Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XIII Tahun 2018 (ReTII)
November 2018, pp. 103~110
ISSN: 1907-5995

Analisis Trend Akun Media Sosial Twitter dengan TF-IDF dan Cosine Similarity

Kristian Adi Nugraha[1], Danny Sebastian[2]
[1,2] Informatika, Universitas Kristen Duta Wacana
Korespondensi : adinugraha@ti.ukdw.ac.id

ABSTRAK

Media sosial saat ini merupakan media digital yang banyak digunakan oleh masyarakat untuk berinteraksi. Informasi mengenai kejadian atau berita yang sedang terjadi sangat cepat terdistribusi melalui media sosial. Hal ini dikarenakan seluruh masyarakat dapat menjadi sumber informasi, tidak hanya sebagai pembaca seperti halnya pada media massa. Pada penelitian ini, penulis memiliki ide untuk menganalisa trend yang ada pada masyarakat melalui akun-akun media sosial yang telah ditentukan sebelumnya. Proses analisa tersebut dilakukan dengan menggunakan metode *TF-IDF* dan *Cosine Similarity* untuk pembentukkan dataset secara aktual. Hasil pengujian menunjukkan bahwa dataset yang dibentuk secara langsung menggunakan data uji menghasilkan nilai kemiripan yang cukup tinggi terhadap data uji, yaitu sebesar 99.14%.
Kata kunci: Cosine Similarity, Media Sosial, TF-IDF

ABSTRACT

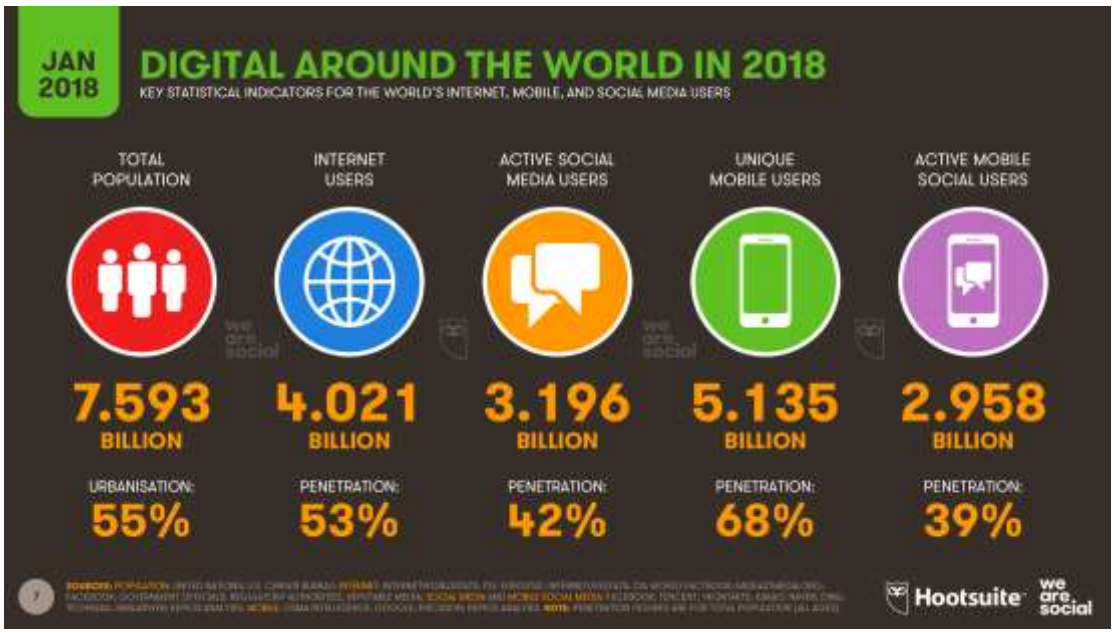
Social media is digital media that most used by people for interaction with other people. Information that already happening is quickly distributed through social media. Because the people can be the source of information, not only as readers like in the mass media. This time, the author has an idea to analyze trends that exist in depends on choosen social media accounts. Analysis process is carried out using the TF-IDF and Cosine Similarity methods to form actual datasets. The test results shows that the dataset produces a high similarity value to the test data, which is 99.14%.
Keyword : Cosine Similarity, Social Media, TF-IDF

