Text Processing Clustering dalam Menentuan Profesi Berdasarkan Data Twitter

Elisabet Sihite[1], Rima Dias Ramadhani[2], Muhammad Zidny Naf'an[3], Rifki Adhitama[4]
[1,2,3]Program Studi Informatika, [4]Program Studi Rekayasa Perangkat Lunak,
[1,2,3,4] Fakultas Teknik Industri dan Informtika, Institut Teknologi Telkom Purwokerto
[1,2,3,4] Purwokerto Kulon, Jawa Tengah, Indonesia
[1]14102059@ittelkom-pwt.ac.id, [2]rima@ittelkom-pwt.ac.id, [3]zidny@ittelkom-pwt.ac.id,
[4]rifki@ittelkom-pwt.ac.id

Abstrak – Saat ini akibat jumlah data yang semakin besar, twitter tidak hanya digunakan untuk menulis pesan microblogging melainkan juga untuk melakukan penemuan pengetahuan untuk menyelidiki fenomena yang terjadi di kehidupan masyarakat. Salah satu fenomena yang terjadi di kehidupan sosial masyarakat adalah pemilihan pekerjaan yang tidak sesuai kemampuan dan keahlian. Dampak buruk dapat terjadi pada perusahaan dan pegawai yang tidak memilih pekerjaan yang sesuai dengan keahlian dan kemampuan yang dimilikinya seperti kerugian besar pada perusahaan dan timbulnya tingkat stress yang tinggi yang tidak menutup kemungkinan timbulnya penyakit yang membahayakan pada pegawai yang tidak memilih pekerjaan sesuai dengan kemampuan dan keahliannya. Terdapat beberapa solusi yang ditawarkan untuk meminimalisir dampak negatif pada pegawai dan perusahaan. Salah satu teknik yang digunakan untuk menyelidiki fenomena yang terjadi dengan menggunakan metode clustering. Pada tahap pengelompokan (clustering) perlu dilakukan text processing hal ini dikarenakan tahapan preprocessing menjadi salah satu penentu untuk menghasilkan nilai keakuratan yang maksimal dan terbaik untuk suatu pengujian metode clustering. Melalui text processing yang telah dilakukan diketahui jika text processing harus dilakukan secara maksimal untuk dapat mencapai nilai keakuratan yang baik pada perancangan dan pengujian metode clustering. Tahapan text processing yang dilakukan dalam hal ini adalah Cleansing, Normalisasi, Transform Case, Stemming, Stop Word dan Tokenize.

Kata kunci – Twitter, TextProcessing, Clustering, Algoritma Nazief & Adriani dan TF-IDF.

Abstract— Currently the data amount of data is getting bigger, twitter is not only to write microblogging messages and also to know what is going on in the community. One of the phenomena that occur in social life is a job that is not in accordance with ability and expertise. Adverse impacts can occur on companies and employees who do not like work that matches the work done such as large losses to the company and the incidence of high stress levels that cannot cover costs arising from difficulties that are not in accordance with ability and expertise. . There are several solutions that are used to minimize negatively on employees and companies. One of the techniques used to use the clustering method. In the clustering stage it is necessary to process this text with the preprocessing stage being one of the determinants to produce the right data. Through text processing that has been done, if the text processing must be done to be able to achieve the right value. Stages of text processing done in this case is Cleansing, Normalization, Transform Case, Stemming, Stop Word and Tokenize.

Keywords Twitter; Text Processing;, Clustering; Algorithm Nazief & Adriani and TF-ID.

