p-ISSN : 2443-2210
e-ISSN : 2443-2229

Jurnal Teknik Informatika dan Sistem Informasi
Volume 5 Nomor 1 April 2019

Implementasi Algoritma *K-Nearest Neighbor* untuk Melakukan Klasifikasi Produk dari beberapa *E-marketplace*

http://dx.doi.org/10.28932/jutisi.v5i1.913

Danny Sebastian[#1]

[#1]*Fakultas Teknologi Informasi, Universitas Kristen Duta Wacana*
Jl. Dr. Wahidin Sudirohusodo 5-25 Yogyakarta 55224
[1]danny.sebastian@staff.ukdw.ac.id

Abstract — **E-marketplace has gained popularity with the Indonesian society resulting in the increment of products offered. Consequently, customers require more effort to search for products. In this study, we classified products from several e-marketplaces. The classification was carried out using TF-IDF method for the weighting, cosine similarity to calculate product similarity distance, and k-nearest neighbor algorithm. Based on the first testing result using 150 product data, the k-nearest neighbor method with k=5 successfully classified 146 data with 4 data classified into the wrong class. This k=5 value gives the best result for this case, with an accuracy of 97.33%. The second testing result using 150 mixed brand product data, the k-nearest neighbor method successfully classified 145 data with 5 data classified into the wrong class. The accuracy of the second testing is 96.67%.**

Keywords— ***e-marketplace, K-Nearest Neighbor, classification, web mining***

I. PENDAHULUAN

E-marketplace adalah istilah yang digunakan untuk menyebut pasar online. *E-marketplace* memungkinkan transaksi antara konsumen dengan konsumen atau *Customer to Customer* (*C2C*). Pada kuartal kedua (Juli) tahun 2018, perkembangan *e-marketplace* mencapai 229 juta pengguna [1], dapat dilihat pada Gambar 1. Saat ini, penduduk Indonesia mulai merubah pola jual beli. Pola jual-beli yang dilakukan di pasar tradisional, berubah menggunakan media elektronik. Dengan adanya perubahan pola jual-beli masyarakat, transaksi dapat dilakukan dengan mudah, dimana saja dan kapan saja melalui media internet. Hal ini memungkinkan penjual memperluas area penjualannya tanpa perlu membuka gerai fisik baru. Oleh sebab itu, banyak sekali penjual yang berlomba-lomba menjual produknya di *e-marketplace*.

Gambar 2 merupakan data transaksi *e-commerce* yang ada di Indonesia berdasarkan kategori. Menurut data tersebut, pada awal tahun 2018, kategori yang paling banyak diminati di Indonesia adalah *fashion* dan kecantikan, *travel*, mainan dan hobi, *home furniture* dan perangkat elektronik [2].

Gambar 1. Peta *e-commerce* Indonesia
Dikutip dari: https://iprice.co.id/insights/mapofecommerce/

Gambar 2. *E-commerce spend by category*
Dikutip dari: https://www.slideshare.net/wearesocial/digital-in-2018-in-southeast-asia-part-2-southeast-86866464

Dengan kemudahan yang ditawarkan, *e-marketplace* menjadi semakin digemari dan semakin bertambah jumlah penjual dan pembelinya. Hal tersebut membuat sebuah permasalahan, yaitu semakin ketat persaingan antar penjual. Karena persaingan tersebut, penjual menjadi harus mencari cara agar produknya menarik dan dibeli oleh pembeli. Salah satunya adalah dengan cara membuat judul/nama produk yang tidak sesuai, seperti menambahkan kata "gratis", "garansi", dan lain sebagainya. Ada juga penjual yang mencoba menambahkan merk lain terkenal dibelakang judul/nama produk untuk meningkatkan peringkat pada proses pencarian, seperti "Apple iPhone 6 bukan Xiaomi, Oppo, Vivo". Pada sisi lain, pembeli jadi membutuhkan usaha ekstra untuk mencari produk yang memang sesuai dengan keinginannya. Selain itu dengan bertambahnya *e-marketplace*, pengguna memerlukan usaha ekstra untuk mencari di beberapa situs *e-marketplace*.

Text data mining atau *text mining* merupakan proses untuk mencari atau melakukan ekstraksi informasi dari data teks [3]. Pada *text mining*, data yang bersifat tidak terstruktur, diolah melalui *text preprocessing* hingga menghasilkan informasi untuk diolah lebih lanjut oleh algoritma *mining*. Salah satu bentuk algoritma *mining* adalah untuk klasifikasi. Dimana klasifikasi bertujuan untuk mengelompokkan data secara otomatis oleh sistem ke kelas-kelas yang sudah ditentukan berdasarkan karakteristik pada suatu kelas tersebut.

Secara umum, *e-marketplace* memiliki konten dalam bentuk teks dan gambar. Konten teks pada *e-marketplace* dapat berisi judul/nama produk, deskripsi produk, maupun komentar atau testimoni pengguna. Hal ini memungkinkan data teks yang berasal dari *e-marketplace* dapat diolah menggunakan algoritma *text mining* menjadi informasi yang berguna. Salah satu pemanfaatan *text mining* pada *e-marketplace* adalah menggunakan *klasifikasi* untuk mengelompokkan data produk yang sama, sehingga pembeli dapat lebih mudah melakukan pencarian produk yang memang sesuai dengan keinginannya.

Berdasarkan latar belakang permasalahan diatas, rumusan permasalahan yang diangkat penulis adalah penerapan algoritma *K-Nearest Neighbor* untuk melakukan klasifikasi data produk dari beberapa *e-marketplace* di Indonesia. Luaran dari penelitian ini adalah penentuan nilai k yang sesuai untuk algoritma *K-Nearest Neighbor* pada kasus klasifikasi produk dari *e-marketplace* di Indonesia dan melihat pengaruh pemberian *nama produk* dalam klasifikasi.

