S. Menaka and N. Radha, "Text classification using keyword extraction technique," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, no. 12, pp. 734-740, 2013.

[34] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and latent semantic indexing," *Journal of King Saud University – Computer and Information Sciences,* vol. 29, no. 2, pp. 189-195, 2017.

[35] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications* , vol. 68, no. 13, pp. 13-18, 2013.