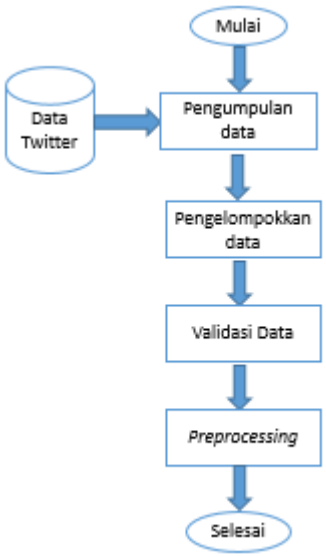
I. PENDAHULUAN

Jejaring sosial twitter telah menjadi media komunikasi digital populer yang digunakan oleh ratusan juta orang dalam beberapa tahun terakhir [1]. Menurut [2] Facebook digunakan untuk menjadwalkan protes, Twitter untuk berkoordinasi. Twitter merupakan salah satu jejaring sosial yang paling banyak digunakan karena bersifat real-time. Twitter dapat dianggap sebagai microblog dikarenakan twitter hanya mampu untuk menampung beberapa pesan teks yang jumlahnya terbatas [3]. Twitter menjadi sebuah sumber data yang paling banyak menyimpan berbagai jenis data [3]. Saat ini akibat jumlah data yang semakin besar, twitter tidak hanya digunakan untuk menulis pesan *microblogging* melainkan juga untuk melakukan penemuan pengetahuan untuk menyelidiki fenomena yang terjadi di kehidupan masyarakat [1]. Salah satu fenomena yang terjadi di kehidupan sosial masyarakat adalah pemilihan pekerjaan yang tidak sesuai kemampuan dan keahlian [4] Dari hasil penelitian yang dilakukan oleh Lee, dkk diketahui bahwa kebanyakan masyarakat tidak memilih

pekerjaan yang sesuai dengan keahlian dan kemampuan yang dimilikinya [5]. Hal ini akan berdampak buruk bagi perusahaan maupun pegawainya. Perusahaan yang memperkerjakan masyarakat yang tidak sesuai bidang kemampuan yang dimilikinya akan merasa dirugikan nantinya dan mungkin akan memberhentikan karyawan maupun pegawainya [6]. Disisi lain, dampak yang akan terjadi pada pegawai yang memilih serta mengerjakan sesuatu yang tidak sesuai dengan kemampuan dan keahliannya akan menyebabkan timbulnya tingkat stress yang tinggi yang tidak menutup kemungkinan timbulnya penyakit yang membahayakan [7]. Berdasarkan latar belakang tersebut, studi tentang penentuan minat dalam pekerjaan diperlukan untuk meminimalisir akibat terjadi dikemudian hari. Penelitian terkait telah banyak dilakukan untuk meminimalisir dampak negatif pada pegawai dan perusahaan. Terdapat beberapa solusi yang ditawarkan diantaranya adalah konsultasi psikologi [8], psikotes [9], pembimbingan dini [10] dan pemanfaatan sistem cerdas. Pemanfaatan sistem cerdas telah banyak dilakukan untuk membantu kasus ini misalkan penelitian yang dilakukan oleh Armen S. Tahirsylaj [11] tentang sistem pendukung keputusan menggunakan teknik text mining [12]. Salah satu teknik digunakan untuk menyelidiki fenomena yang terjadi adalah metode clustering [13].

II. METODE PENELITIAN

Pada penelitian ini dilakukan metode penelitian untuk melakukan *text processing* untuk melakukan *clustering* dalam menentukan profesi berdasarkan data twitter dapat diketahui pada Gambar 2.1 berikut ini.



Gambar 2.1 Metode Penelitian

A. Pengumpulan Data

Pengumpulan data dilakukan dengan dengan mengumpulkan data yang memiliki karakteristik/kemiripan antara kata yang satu dengan kata yang lain.

B. Pengelompokan Data

Pengelompokan data dilakukan dengan melihat tingkat kemiripan teks dengan teks yang lainnya dengan menyusun teks dalam bentuk kuisioner. Hasil penilai kuisioner berdasarkan jumlah tanggapan 45 responden yang menentukan kelompok data pada 4 kelompok *cluster* yang telah ditentukan.