II. TINJAUAN PUSTAKA

Pada tahun 2015, Lopes, Prajyoti, and Bidisha Roy melakukan penelitian untuk membuat sistem rekomendasi berdasarkan data navigational dari *website* menggunakan metode *web mining*. Penulis menggumpulkan *user click stream data* dari sebuah website *e-commerce*. Data mentah tersebut dilakukan *preprocessing* berupa *field separation, data cleaning, user differentiation, session identification, session clustering*, dan *data formatting*. Simpulan dari penelitian ini adalah system rekomendasi dapat memberikan hasil yang baik berdasarkan pola pengguna. Hasil dari system mengurangi kesalahan klasik dari system rekomendasi, yaitu pengurangan hasil yang *false positif*.

Penelitian yang dilakukan oleh Adeniyi DA, Wei Z, dan Yongquan Y [4] bertujuan untuk membuat system rekomendasi berita menggunakan data yang didapatkan dari *Really Simple Syndication* (RSS). Algoritma klasifikasi *K-Nearest Neighbor* dilatih untuk dapat melihat pola klik dari pengguna, kemudian mengelompokkannya ke sebuah user group, kemudian memberikan rekomendasi informasi browsing sesuai dengan karakteristik pengguna. Proses penelitian dilakukan dengan data RSS di-ekstraksi, lalu dibersihkan datanya, diformat, dan dikelompokkan sehingga

membentuk sebuah dataset yang dapat digunakan. Untuk menghitung jarak antar node, digunakan metode *eucledian distance*. Hasil dari penelitian ini, imeplementasi algoritma *K-Nearest Neighbor* dapat memberikan manfaat dan memberikan akurasi yang baik, sehingga rekomendasi yang dihasilkan sesuai dengan kebutuhan atau minat pengguna.

Penelitian terdahulu yang pernah dilakukan oleh penulis adalah "Rancang Bangun Website Klasifikasi untuk Pencarian Produk Pasar Online menggunakan Algoritma *K-Nearest Neighbor*" [5]. Pada penelitian tersebut dilakukan simulasi perhitungan menggunakan data produk handphone. Hasil dari penelitian tersebut adalah pembuktian secara simulasi bahwa metode K-nearest neighbour dapat digunakan untuk kasus tersebut dan perancangan yang digunakan untuk membuat aplikasi website pada penelitian aktual.

Penelitian menggunakan algoritma *K-Nearest Neighbor* juga dilakukan oleh W. E. Nurjanah, R. S. Perdana dan M. A. Fauzi pada tahun 2017 [6]. Penulis mengumpulkan data tweet dari media sosial twitter, dimulai dari tahapan *preprocessing*. Pada tahapan *preprocessing*, dilakukan tokenisasi, *data cleansing, case folding, filterisasi*, dan *stemming*. Tahapan selanjutnya adalah pembobotan menggunakan metode *TF-IDF*. Hasil pembobotan dihitung jarak kemiripannya menggunakan *cosine similarity*, kemudian diolah menggunakan algoritma *K-Nearest Neighbor*. Tahap selanjutnya adalah melakukan pembobotan jumlah *Retweet* dan dinormalisasi menggunakan metode *min-max*. Kemudian hasil penggabungan tersebut mengahasilkan sebuah nilai, yang menunjukkan hasil klasifikasi dokumen positif atau negatif.

TABEL I
PERBEDAAN PENELITIAN AKTUAL DENGAN PENELITIAN TERDAHULU

Judul	*Dynamic Recommendation System Using Web Usage Mining for E-commerce Users*	*Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classificationmethod*	Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode *K-Nearest Neighbor* dan Pembobotan Jumlah *Retweet*	Klasifikasi Dokumen Menggunakan *Metode K-nearest neighbour* (KNN) dengan *Information Gain*.
Oleh	P. Lopes and B. Roy	Adeniyi DA, Wei Z, Yongquan Y	W. E. Nurjanah, R. S. Perdana and M. A. Fauzi	P. D. Nugraha, S. A. Faraby and Adiwijaya
Tahun	2015	2016	2017	2018
Metode	*User Interest Measure, Wish list buffer, Similarity Measures*	*Eucledian distance, K-Nearest Neighbor*	*K-Nearest Neighbor,* dengan pembobotan jumlah *retweet* menggunakan *min-max*	*K-Nearest Neighbor* dan *Information Gain*
Obyek	*Navigational data* di *website*	*Really Simple Syndication* (RSS)	Tweet dari media sosial twitter	*Text Categorization Collection Data Set*

TABEL II
PERBEDAAN PENELITIAN AKTUAL DENGAN PENELITIAN TERDAHULU (LANJUTAN)

Judul	Rancang Bangun Website Klasifikasi untuk Pencarian Produk Pasar Online menggunakan Algoritma *K-Nearest Neighbor*	Implementasi Algoritma *K-Nearest Neighbor* Untuk Melakukan Klasifikasi Produk dari Beberapa *E-marketplace*.
Oleh	Danny Sebastian	Danny Sebastian
Tahun	2017	2018
Metode	*TF-IDF, Eucledian distance,* dan *K-Nearest Neighbor*	*TF-IDF, Eucledian Distance* dan *K-Nearest Neighbor*
Obyek	Produk dari tokopedia dan bukalapak	Produk dari tokopedia dan bukalapak.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

Pada tahun 2018, P. D. Nugraha, S. A. Faraby and Adiwijaya melakukan penelitian klasifikasi dokumen dan dikombinasikan dengan metode *Information Gain* [7]. Penelitian tersebut dilakukan dengan cara mencari atribut dari dataset, kemudian melakukan seleksi fitur menggunakan metode *information gain* kemudian diolah menggunakan klasifikasi *K-Nearest Neighbor*. Evaluasi dilakukan dengan cara membandingkan metode *K-Nearest Neighbor* dan kombinasi *K-Nearest Neighbor* dengan *information gain*. Kedua skenario tersebut dibandingkan waktu pemrosesan dan akurasi hasil pemrosesan.