1. PENDAHULUAN

Media sosial adalah aplikasi yang berjalan diatas internet dan dibangun berdasarkan teknologi *website 2.0*. Saat ini, pengguna media sosial berasal dari seluruh dunia dengan karakter yang beraneka ragam, mereka saling bertukar informasi, berkolaborasi, dan berbagi konten antar pengguna [1] [2]. Berdasarkan data statistik, jumlah pengguna media sosial aktif pada januari 2018 adalah 3,196 milyar orang [3], dapat dilihat pada Gambar 1. Data yang dihasilkan oleh pengguna media sosial menjadi sangat besar, sehingga dari data tersebut dapat mencerminkan *trend* topik yang sedang banyak dibicarakan oleh masyarakat. Konten yang terdapat pada media sosial terdiri dari beraneka ragam, meliputi promosi, pekerjaan, hobi, bahkan dunia politik. Selain itu, konten media sosial di setiap negara berbeda-beda sesuai dengan kondisi dan kejadian yang ada di negara tersebut. Faktor waktu juga turut mempengaruhi makna konten media sosial, karena sebuah konten yang sama akan memiliki makna berbeda jika berada pada waktu yang berbeda.

Pada beberapa tahun terakhir, obyek dari *text mining* yang banyak diteliti adalah *website* atau *world wide web* [6] [7]. *World wide web* memiliki banyak konten dokumen teks yang dapat diolah lebih lanjut menggunakan *text mining*, khususnya pada media social di mana terdapat banyak pengguna aktif dan data di dalamnya terus bertambah setiap saat [1] [8] [9]. Untuk dapat melihat *trend* yang terjadi berdasarkan data atau konten di media sosial, dibutuhkan metode atau algoritma yang dapat melakukan ekstraksi informasi atau trend secara otomatis [4]. Salah satu metode yang digunakan untuk melakukan analisis terhadap konten media sosial yang berbentuk teks adalah dengan *text mining* [5]. Saat ini pada beberapa media sosial telah memiliki fitur untuk melihat *trend* secara global di seluruh dunia, namun untuk belum terdapat fitur untuk melihat *trend* pada cakupan yang lebih sempit misalnya pada satu wilayah negara saja atau bahkan pada satu akun saja. Salah satu kendala untuk pengolahan *text mining* pada media sosial adalah fitur yang digunakan untuk proses ekstraksi cukup kompleks, karena fitur tersebut harus dapat menyesuaikan kondisi wilayah di mana konten tersebut berada serta waktu di mana konten tersebut muncul.

Pada penelitian ini, penulis akan merancang sistem yang dapat digunakan untuk mengetahui *trend* dari akun yang dipilih. Pembentukan fitur dilakukan menggunakan data aktual dengan metode *TF-IDF* dan *Cosine Similarity*, sehingga bisa menyesuaikan dengan kondisi yang ada pada waktu tersebut. Dengan demikian, hasil trend yang didapat memiliki tingkat akurasi yang lebih tinggi.



Gambar 1. Statistik pengguna media digital 2018

Dikutip dari: D. Chaffey, "Smart Insights - Global social media research summary 2018," Smart Insights, 28 March 2018. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Acesso em 10 September 2018]

Berdasarkan latar belakang diatas, tim penulis mencoba berkontribusi dengan melakukan penelitian untuk melihat trend yang ada di media sosial twitter. Penulis membangun sistem yang dapat melakukan pembobotan menggunakan TF-IDF dan menghitung cosine similarity untuk menentukan trend yang ada. Luaran dari penelitian ini adalah sistem yang dapat menentukan trend yang ada di media sosial, khususnya media sosial twitter.

2. LANDASAN TEORI

A. 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 [10]. Proses yang dilakukan dalam *text preprocessing* adalah:

1. Tokenisasi
   Dalam proses tokenisasi, dokumen teks akan dipecah menjadi sebuah *token* atau sebuah kata [11]. 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* [10]. *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 [12]. *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* [12]. 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 [10]. 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.

B. 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 [12] [13]. 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-ID*F 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

C. Cosine Similarity

Cosine similarity merupakan metode pengukuran yang banyak digunakan di *pattern recognition* dan *text classification* [14]. *Cosine similarity* mengukur kemiripan dua buah vektor dalam sebuah *product space* dengan mengukur cosine dari sudut kedua vektor [15]. 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.

3. METODE PENELITIAN

Penelitian ini dilakukan dengan 6 tahap seperti yang tertulis pada Gambar 2. Berikut ini penjelasan untuk masing-masing tahap penelitian.