C. Validasi Data

Validasi Data dilakukan dengan menghitung nilai keakuratan pada data dengan menggunakan perhitungan reliability pada software SPSS. Hasil perhitungan reliability menggunakan software SPSS menunjukkan jika nilai yang dihasilkan 0,966 yang sangat memuaskan sehingga diketahui jika nilai keakuratan data sangat memuaskan karena sudah mendekati nilai 1 yang menunjukkan kesempurnaan.

Tabel 2.1 *Reliability Statistics*

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.966	.966	45

Hasil perhitungan fungsi pendukung reliability pada software SPSS dapat diketahui seperti pada Gambar 2.2 berikut ini.

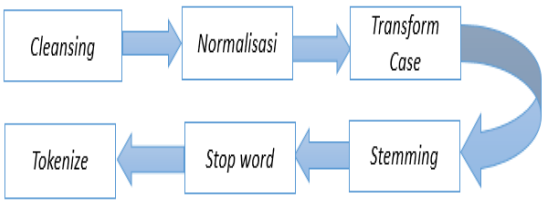
Tabel 2.2 *Scale Statistics*

Mean	Variance	Std. Deviation	N of Items
1,2383E2	988.878	31.44643	45

Tabel 2.2 menunjukan jika nilai *mean* pada validasi data sebesar 1,2383E2 dengan nilai *variance* sebesar 988.878 dan nilai *std deviation* dengan jumlah responden sebanyak 45 responden.

D. Prepocessing

Tahapan *preprocessing* merupakan tahapan awal untuk melakukan pengolahan data. Beberapa tahapan *preprocessing* dapat diketahui pada Gambar 2.2 diantaranya yaitu: *Cleansing*, Normalisasi, *Transform case*, *Stemming*, *Stop word* dan *Tokenizing*.



Gambar 2.2 Tahapan *preprocessing*

a) Cleansing

Cleansing merupakan suatu teknik yang dilakukan untuk menghilangkan kata-kata yang tidak dperlukan seperti kata URL, hastag (#), username (@), angka besera simbol dan tanda baca pada data twitter untuk dilakukan pengolahan pengolahan data selanjutnya. [14]

b) Normalisasi

Normalisasi merupakan sebuah proses pengolahan data dengan menggunakan kamus untuk mengubah setiap kata yang tidak sesuai dengan arti dan makna yang sebenarnya. Pengunaan normalisasi dilakukan dengan menggunakan kamus manual yang akan menunjukkan penerjemahan arti dan makna dari sebuah kata slank "gaul", singkatan (akronim) menjadi kata yang memiliki makna sesuai pada kamus besar bahasa Indonesia. Kamus Normalisasi yang dilakukan secara manual sebagai berikut :


```
{
  "_id": "48",
  "_takBaku": "bhs",
  "_baku": "bahasa"
},
{
  "_id": "49",
  "_takBaku": "dg",
  "_baku": "dengan"
},
{
  "_id": "50",
  "_takBaku": "dgn",
  "_baku": "dengan"
   },
```


c) Transform Case

Transform case adalah sebuah proses pengolahan data dengan melakukan perubahan setiap huruf besar *(uppercase)* menjadi huruf kecil (*lowercase*) sehingga menjadi satu jenis huruf yaitu huruf kecil seluruhya.

d) Stemming

Stemming merupakan proses pengolahan data selanjutnya yang dilakukan untuk mengubah kata-kata bahasa Indonesia menjadi kata dasar yang tidak memiliki imbuhan awalan (prefiks), sisipan (infiks), akhiran (sufiks) dan gabungan awalan akhiran (konfiks). *Stemming* yang digunakan menggunakan algoritma Nazief & Adriani. Penggunaan Algoritma Nazief & Adriani dikarenakan algoritma ini diketahui memiliki tingkat akurasi yang sangat tinggi yaitu 95,26% dibandingkan dengan algoritma Porter yang hanya memiliki nilai akurasi sebesar 79,13%. [15]

e) Stop word

Stop word merupakan fungsi pengolahan data yang menghapus kata-kata penghubung yang tidak berhunungan dengan klasifikasi yang mengurangi dimensi teks namun tidak mengurangi isi sentimen dari teks.