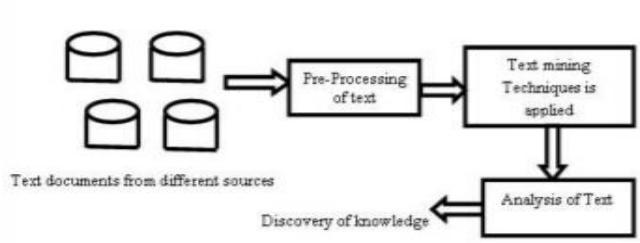
Berdasarkan tinjauan pustaka yang sudah dilakukan, maka diperoleh kesimpulan bahwa algoritma klasifikasi *K-Nearest Neighbor* dapat digunakan untuk mengelompokkan produk dari beberapa *e-marketplace*. Oleh karena itu, penelitian aktual dilakukan menggunakan metode pembobotan TF-IDF dan algoritma klasifikasi *K-Nearest Neighbor*. Perbedaan penelitian aktual dengan penelitian terdahulu dapat dilihat pada TABEL I dan TABEL II.

III. LANDASAN TEORI

A. *Text Mining*

Text mining atau *text data mining*, adalah bidang pengetahuan yang mencakup area *information retrieval* [8], *text analysis*, *information extraction* [9], *clustering* [10], *categorization* [3] [5], *visualization*, *database technology*, *machine learning*, dan *data mining* [11].

Dalam beberapa tahun terakhir, obyek dari *text mining* yang banyak diteliti adalah aplikasi *website* atau *world wide web* [12] [13]. Pada umumnya, aplikasi website memiliki banyak konten dokumen teks yang dapat diolah lebih lanjut menggunakan algoritma *text mining*. Beberapa bidang aplikasi website yang digunakan menjadi objek penelitian antara lain website berita [14], media sosial [15] [16] [17], *e-commerce* [18] , dan lainnya.

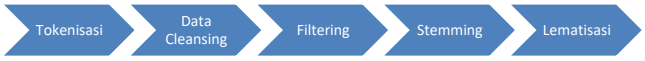


Gambar 3. Proses *text mining*
Dikutip dari: Vijayarani, S., Ms J. Ilamathi, and Ms Nithya, 2015, *Preprocessing Techniques for Text Mining – An Overview*

Tahapan dari *text mining* dapat dilihat pada Gambar 3. Adapun langkah pertama dari *text mining* adalah *data acquisition*. Dimana *data* tidak terstruktur dikumpulkan dari satu sumber atau beberapa sumber. Kemudian data tersebut diolah melalui tahap *text preprocessing*. Pada tahap text preprocessing, *data* tidak terstruktur yang sudah dikumpulkan akan diolah menjadi *data* terstruktur. Hasil dari tahap text preprocessing akan diolah kembali menggunakan algoritma *text mining*, seperti *clustering*, *information extraction*, *classification*, dan lain sebagainya. Keluaran dari algoritma tersebut dianalisis untuk menjadi pengetahuan.

B. *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 [19]. Langkah dari text preprocessing dapat dilihat pada Gambar 4.



Gambar 4. Langkah *text preprocessing*

Adapun penjelasan untuk masing-masing proses yang dilakukan dalam *text preprocessing* adalah:

1. Tokenisasi
   Dalam proses tokenisasi, dokumen teks akan dipecah menjadi sebuah *token* atau sebuah kata [20]. Dalam proses tokenisasi, dilakukan penyesuaian tipe kapitalisasi teks.
2. *Data Cleansing*
   Token yang dihasilkan dari proses tokenisasi akan dihapus karakter *special* dan tanda baca. Pada tahap ini, dapat dilakukan perubahan simbol yang berupa fitur menjadi token yang berguna dan memiliki makna untuk pemrosesan selanjutnya.
3. *Filtering*
   Setiap *token* yang sudah dihasilkan pada tahap sebelumnya, akan dibersihkan dari *stop word*, atau disebut juga dengan proses *filtering* [19]. *Stop word* merupakan kata yang dianggap tidak mencerminkan *keyword* dari dokumen. Manfaat menghilangkan *stop word* adalah mengurangi dimensi dari jumlah kata yang akan diolah, hal ini dapat mempercepat proses analisis menggunakan algoritma *text mining* [3]. Daftar stop word akan disesuaikan dengan Bahasa dan karakteristik dari dokumen yang di proses.
4. *Stemming*
   Stemming adalah metode yang digunakan untuk menghasilan *stem/root*/kata dasar dari sebuah *token* [3]. Tujuan dari proses *stemming* adalah menghilangkan imbuhan, sehingga mengurangi jumlah dari kata yang diproses dalam *text mining*, menghemat waktu, dan menghemat memori.

5. *Lematisasi*
   Lematisasi adalah proses mengubah sebuah kata menjadi bentuk yang sesuai (*lemma*), sehingga dapat dikelompokkan dengan kata lain yang sama [19]. 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.

C. Klasifikasi

Klasifikasi teks merupakan kegiatan untuk menentukan sebuah dokumen tergabung kedalam sebuah kelas dokumen [9] [10]. Dengan berkembangnya jumlah dokumen, diperlukan sebuah tools untuk melakukan *automatic classification*. Automatic classification adalah proses klasifikasi yang dilakukan oleh komputer [21].

Berdasarkan jumlah kelas terdapat 2 tipe klasifikasi, yaitu *binary classification* dan *multi-class classification* [22]. Binary classification merupakan klasifikasi sebuah obyek ke salah satu kelas dari dua kelas yang ditentukan. Sedangkan *multi-class classification* adalah klasifikasi sebuah objek ke satu atau lebih kelas.

Dalam klasifikasi teks, obyek *data* dibagi menjadi 2, yaitu *data* latih dan *data* uji [22]. *Data* latih adalah *data* dokumen yang sudah diklasifikasikan secara manual, sedangkan *data* uji adalah dokumen yang belum diklasifikasikan dan akan digunakan sebagai data pengujian. Tujuan dari membagi *data* ini adalah mendapatkan pengetahuan karakteristik kelas berdasarkan *data* latih, dan menerapkan nya ke *data* uji secara akurat.