Gambar 2. Tahapan penelitian

A. Studi Pustaka

Studi pustaka adalah tahapan awal penelitian, tim penulis melakukan pencarian refrensi terkait media sosial, TF-IDF, *document similarity*, dan analisis *trend*. Selain itu tim penulis mempelajari karakteristik meta data yang ada di konten twitter, hal ini diperlukan untuk mendapatkan metode meta data yang dapat diolah dalam penelitian ini.

B. Pengumpulan Data

Tahap pengumpulan data dilakukan hampir bersamaan dengan studi pustaka, dimana tim penulis mengumpulkan data twit yang akan digunakan untuk penelitian. Akun twitter yang digunakan pada penelitian ini dibatasi dengan akun media elektronik yang terverifikasi. Informasi verifikasi akun resti atau official ditandai dengan tanda centang dengan latar belakang biru disebelah akun twitter. Akun yang dipilih dapat dilihat pada Tabel 1.

C. Perancangan Sistem

Setelah pengumpulan data, tim penulis melakukan perancangan system yang akan dibangun. Hal yang diperhatikan pada tahap ini adalah struktur data yang diperlukan untuk menyimpan informasi data twit. Selain itu system dirancang dapat mengolah data twit menjadi trend menggunakan metode pembobotan TF-IDF dan *Cosine similarity*.

D. Pembangunan Sistem

Sistem dibangun dalam platform aplikasi website menggunakan bahasa pemrograman *php* dan *frameworks Laravel*. *Text editor* yang digunakan untuk pengembangan system adalah *Jetbrains PHP Storm*. *Web server* yang digunakan adalah apache, sedangkan *database server* yang digunakan adalah MySQL.

E. Pengujian Sistem

Data tweet akan dikumpulkan dari tanggal 10 September 2018 sampai 16 September 2018. Setengah data akan dipilih untuk menjadi data latih, sedangkan setengah data akan digunakan untuk data uji. Cara pembagian data dilakukan dengan cara bergantian, misalnya data pertama untuk data latih, data kedua untuk data uji, data ketiga untuk data latih, dan seterusnya. Daftar akun twitter dan jumlah data twit dapat dilihat pada Tabel *1*.

Tabel 1 daftar akun twitter, data latih, dan data uji

Akun Twitter	Jumlah data	Jumlah data latih	Jumlah data uji
BBC Indonesia (@bbcindonesia)	116	58	58
CNN Indonesia (@CNNIndonesia)	1255	628	627
Detik (@detikcom)	2262	1131	1131
JPNN (@jpnncom)	1372	686	686
Kompas (@kompascom)	2773	1387	1386
Kontan (@kontannews)	1097	549	548
Koran Tempo (@korantempo)	2021	1011	1010
Media Indonesia (@mediaindonesia)	306	153	153
Okezone News (@okezonenews)	554	277	277
Republika Online (@republikaonline)	790	395	395
Tempo (@tempodotco)	2088	1044	1044

Akun-akun yang ditunjukkan pada Tabel 1 merupakan akun-akun media massa pada platform Twitter yang kredibel (terverifikasi oleh Twitter). Penulis memilih akun media massa dengan pertimbangan konten di dalamnya menggunakan Bahasa Indonesia yang baku dan formal, sehingga akan lebih mudah diolah. Selain

itu media massa merupakan cerminan dari kondisi yang ada pada suatu negara, dengan demikian apa yang menjadi konten di dalamnya merupakan representasi dari apa yang sedang terjadi di masyarakat.

Data tweet dari seluruh akun yang telah dipilih untuk menjadi data uji akan diolah lebih lanjut menggunakan TF-IDF dan cosine similarity untuk menghasilkan dataset fitur secara otomatis. Dataset berisi kelompok-kelompok trend yang terbentuk berdasarkan tweet-tweet sejenis menurut perhitungan TF-IDF. Seluruh tweet di dalam masing-masing kelompok akan diambil kata kuncinya, kemudian diberi label yang sesuai dengan isi dari kelompok tersebut. Dengan demikian, dataset fitur tersebut nantinya akan berisi kata kunci untuk masing-masing trend yang ada. Proses pengujian akhir dilakukan dengan cara menguji data yang terpilih sebagai data latih dan menguji data yang terpilih dengan data uji, kemudian membandingkan hasil persentase keduanya.