f) Tokenize

Tokenize merupakan proses preprocessing yang terakhir yang dilakukan untuk mengubah kalimat menjadi satu-satu kata secara terpisah sehingga nantinya dapat dihitung nilai frekuensi setiap kata yang digunakan. Tahapan preprocessing ini dilakukan dengan menggunakan ekstraksi fitur TF-IDF sebagai cara untuk menghitung nilai frekuensi setiap kata yang dilakukan dengan menggunakan *software* RapidMinner. Tokenize menunjukan jika nilai frekuensi pada masing-masing data diketahui dengan menggunakan rumus sebagai berikut untuk mengetahui nilai frekuensi kemunculan term. Untuk mengetahui nilai TF, IDF dan TF-IDF dapat diketahui berdasarkan persamaan dibawah ini.

Nilai *term frequency (TF)*

$$W(d, t) = TF(d, t) \quad (1)$$

Nilai *Inverse Document Frequency* (IDF)

$$IDF = LOG(N/(DF(t))) \quad (2)$$

Nilai TF-IDF dilakukan berdasarkan pengabungan nilai metode TF dan metode IDF dengan persamaan sebagai berikut [15].

$$TF\text{-}IDF(d,t) = TF(d, t).\ IDF(t) \quad (3)$$

III. HASIL PENELITIAN

Hasil penelitian yang dihasilkan oleh setiap tahapan *preprocessing* adalah sebagai berikut.

A. Hasil *preprocessing cleasing*

Tabel 3.1 *Preprocessing Cleansing*

No	Data sebelum *Cleansing*	Data sesudah *Cleansing*
1.	Ini hanya lah sebuah awal dari karir saya nanti bukan untuk menyombongkan hati tapi untuk memotivasi diri sendiri:	Ini hanya lah sebuah awal dari karir saya nanti bukan untuk menyombongkan hati tapi untuk memotivasi diri sendiri

2.	@merryriana: pada diriku, pendorong yang kuat itu adalah untuk membahagiakan orangtua. Hal 234	merryriana diri dorong kuat bahagia orangtua

B. Hasil *preprocessing* normalisasi

Tabel 3.2 *Preprocessing* normalisasi

No	Data sebelum normalisasi	Kamus Manual	Data sesudah normalisasi
1.	dewasa itu ketika bisa taat dan memegang teguh komitmen	memegang = pegang	dewasa itu ketika bisa taat dan pegang teguh komitmen
2.	hilangkan perut buncit dalam 5 detik	hilangkan = hilang	hilang perut buncit dalam detik
3.	mendadak hari ini kantor kedatangan engineer baru	mendadak = dadak engginee r = insinyur	dadak hari ini kantor datang insinyur baru
4.	namun bagi seniman rekaman yg hanya berkarya di studio dengan mengeluarkan investasi yg besar, music streaming tidak terlalu menjanjikan.	rekaman = rekam berkarya = karya mengeluarkan = keluar music = musik stereaming = tonton menjanjikan = janji	namun bagi seniman rekam yang hanya karya di studio dengan keluar investasi yang besar musik tonton tidak terlalu janji

C. Hasil *preprocessing tranform case*

Tabel 3.3 *Preprocessing tranform case*

No	Kelompok Data	Data sebelum *Transform case*	Data sesudah *Transform case*
1.	Pengusaha	Saya	saya
2.	Pendidik	Rindu	rindu
3.	Karyawan Perkantoran	Rapat	rapat
4.	Seniman	Namun	namun

Tabel 2.3 menunjukan perubahan jenis huruf *capital (uppercase)* menjadi huruf kecil (*lowercase)* pada masing-masing teks pada setiap kelompok yang telah ditentukan.