D. Term Frequency – Inverse Document Frequency

Terms Frequency & *Inverse Document Frequency* (*TF-IDF)* merupakan metode pembobotan secara statistic. Metode *TF-IDF* menunjukkan seberapa penting sebuah kata pada sebuah dokumen yang terletak pada sebuah kelompok [3] [23]. Metode pembobotan *TF-IDF* biasanya digunakan dalam *text mining*. *Term frequency (TF)* adalah jumlah sebuah kata pada sebuah 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. Semakin kecil nilai *IDF* menunjukkan suatu kata muncul pada banyak dokumen. Sedangkan nilai IDF akan semakin besar apabila suatu kata hanya muncul pada 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

E. Cosine Similarity

Cosine similarity merupakan metode pengukuran yang banyak digunakan di *pattern recognition* dan *text classification* [24]. *Cosine similarity* mengukur kemiripan dua buah vektor dalam sebuah *product space* dengan mengukur cosine dari sudut kedua vektor [25]. 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.

F. K-Nearest Neighbor

K-Nearest Neighbor atau *KNN* merupakan metode klasifikasi dengan berdasarkan jarak data baru ke beberapa data atau tetangga terdekat [26]. Pendekatan *Nearest Neighbor* adalah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah

p-ISSN : 2443-2210
e-ISSN : 2443-2229

fitur yang ada [26] [27]. Tujuan dari algoritma *Nearest Neighbor*, jarak antara satu data dengan data yang lain dapat dihitung. Nilai jarak hasil perhitungan dapat digunakan sebagai nilai kedekatan atau kemiripan antara data uji dengan data latih [26].

Nilai K Pada *Nearest Neighbor* berarti K-data terdekat dari data uji [26]. Jika K bernilai 2, akan diambil 2 tetangga terdekat dari data latih, begitu juga bila K bernilai n maka akan diambil sejumlah n tetangga terdekat dari data latih. Permasalahan dari metode *nearest neighbor* adalah pemilihan nilai K yang tepat.

Pengkategorian teks adalah proses pengelompokan document teks menjadi satu atau lebih kategori yang telah ditetapkan berdasarkan isi kontennya [26] [28]. Sejumlah teknik klasifikasi statisik dan mesin pembelajaran telah ditetapkan untuk pengkategorian teks. Langkah pertama dalam pengkategorian teks adalah mengubah dokumen yang biasanya merupakan string karakter menjadi sebuah representasi yang cocok untuk algoritma pembelajaran dan tugas pengklasifikasian.

Langkah dari metode *KNN* [29] adalah:

1. Diasumsikan ada j training kategori (C_1, C_2,…,C_j) dan total seluruh sampel training adalah N. Kedua komponen tersebut menjadi *m-dimension feature vector*, yaitu sebuah bidang *vector* dimana hasil klasifikasinya akan ditandai di bidang tersebut.
2. Jadikan sampel X menjadi sama dengan *feature selection vector* pada bidang tersebut (X_1, X_2, …X_m)
3. Hitung kesamaan antara seluruh training sampel dengan X. Untuk menghitung kesamaan dapat dilakukan dengan rumus *Cosine Similarity* [29] [30].
4. Lakukan perhitungan kecenderungan dari X untuk setiap kategori menggunakan rumus dibawah ini [29] [31].

$$P(X, C_j) = \sum_d Cosine(x, d_i).y(y_i, C_j) \qquad [5]$$

Dengan:
y (y_i, C_j) = fungsi kategori atribut

5. Tentukan X masuk pada kategori apa berdasarkan bilangan terbesar pada perhitungan P (X, C_j).

IV. METODE PENELITIAN

Penelitian yang akan dilakukan oleh penulis terdiri dari enam tahap seperti yang ditunjukkan pada Gambar 5. Pada bagian ini akan dijelaskan tahapan penelitian yang dilakukan.

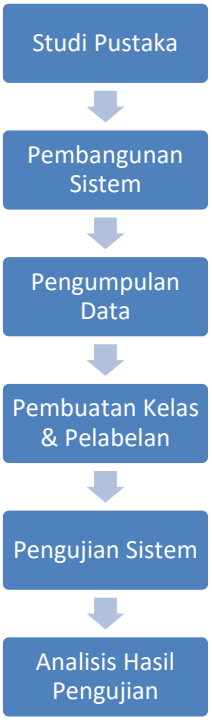
A. Studi Pustaka

Penulis melakukan studi mengenai metode dan algoritma yang akan digunakan, yaitu metode TF-IDF, *cosine similarity*, dan algoritma *K-Nearest Neighbor*. Selain itu penulis juga mempelajari tentang *struktur data* dari produk di *e-marketplace* tokopedia dan bukalapak yang akan diteliti.

B. Pembangunan Sistem

Pada tahap ini, penulis melakukan pembangungan sistem sesuai dengan hasil penelitian terdahulu, yang berjudul "Rancang Bangun Website Klasifikasi untuk Pencarian Produk Pasar Online menggunakan Algoritma *K-Nearest Neighbor*". Sistem dibangun berbasis *website*, menggunakan Bahasa pemrograman PHP dengan *frameworks* Laravel. Penulis memilih *platform website* karena *platform website* secara langsung terhubung ke jaringan *internet*, sehingga proses komunikasi dengan *server e-marketplace* dapat dilakukan dengan 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



Gambar 5. Langkah penelitian

C. Pengumpulan Data

Pada penelitian ini, penulis menentukan sebuah kategori produk yang akan diteliti, yaitu kategori *handphone*. Kategori *handphone* dipilih karena merupakan salah satu kategori yang banyak diminati oleh masyarakat di Indonesia, dan memiliki nama produk yang ditentukan oleh produsen *handphone*. Nama produk tersebut akan digunakan menjadi nama kelas klasifikasi.

Pada tahap ini, penulis mengumpulkan data produk dari kategori yang ditentukan, yaitu kategori handphone. Dari semua merk *handphone*, akan dipilih 5 merk yang akan

dijadikan obyek penelitian, yaitu Apple, Asis, Oppo, Vivo, dan Xiaomi.