4. HASIL DAN ANALISIS

Hasil pengujian yang telah dilakukan oleh penulis terhadap data tweet tanggal 10 hingga 16 September 2018 seperti yang ditunjukkan pada Tabel 2 dan 3. Pada tabel tersebut dapat terlihat trend gabungan dari seluruh akun yang telah ditentukan sebelumnya. Pada tabel tersebut, dapat terlihat bahwa topik yang cukup konsisten dengan nilai persentase cukup tinggi dari awal hingga akhir adalah Asian Games, Ekonomi, dan Pemilihan Presiden (Pilpres). Hal ini cukup wajar karena pada tanggal tersebut, masyarakat di Indonesia banyak membahas mengenai topik tersebut.

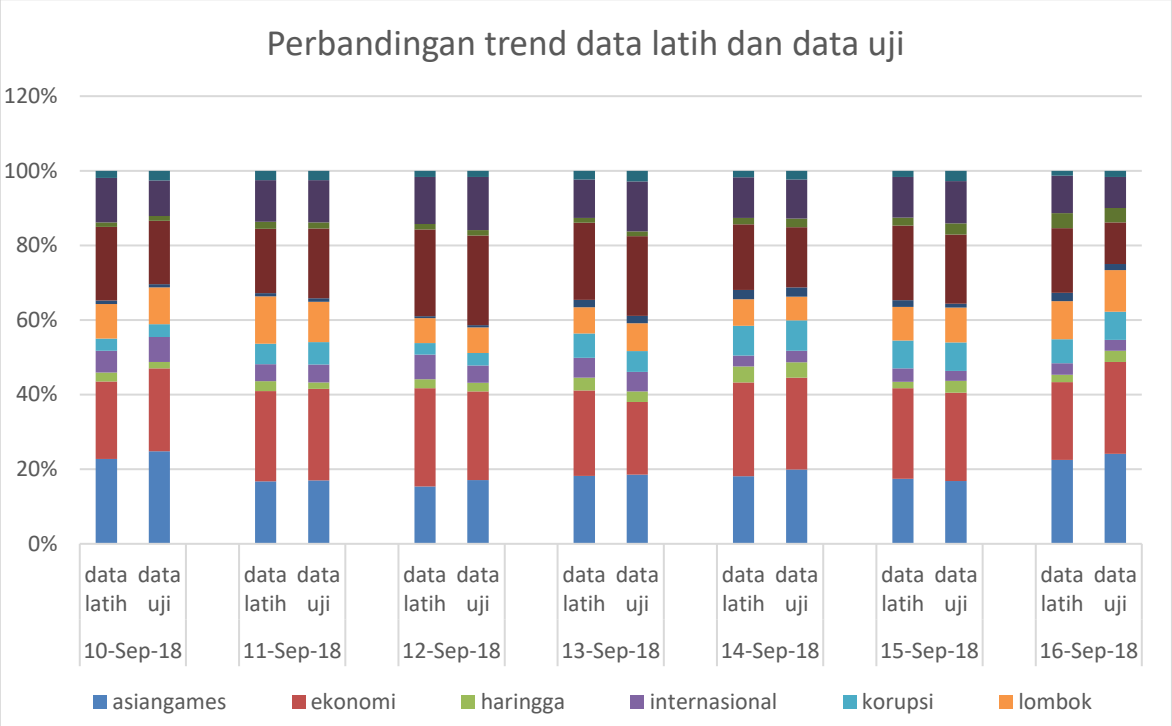
Tabel 2 Hasil pengujian trend tanggal 10-13 September 2018

	10-Sep-18			**11-Sep-18**			**12-Sep-18**			**13-Sep-18**		
	data latih	**data uji**	**selisih**	**data latih**	**data uji**	**selisih**	**data latih**	**data uji**	**selisih**	**data latih**	**data uji**	**selisih**
asiangames	23%	25%	2%	17%	17%	0%	15%	17%	2%	18%	19%	0%
ekonomi	21%	22%	1%	24%	25%	0%	26%	24%	3%	23%	19%	3%
sepak bola	2%	2%	1%	3%	2%	1%	2%	2%	0%	3%	3%	1%
internasional	6%	7%	1%	5%	5%	0%	7%	5%	2%	5%	5%	0%
korupsi	3%	3%	0%	5%	6%	0%	3%	3%	0%	7%	6%	1%
lombok	9%	10%	1%	13%	11%	2%	7%	7%	0%	7%	7%	0%
narkoba	1%	1%	0%	1%	1%	0%	0%	1%	0%	2%	2%	0%
pilpres	20%	17%	3%	17%	19%	1%	23%	24%	1%	21%	21%	1%
pns	1%	1%	0%	2%	2%	0%	1%	1%	0%	1%	1%	0%
politik	12%	10%	2%	11%	11%	0%	13%	14%	2%	10%	13%	3%
reklamasi	2%	3%	1%	3%	3%	0%	2%	2%	0%	2%	3%	0%
RATA-RATA SELISIH			**1.09%**			**0.36%**			**0.91%**			**0.82%**