D. Hasil *preprocessing stemming*

Tabel 3.4 *Preprocessing Stemming*

No	Kelompok Data	Data sebelum *Stemming*	Data sesudah *Stemming*
1.	Pengusaha	pengaduan	adu
2.	Pendidik	memiliki	milik
3.	Karyawan Perkantoran	pemimpin	pimpin
4.	Seniman	perayaan	raya

E. Hasil *preprocessing stop word*

Tabel 3.5 *Preprocessing Stop word*

No	Kelompok Data	Data sebelum *Stop word*	Data sesudah *Stop word*
1.	Pengusaha	Yang belum menang ikutan lagi ya besok masih ada dua hari lagi semangat semoga beruntung	menang ikut besok semangat moga untung
2.	Pendidik	taat konstitusi bukan konstituen	taat konstitusi konstituen

3.	Karyawan Perkantoran	jadi teman mari lupakan pekerjaan nikmati liburan	teman mari lupa kerja nikmat libur
4.	Seniman	saya percaya air mata adalah sebuah ekspresi ungkapan hati yg tulus kini air mata kebahagiaan itu jatuh	percaya air ekspresi ungkap hati yg tulus air bahagia jatuh

F. Hasil *preprocessing tokenize*

Tabel 3.6 *Preprocessing Tokenize*

informasi	insinyur	kantor	musik	pasar	rekam	rupiah	seleksi
0	0	0	0	0.707	0	0.707	0
0.707	0	0	0	0	0	0	0.707
0	0.707	0.707	0	0	0	0	0
0	0	0	0.707	0	0.707	0	0

IV. PEMBAHASAN

Tahapan *Preprocessing* atau yang biasa disebut juga Pra-proses merupakan tahap awal pemerosesan pengolahan data. Tahap ini dilakukan untuk mempersiapkan data sehingga siap untuk digunakan pada tahap selanjutnya yaitu tahap pengembangan model deteksi untuk mengelompokkan data dengan mneggunakan metode *clustering*. Algoritma yang digunakan pada tahap *preprocessing stemming* yaitu Algoritma Nazief & Adriani. Algoritma ini memiliki kelebihan dan kekurangan pada saat melakukan pengubahan kata yang memiliki imbuhan menjadi kata dasar. Berdasarkan *study literature* diketahui jika pengolahan data dengan menggunakan algoritma stemming Nazief & Adriani pada bahasa Indonesia memiliki nilai akurasi yang sangat besar dengan nilai akurasi sebesar 95,26% dibandingkan nilai akurasi yang dihasilkan algoritma steming porter pada bahasa Indonesia yaitu bernilai sebesar 79,13% hal ini dikarenakan adanya pembacaan kamus pada Algoritma Nazief & Adriani. Algoritma Nazief & Adriani menggunakan kamus untuk dapat melakukan perubahan kata yang memiliki imbuhan menjadi sebuah kata dasar. Setiap tahapan pada pengolahan data menggunakan *text processing* akan menunjukkan keterkaitan antara tahapan yang satu dengan yang lainnya (penyempurnaa) dalam pengolahan data dimana tahapan-tahapan yang dimiliki pada *text processing* digunakan untuk membuat kata menjadi sesuai karakteristik data yang sesuai dengan kebutuhkan pengelompokan data (*clustering).*

V. PENUTUP

A. Kesimpulan

Berikut beberapa kesimpulan yang diketahui yaitu :