Pada penelitian ini, penulis menggunakan *stopword* yang berasal dari Tala, F.Z [32]. Berdasarkan data yang dikumpulkan, penulis melakukan analisis terkait *stopword* yang perlu ditambahnkan. Seperti kata "free", "gratis", "ori", "dijamin", "garansi", dan lain-lain.

D. Pembuatan Kelas & Pelabelan

Tahap selanjutnya adalah membuat kelas yang sesuai dengan merk yang dipilih. Kelas klasifikasi yang dibuat berasal dari nama produk, karena nama produk adalah nama yang memang diberikan/ditentukan oleh produsen *handphone*. Sehingga ketika produk sejenis dikelompokkan pada kelas yang sama, pembeli akan lebih mudah mencari produk yang memang sesuai dengan kemauannya. Contoh nama kelas yang dibuat adalah "iPhone", "iPhone 3G", "iPhone 3GS", "iPhone 4", "iPhone 4S", dan lain-lain.

Pada tahap ini, penulis juga melakukan pelabelan 300 data secara manual, dapat disebut dengan proses klasifikasi manual. Pada proses pelabelan, semua data produk akan diklasifikasi ke kelas yang sudah ditentukan pada tahap sebelumnya. Pelabelan disini digunakan untuk menentukan apakah hasil klasifikasi sistem sesuai dengan yang seharusnya.

E. Pengujian Sistem

Data produk yang dikumpulkan, akan digunakan untuk pengujian system. Untuk masing-masing merk, penulis mengambil 90 data produk dari gabungan tokopedia dan bukalapak, sehingga dikumpulkan 450 data produk. Dari 450 data produk tersebut, dibagi menjadi 3 kelompok data masing-masing 150, yaitu data latih, data uji 1, dan data uji 2. Penulis tidak membatasi secara pasti jumlah produk dari tokopedia maupun bukalapak.

Pengujian dibagi dalam 2 kali pengujian,

1. Pengujian 1 dilakukan menggunakan 150 data uji, Data uji terdiri dari masing-masing 30 data produk dari masing-masing merk. Data uji yang digunakan adalah datau uji 1. Pengujian ini akan menghasilkan nilai k dengan akurasi yang paling tinggi. Pengujian dilakukan dengan nilai k=1 sampai dengan 15.
2. Pengujian 2 dilakukan menggunakan 150 data uji yang terdiri dari campuran kelima merk yang dipilih (*multi-brand*). Data uji tersebut dibagi kedalam 5 set, masing-masing 30 data. Pengujian 2 dilakukan menggunakan nilai k yang diperoleh dari pengujian 1.

Proses klasifikasi dilakukan terhadap data judul dari sebuah produk. Hal ini dilakukan karena data judul sebuah produk sudah mencerminkan *keyword* sebuah produk.

F. Analisis Hasil Pengujian

Setelah penulis mendapatkan hasil klasifikasi produk yang sudah dilakukan oleh system. Pertama-tama analisis dilakukan dengan cara membandingkan hasil klasifikasi yang dilakukan oleh system dan klasifikasi yang dilakukan oleh penulis secara manual. Dari hasil perbandingan tersebut, akan dihitung akurasi untuk masing-masing pengujian, dalam hal ini masing-masing merk.

Setelah dihitung akurasi, penulis akan melihat secara detil satu-persatu data yang tidak sesuai hasil klasifikasi system dengan hasil klasifikasi secara manual. Kesalahan klasifikasi tersebut akan dianalisis penyebabnya, dan digunakan untuk masukan bagi penelitian selanjutnya.

V. Analisis & Pembahasan

Penulis melakukan pengujian terhadap data produk yang telah berhasil dikumpulkan sebelumnya. Data produk tersebut berasal dari tokopedia dan bukalapak. Data produk yang diuji berasal dari kategori *handphone* yang terdapat di kedua *e-marketplace*.

A. Pengujian 1

Pengujian pertama dilakukan menggunakan 150 data uji, dengan 5 kali proses klasifikasi untuk masing-masing merk yang dipilih. Hasil klasifikasi pengujian 1 dengan nilai k=1 dapat dilihat pada Tabel IV.

Tabel III
Hasil Klasifikasi pengujian 1 dengan k=1

No	Merk	Jumlah Data	Klasifikasi Benar	Klasifikasi Salah
1	Apple	30	28	2
2	Asus	30	29	1
3	Oppo	30	25	5
4	Vivo	30	16	14
5	Xiaomi	30	19	11
TOTAL		**150**	**117**	**33**

Hasil klasifikasi pada merk Apple, 28 dari 30 data produk berhasil diklasifikasi dengan benar oleh sistem. Sedangkan 2 data produk diklasifikasikan ke kelas yang tidak sesuai oleh sistem. Sedangkan pada merk Asus, 29 data produk berhasil diklasifikasikan, dan 1 data produk diklasifikasn ke kelas yang tidak sesuai. Pada merk Oppo, 25 data produk diklasifikasikan ke kelas yang sesuai, sedangkan 5 data produk diklasifikasikan ke kelas yang tidak sesuai. Pada merk Vivo, 16 data produk diklasifikasikan ke kelas yang sesuai, dan 14 data produk diklasifikasikan ke kelas yang tidak sesuai. Pada merk Xiaomi, 19 data produk berhasil diklasifikasikan ke kelas yang sesuai, sedangkan 11 data produk diklasifikaskan ke kelas yang tidak sesuai.