Tabel 3 Hasil pengujian trend tanggal 14-16 September 2018

	14-Sep-18			**15-Sep-18**			**16-Sep-18**		
	data latih	**data uji**	**selisih**	**data latih**	**data uji**	**selisih**	**data latih**	**data uji**	**selisih**
asiangames	18%	20%	2%	17%	17%	1%	22%	24%	2%
ekonomi	25%	25%	0%	24%	24%	0%	21%	25%	4%
sepak bola	4%	4%	0%	2%	3%	1%	2%	3%	1%
internasional	3%	3%	0%	4%	3%	1%	3%	3%	0%
korupsi	8%	8%	0%	7%	8%	0%	6%	8%	1%
lombok	7%	6%	1%	9%	9%	0%	10%	11%	1%
narkoba	2%	2%	0%	2%	1%	1%	2%	2%	1%
pilpres	18%	16%	1%	20%	19%	1%	17%	11%	6%
pns	2%	2%	1%	2%	3%	1%	4%	4%	0%
politik	11%	10%	0%	11%	11%	0%	10%	8%	2%
reklamasi	2%	2%	1%	2%	3%	1%	1%	2%	0%
RATA-RATA SELISIH			**0.55%**			**0.64%**			**1.64%**

Secara keseluruhan, rata-rata selisih persentase dari data latih dan data uji adalah 0.86%. Hal ini menunjukkan bahwa hasil dataset yang dilatih dengan menggunakan data latih yang terpilih dapat mencakup kata kunci dari seluruh data uji yang ada, dengan tingkat kesamaan trend antara data latih dan data uji sebesar 99.14%. Proporsi dari trend yang dihasilkan dapat dilihat pada Gambar 3. Pada gambar tersebut terdapat grafik batang yang menunjukkan proporsi dari trend seluruh akun terpilih antara data latih dan data uji. Perbandingan kedua grafik memiliki perbedaan yang kecil (0.86%).



Gambar 3. Grafik perbandingan trend data uji dan data latih

5. KESIMPULAN

Berdasarkan hasil pengujian yang telah dilakukan oleh penulis, sistem yang dibangun telah mampu menghasilkan trend dari akun-akun media sosial yang terpilih dengan tingkat akurasi sebesar 99.14% dengan dataset yang dibentuk secara otomatis menggunakan data aktual atau terbaru. Pengujian tersebut dilakukan pada rentang waktu satu minggu pada akun-akun media massa. Pada penelitian berikutnya, jenis akun yang digunakan dapat diperluas mencakup akun-akun personal atau akun jenis yang lain dengan penambahan metode normalisasi di dalamnya karena konten-konten pada akun tersebut belum tentu menggunakan bahasa yang standar dan baku.

UCAPAN TERIMAKASIH

Tim penulis mengucapkan terima kasih kepada Kementrian Riset Teknologi dan Pendidikan Tinggi Republik Indonesia (Ristekdikti) dan Fakultas Teknologi Infomasi Universitas Kristen Duta Wacana (UKDW) 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] A. M. Kaplan e M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Business horizons,* vol. 53, nº 1, pp. 59-68, 2010.

[2] 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.

[3] D. Chaffey, "Smart Insights - Global social media research summary 2018," Smart Insights, 28 March 2018. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/. [Acesso em 10 September 2018].

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

[5] W. He, S. Zha e 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.

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

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

[8] A. R. Chrismanto e 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.

[9] X. Chen, M. Vorvoreanu e 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.

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

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

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

[13] S. Menaka e 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.

[14] F. S. Al-Anzi e 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.

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

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

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

[18] A. J. Kim e 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.

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

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

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

[22] K. Crowston e 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.

[23] D. Jiang, X. Luo, J. Xian e Z. Xu, "Sentiment Computing for the News Events Based on the Social Media Big Data," *IEEE Access,* vol. 5, pp. 2373-2382, 2017.

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

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

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

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

[28] 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.

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

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

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

[32] V. Gupta e 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.