1) Tahapan *preprocessing* yang dilakukan akan mempengaruhi nilai akurasi pada pengujian metode nantinya hal ini disebabkan oleh karena data yang akan digunakan untuk mekakukan pengujian metode semakin baik semakin sesuai karektersitik pengelompokan data maka nilai akurasi akan semakin meningkat namun begitu juga sebaliknya.
2) Penggunaan kamus pada algoritma *stemming* Nazief & Adriani sangat mempengaruhi nilai akurasi, semakin akurat kamus yang digunakan maka akan semakin akurat juga nilai akurasi algoritma Nazief & Adriani.
3) Melaui *study literature* diketahui jika algoritma Nazief & Adriani memiliki tingkat akurasi yang sangat tinggi sebesar 95,26 % dibandingkan dengan penggunaan algoritma porter pada bahasa indonesia yang memiliki nilai sebesar 79,13% namun untuk waktu pengeksekusian program *stemming* diketahui jika algoritma porter memiliki waktu pengeksekusian program yang lebih baik karena tidak membutuhkan waktu yang lama sehingga sangat efisien untuk digunakan pada *stemming* bahasa Indonesia.

B. Saran

Melalui penelitian ini disarankan untuk menggunakan tahap *preprocessing* secara maksimal untuk dapat meningkatkan nilai akurasi pada penggunaan metode *cluster* nantinya.

DAFTAR PUSTAKA

[1] C. Chen et al., “Investigating the deceptive information in Twitter spam,” Futur. Gener. Comput. Syst., vol. 72, pp. 319–326, 2017.

[2] M. N. M. Klinczak dan C. A. A. Kaestner, “A study on topics identification on Twitter using clustering algorithms,” 2015 Latin-America Congr. Comput. Intell. LA-CCI 2015, 2016.

[3] M. Choi,, Y. Sang,, dan H. Park.”Exploring political discussions by Korean Twitter users: A look at opinion leadership dan homophily phenomenon,”

[4] J. PAN, X. QIN, dan G. G. LIU, “The impact of body size on urban employment: Evidence from China,” China Econ. Rev., vol. 27, pp. 249–263, 2013.

[5] N. Y. Lee, Y. Kim, dan Y. Sang, “How do journalists leverage Twitter? Expressive dan consumptive use of Twitter,” Soc. Sci. J., 2015.

[6] Z. Megri, “The Impact of Talent Management System on the Enterprise Performance: a Study on a Sample of Workers in National Company of Juice dan Canned-food Unit MANAA (Batna),” Arab Econ. Bus. J., vol. 9, no. 2, pp. 156–165, 2014.

[7] Y. Chen, B. McCabe, dan D. Hyatt, "Impact of individual resilience dan safety climate on safety

[8] L. Lazuras dan A. Dokou, "Mental health professionals' acceptance of online counseling," Technol. Soc., vol. 44, pp. 10–14, 2016.

[9] Q. Wang, "Coaching for Learning: Exploring Coaching Psychology in Enquiry-based Learning dan Development of Learning Power in Secondary Education," Procedia - Soc. Behav. Sci., vol. 69, no. Iceepsy, pp. 177–186, 2012.

[10] F. Gorunescu, Data Mining : Concepts, Model dan Techniques, New York: Springer-Verlag, 2011.

[11] A. S. Tahirsylaj, "Stimulating creativity dan innovation through Intelligent Fast Failure," Think. Ski. Creat., vol. 7, no. 3, pp. 265–270, 2012.

[12] K. Singh, H. K. Shakya, dan B. Biswas, "Clustering of people in social network based on textual similarity," Perspect. Sci., vol. 8, pp. 570–573, 2016.

[13] F. Goronescu, Data Mining : Concepts, Model dan Techniques, New York : Springer-Verlag, 2011.

[14] D. J. Suri and K. K. Purnamasari, "Perbandingan Seleksi Fitur Untuk Klasifikasi Sentimen SVM Pada Twitter."

[15] L. Agusta, U. Kristen, and S. Wacana, "PERBANDINGAN ALGORITMA STEMMING PORTER DENGAN ALGORITMA NAZIEF & ADRIANI UNTUK STEMMING DOKUMEN TEKS BAHASA INDONESIA," pp. 196–201, 2009.