Tabel IV
Hasil Klasifikasi pengujian 1 dengan k=5

No	Merk	Jumlah Data	Klasifikasi Benar	Klasifikasi Salah
1	Apple	30	28	2
2	Asus	30	30	0
3	Oppo	30	30	0
4	Vivo	30	29	1
5	Xiaomi	30	29	1
TOTAL		**150**	**146**	**4**

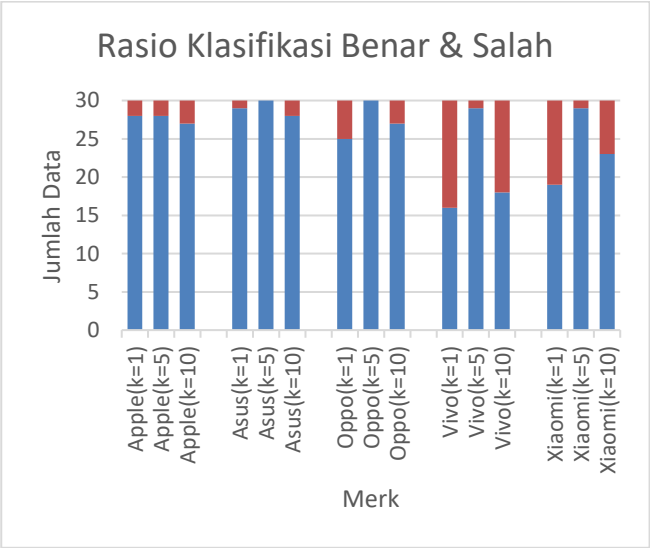
p-ISSN : 2443-2210
e-ISSN : 2443-2229

TABEL IV menunjukkan hasil klasifikasi pengujian 1 oleh sistem dengan k=5. Hasil klasifikasi pada merk Apple, 28 dari 30 data produk berhasil diklasifikasi dengan benar oleh sistem. Sedangkan 2 data produk diklasifikasikan ke kelas yang tidak sesuai oleh sistem. Sedangkan pada merk Asus dan Oppo, seluruh data berhasil diklasifikasikan dengan benar oleh sistem. Pada merk Vivo, 29 dari 30 data produk berhasil diklasifikasikan ke kelas yang benar oleh sistem. Pada merk Xiaomi, 29 dari 30 data produk berhasil diklasifikasikan ke kelas yang benar oleh sistem.

TABEL V
HASIL KLASIFIKASI PENGUJIAN 1 DENGAN K=10

No	Merk	Jumlah Data	Klasifikasi Benar	Klasifikasi Salah
1	Apple	30	28	2
2	Asus	30	29	1
3	Oppo	30	27	3
4	Vivo	30	26	4
5	Xiaomi	30	28	2
TOTAL		**150**	**138**	**12**

TABEL V menunjukkan hasil klasifikasi pengujian 1 dengan k=10. Hasil klasifikasi untuk merk Apple adalah 28 klasifikasi sesuai dan 2 klasifikasi tidak sesuai. Pada merk Asus, klasifikasi yang dihasilkan adalah 29 data produk diklasifikasikan ke kelas yang sesuai, dan 1 data produk diklasifikasikan ke kelas yang tidak sesuai. Sedangkan pada merk Oppo, 27 data produk diklasifikasikan ke kelas yang sesuai dan 3 data produk diklasifikasikan ke kelas yang tidak sesuai. Sedangkan pada merk Vivo, 26 data produk diklasifikasikan ke kelas yang sesuai, dan 4 data produk diklasifikasikan ke kelas yang tidak sesuai. Pada merk Xiaomi, 28 data produk diklasifikasikan ke kelas yang sesuai dan 2 data produk di kelas yang tidak sesuai.



Gambar 6. Rasio klasifikasi benar dan salah
Sumber: Hasil pengolahan data

Grafik rasio jumlah klasifikasi benar dan salah dapat dilihat pada Gambar 6. Berdasarkan data hasil klasifikasi oleh sistem, dihitung akurasi atau persentase keberhasilan sistem melakukan klasifikasi dengan benar.

TABEL VI
PERHITUNGAN AKURASI PENGUJIAN 1

	k=1	k=5	k=10
Klasifikasi benar	117	146	138
Klasifikasi salah	33	4	12
Akurasi	78.00%	97.33%	92.00%

TABEL VI menunjukan hasil perhitungan akurasi dari pengujian 1. Nilai akurasi terbaik adalah 97.33%, sedangkan nilai akurasi terburuk adalah 78%. Berdasarkan hasil perhitungan akurasi, nilai k terbaik adalah 5. Nilai k=5 akan digunakan untuk pengujian 2, klasifikasi *multi-brand*.

B. Pengujian 2

Pada pengujian kedua, klasifikasi dilakukan dengan 150 data uji kedua. Secara acak 150 data dibagi kedalam 5 set data, masing-masing 30 data produk yang berisi kelima merk yang dipilih secara acak. Nilai k dalam pengujian ini adalah 5. Hasil dari pengujian 2, dapat dilihat pada TABEL VII dibawah ini.

TABEL VII
HASIL KLASIFIKASI PENGUJIAN 2 DENGAN K=5

Set ke-	Jumlah Data	Klasifikasi Benar	Klasifikasi Salah
1	30	28	2
2	30	30	0
3	30	29	1
4	30	28	2
5	30	30	0
TOTAL	**150**	**145**	**5**

Berdasarkan hasil pengujian 2, data produk pada set pertama dan keempat menghasilkan 28 klasifikasi sesuai dan 2 klasifikasi salah. Sedangkan set kedua dan kelima menghasilkan 30 data diklasifikasikan dengan sesuai, sedangkan set ketiga menghasilkan 29 data diklasifikasikan dengan sesuai dan 1 data diklasifikasikan tidak sesuai. Secara keseluruhan, 145 data berhasil diklasifikasikan ke kelas yang sesuai, sedangkan 5 data diklasifikasikan ke kelas yang tidak sesuai. Berdasarkan hasil perhitungan, didapatkan nilai akurasi adalah sebesar 96.6%. Berikut ini perhitungan nilai akurasi dari pengujian 2 *multi-brand*:

$$Akurasi = \frac{145}{150} x100\% = 96.67\%$$

C. *Analisis hasil*

Setelah melakukan evaluasi sistem, penulis melakukan analisis untuk masing-masing data produk yang

diklasifikasikan ke kelas yang salah oleh sistem. Data produk hasil klasifikasi yang salah oleh sistem dapat dilihat pada TABEL VIII.

TABEL VIII
HASIL KLASIFIKASI OLEH SISTEM YANG SALAH

No	Nama Produk	Kelas Klasifikasi	Hasil Klasifikasi Sistem
1	Apple iphone 6s+ 16gb rosegold	Apple Iphone 6S Plus	Apple Iphone
2	iPhone 7 Plus, JET Black, 128GB Garansi Apple Internasional	Apple Iphone 7 Plus	Apple Iphone 7
3	hp vivo 5/ hp vivo second/ hp vivo murah/ hp second murah	Vivo V5	Vivo V3
4	[NEW] XIAOMI MI6 / MI 6 PRO RAM 6GB INTERNAL 128GB	Xiaomi Mi 6 Pro	Xiaomi Mi 6
5	iphone 7+ jet black	Apple Iphone 7 Plus	Apple Iphone 7
6	iPhone 6s+ Garansi Apple Internasional	Apple Iphone 6S Plus	Apple Iphone
7	VIVO V5 LITE RAM 3/32GB. PAKET GIFT BOX V5 NEW	Vivo V5 Lite	Vivo V5
8	vivo Y55S.	Vivo Y55S	Vivo Y55
9	HP OPPO F1s/ Oppo A57 3/32GB 4G LTE NEW [GOLD DAN BLACK]	Oppo F1s	Oppo A57

Terdapat 9 data yang diklasifikasikan ke kelas yang salah oleh sistem. 3 data produk berasal dari merk Apple, 3 data produk berasal dari merk Vivo, 1 data produk berasal dari merk Oppo, dan 1 data produk berasal dari merk Xiaomi.

Data produk pertama adalah "apple iphone 6s+ 16gb rosegold". Hasil klasifikasi oleh sistem adalah "Apple Iphone", sedangkan kelas yang seharusnya adalah "Apple Iphone 6S Plus". Kesalahan klasifikasi terjadi karena sistem tidak dapat memahami kesamaan symbol tambah (+) dan kata "*Plus*" pada judul produk. Pada penelitian ini, penulis menghilangkan karakter simbol pada tahap *preprocessing*. Oleh karena itu, sistem mengelompokkan data produk ke kelas yang salah. Kesalahan ini dapat diperbaiki dengan cara pengenalan symbol atau melengkapi data latih yang ada. Solusi pengenalan symbol dapat dilakukan dengan cara melakukan merubah simbol-simbol yang merupakan keyword fitur menjadi teks, seperti symbol "+" dirubah ke kata "plus". Kesalahan ini juga terjadi pada data produk kelima "iphone 7+ jet black" dan keenam "iPhone 6s+ Garansi Apple Internasional".

Data kedua adalah "Apple iPhone 7 Plus, JET Black, 128GB Garansi Apple Internasional". Pada data kedua, data produk seharusnya diklasifikasi ke kelas "Apple Iphone 7 Plus", tetapi sistem mengklasifikasikan ke kelas "Apple Iphone 7". Kesalahan terjadi karena pada kata "Plus," terdapat sebuah karakter koma yang terhubung ke kata "Plus". Langkah preprocessing yang dilakukan oleh penulis adalah melakukan tokenisasi, kemudian setelah dihasilkan token/kata, sistem menghapus token yang hanya berisi symbol. Di sini, sistem menganggap kata "Plus," (dengan koma) dan "Plus" adalah kata yang berbeda, sehingga pada proses *K-Nearest Neighbor* membentuk dimensi vektor yang berbeda. Kesalahan yang sama terjadi pada data kedelapan, "vivo Y55S.". Kesalahan ini dapat diperbaiki dengan menambah proses menghilangkan karakter symbol walau terhubung dengan kata dan/atau melakukan pengenalan simbol yang dapat mencerminkan nama produk.

Data ketiga adalah "hp vivo 5/hp vivo second/ hp vivo murah/ hp second murah". Data ketiga seharusnya diklasifikasikan ke kelas "Vivo V5", tetapi sistem mengklasifikasi ke kelas Vivo V3. Kesalahan terjadi karena judul produk tidak tertulis dengan benar. Penjual hanya menuliskan "vivo 5" bukan "Vivo V5", sehingga sistem melakukan kesalahan klasifikasi. Hal ini juga berhubungan dengan data training yang ada di sistem, dimana sistem tidak memiliki data latih yang berisi "vivo 5" pada kelas "Vivo V5". Apabila diperiksa pada deskripsi, penjual ada menuliskan data produk "Vivo V5", tetapi pada penelitian ini, penulis tidak melakukan klasifikasi terhadap deskripsi produk. Untuk pengembangan selanjutnya, penulis menyarankan klasifikasi juga dilakukan ke deskripsi produk.

Data keempat adalah "[NEW] XIAOMI MI6 / MI 6 PRO RAM 6GB INTERNAL 128GB". Pada kasus keempat, hasil klasifikasi oleh sistem adalah "Xiaomi Mi 6", sedangkan seharusnya data produk masuk ke kelas "Xiaomi Mi 6 Pro". Kesalahan klasifikasi terjadi karena data latih yang terdapat di judul produk berasal dari 2 kelas, "Xiaomi Mi 6" dan "Xiaomi Mi 6 Pro". Apabila dilihat secara keyword, 2 keyword berasal dari "Xiaomi Mi 6", dan 1 keyword berasal dari "Xiaomi Mi 6 Pro". Sehingga apabila dilakukan klasifikasi menggunakan *K-Nearest Neighbor*, dimensi vector yang terbentuk oleh judul produk akan cenderung dekat dengan "Xiaomi Mi 6", sehingga sistem akan mengelompokkannya ke kelas "Xiaomi Mi 6". Solusi untuk permasalahan ini adalah dengan melakukan klasifikasi ke deskripsi produk.

Data ketujuh "VIVO V5 LITE RAM 3/32GB. PAKET GIFT BOX V5 NEW" dan kesembilan "HP OPPO F1s/ Oppo A57 3/32GB 4G LTE NEW [GOLD DAN BLACK]" diklasifikasikan ke kelas yang salah. Kesalahan klasifikasi terjadi karena pada judul terdapat lebih dari 1 nama produk. Akan tetapi apabila dilihat kedalam teks deskripsi, data

produk hanya mendeskripsikan salah satu produk saja. Kesalahan yang terjadi ini adalah kesalahan pengguna, dimana pengguna menuliskan lebih dari 1 produk untuk judul produk yang dijual di *e-marketplace*. Solusi yang dapat dilakukan untuk mengatasi permasalahan ini adalah melakukan klasifikasi ke deskripsi produk.

VI. KESIMPULAN & SARAN

Berdasarkan hasil pengujian, metode *K-Nearest Neighbor* dapat melakukan klasifikasi produk dari *e-marketplace*, khususnya tokopedia dan bukalapak. Akurasi yang dihasilkan dari pengujian 1, pemilihan nilai k=1, 5, atau 10 adalah 78%, 97,33% dan 92%. Berdasarkan pengujian 1 disimpulkan nilai k yang optimal untuk kasus ini adalah 5. Pada pengujian 2, *multi-brand*, akurasi yang dihasilkan adalah 96.67%. Nilai akurasi ini dapat dikatakan baik karena melebihi 90%. Akurasi dari algoritma *K-Nearest Neighbor* sangat dipengaruhi oleh data latih. Semakin lengkap data latih, maka akurasi akan semakin baik.

Pada penelitian ini, satu kesalahan klasifikasi terjadi karena penjual/pengguna *e-marketplace* tidak memberikan nama produk atau judul dengan benar. Sehingga algoritma *K-Nearest Neighbor* melakukan kesalahan klasifikasi.

Beberapa saran pengembangan yang dapat dilakukan untuk penelitian selanjutnya adalah

1. Menambah proses normalisasi kata, terutama untuk kata dalam Bahasa asing, kata-kata yang disingkat, atau kata-kata tidak resmi (*slang*), dan simbol.
2. Lakukan pengujian ke data konten deskripsi produk. Karena dimungkinkan melihat sebuah produk *handphone* melalui spesifikasi yang dituliskan pada deskripsi sebuah produk.
3. Lakukan pengujian ke data produk selain kategori *handphone* atau pengujian ke data produk yang berasal dari *e-marketplace* selain tokopedia dan bukalapak.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Fakultas Teknologi Infomasi Universitas Kristen Duta Wacana yang telah mendukung kegiatan penelitian ini sehingga dapat terlaksana dengan baik.

DAFTAR PUSTAKA

[1] iprice, "Peta E-Commerce Indonesia - Q2 2018," iprice, 2018. [Online]. Available: https://iprice.co.id/insights/mapofecommerce/. [Accessed 2018 Oktober 2018].

[2] wearesocial, "Digital in 2018 in Southeast Asia Part 2 - South-East," wearesocial, 29 Jan 2018. [Online]. Available: https://www.slideshare.net/wearesocial/digital-in-2018-in-southeast-asia-part-2-southeast-86866464. [Accessed 1 Oktober 2018].

[3] 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.

[4] D. Adeniyi, Z. Wei and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Applied Computing and Informatics,* vol. 12, no. 1, pp. 90-108, 2016.

[5] 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.

[6] W. E. Nurjanah, R. S. Perdana and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter Menggunakan Metode K-Nearest Neighbor dan Pembobotan Jumlah Retweet," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer,* vol. 1, no. 12, 2017.

[7] P. D. Nugraha, S. A. Faraby and Adiwijaya, "Klasifikasi Dokumen Menggunakan Metode k-Nearest Neighbor (kNN) dengan Information Gain," *eProceedings of Engineering,* vol. 5, no. 1, 2018.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] E. Han, G. Karypis and V. Kumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," Minneapolis, 1999.

[22] B. Sriram, D. Fuhry, E. Demir, H. Ferhastosmanoglu and M. Demirbas, "Short Text Classification in Twitter to Improve Information Fitltering," in *SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development*

in information retrieval, Geneva, 2010.

[23] 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.

[24] 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.

[25] 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.

[26] E. W. Jumadi, "Penggunaan KNN (K-Nearest Neighbor) Untuk Klasifikasi Teks Berita yang Tak-Terkelompokkan pada Saat Pengklasteran Oleh STC (Suffix Tree Clustering)," *ISTEK*, vol. IX, pp. 50-81, Juni 2015.

[27] Kusrini and Luthfi, Algoritma Data Mining, Yogyakarta: Andi Offset, 2009.

[28] Y. Liao, "Review of K-Nearest Neigbor Text Categorization Method," 2002. [Online]. Available: https://www.usenix.org/legacy/event/sec02/full_papers/liao /liao_html/node4.html. [Accessed 10 March 2017].

[29] A. Ardiyanto, "Klasifikasi Komentar pada Dataset Pemilu Presiden Indonesia 2014 dengan Metode Improved K-Nearest Neighbor," Yogyakarta, 2017.

[30] C. Manning and H. Schutze, Foundation of Statistical Natural Language Processing, Cambridge: MIT Press, 2000.

[31] N. Suguna and K. Tanushkodi, "An Improved k-Nearest Neighbor Classification Using Genetic Algorithm," *International Journal of Comupter Science Issue,* vol. 7, no. 2, pp. 18-21, 2010.

[32] F. Tala, "A study of stemming effects on information retrieval in Bahasa Indonesia.," Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands, 2003.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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].

[39] 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].

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] P. Lopes and B. Roy, "Dynamic Recommendation System Using Web Usage Mining for E-commerce Users," *Procedia Computer Science,* vol. 45, pp. 60-69, 2015.

[51] V. Gupta, H. Karnick, A. Bansal and P. Jhala, "Product classification in e-commerce using distributional semantics," *arXiv preprint arXiv:1606.06083*, 2016.