PENGARUH *STEMMING* TERHADAP EKSTRAKSI TOPIK MENGGUNAKAN METODE *TF*IDF*DF* PADA APLIKASI *PDS*

Luthfan Hadi Pramono[1)], dan Cuk Subiantoro[2)]
[1, 2)]Teknik Informatika STMIK AKAKOM
[1, 2)]Jl. Raya Janti Karang Jambe No. 143 Yogyakarta 55198, Indonesia
e-mail: luthfanhp@akakom.ac.id[1)], cuks@akakom.ac.id[2)]

ABSTRAK

Personal Digital Secretary (PDS) merupakan sistem yang dikembangkan untuk menjadi "sekretaris pribadi" yang bekerja mendampingi pengguna secara digital. PDS menyampaikan informasi kepada penggunanya berupa email, media sosial dan berita. Untuk mengetahui informasi berita dari luar, harus dilakukan penelusuran topik pengguna melalui data email dan media sosial, sehingga informasi berita akan memiliki hubungan keterkaitan dengan pengguna. Penelusuran topik pengguna pada data email dan media sosial pada PDS menggunakan metode modifikasi pembobotan dalam algoritma TF*IDF yang dinamakan TF*IDF*DF. Pada pengembangan lebih lanjut, ditambahkan proses stemming dengan harapan untuk mendapatkan topik yang sesuai. Dari penelitian yang telah dikerjakan, terdapat perbedaan terms yang didapatkan dari ekstraksi topik tanpa penambahan proses stemming dan dengan penambahan proses stemming. Informasi berita yang didapatkan dengan penambahan proses stemming memiliki hasil lebih terfokus dibandingkan dengan informasi berita yang didapatkan dari ekstraksi topik tanpa penambahan proses stemming. Dengan penambahan proses stemming pada algoritma TF*IDF*DF menunjukkan bahwa kata (term) hasil dari proses ekstraksi yang didapatkan telah menjadi kata dasar karena adanya proses stemming. Kata dasar disini merupakan bentuk dasar yang merupakan indikasi sebuah topik.

***Kata Kunci**: Topik pengguna, ekstraksi topik, TF*IDF, model topik, seleksi fitur.*

ABSTRACT

Personal Digital Secretary (PDS) is a system that was developed to be a "personal secretary" who work alongside users digitally. PDS convey information to users in the form of email, social media and news. In order to know the information and news from the outside, it must be done by extracting user topics through email and social media, with the result that news information will have corresponding relationships with users. User topic extraction through email and social media in PDS is using modified weighting method in TF*IDF algorithm named TF*IDF*DF. In the further development, added stemming process in hopes of obtaining an appropriate topic. From the research that has been done, there are differences in terms obtained from the topic extraction without addition stemming process and with addition of stemming process. News information obtained by the addition of stemming process has more focused results than the news information obtained from the topics extraction without additional stemming process. With the addition of stemming process on the TF*IDF*DF algorithm indicates that the word (terms) results obtained from the extraction process has become the basic words because of stemming process. These Basic words are the basic form that an indication of a topic

***Keywords**: User topic, topic extraction, TF*IDF, topic model, fiture selection.*

I. PENDAHULUAN

Personal Digital Secretary (*PDS*) merupakan pengembangan dari sistem yang pernah dikembangkan sebelumnya [1]. *PDS* adalah sistem yang dikembangkan untuk menggantikan peran sekretaris dalam kegiatan sehari-hari. Sistem ini diharapkan dapat bekerja secara penuh selama 24 jam setiap hari dan dapat menggantikan peran sekretaris yang tidak bisa bekerja secara penuh karena memiliki batasan waktu dalam bekerja. *PDS* diharapkan akan bisa mendampingi pengguna, baik di rumah, di kantor maupun aktifitas di luar. *PDS* ditujukan untuk pengguna yang aktif sehingga harus bisa menangani informasi dari luar seperti berita, media sosial, dan email. Informasi-informasi tersebut akan disampaikan ke pengguna secara tepat seperti halnya seorang sekretaris.

Data informasi email bisa digunakan untuk klasifikasi topik. Fitur pada email diekstrak dari *email content* atau *body*, judul atau subyek, ataupun *metadata* yang lain misalnya pengirim, penerima, bcc, tanggal pengiriman, tanggal penerimaan, jumlah penerima dan yang lainnya [2]. Twiter, merupakan media sosial dan sebuah layanan *microblogging* yang populer. Dari informasi tersebut, twitter telah banyak dianalisa dan salah satunya digunakan sebagai pencarian topik minat dari penggunanya [3]. Ekstraksi topik pengguna pada *PDS* menggunakan data informasi awal yang dimiliki pengguna yaitu email dan media sosial twitter. Dokumen dari email dan media sosil twitter diproses melalu *text preprocessing* kemudian diberikan pembobotan dan dirangking sehingga didapatkan kata-kata tertentu yang mewakili topik pembahasan dari pengguna.

Penyampaian informasi yang tepat kepada pengguna melibatkan pengolahan data informasi serta melalui tahapan-tahapan pengelompokan. Ada tiga macam sumber data informasi yang diolah oleh sistem *PDS* yaitu email, media sosial serta berita. Dalam hal ini email dan media sosial terkait dengan pengguna karena masing-masing memiliki akun sendiri, sehingga informasi yang masuk melalui email dan media sosial akan masuk ke pengguna. Sedangkan untuk informasi berita, informasi ini bersifat umum dan tidak ada keterhubungan terhadap pengguna. Informasi berita ini harus dikaitkan dengan pengguna dengan menggunakan suatu cara tertentu yaitu penelusuran atau ekstraksi topik pengguna terhadap informasi [1].

Dari penelitian yang telah dikerjakan masih terdapat *noisy text* yang muncul pada ekstraksi topik yang mengakibatkan munculnya *term* yang tidak sesuai dengan tata penulisan. Oleh karena itu, perlu dilakukan proses lebih lanjut untuk memperbaiki *noisy text* sehingga *term* yang didapatkan akan lebih sempurna.

Berdasarkan penelitian yang telah dikerjakan oleh [4], teknik pemodelan ekstraksi topik dibagi menjadi 3 kategori yaitu pendekatan *information retrieval*, pendekatan *mechine learning* dan pendekatan *ontology based*.

Dalam pendekatan *information retrieval*, dokumen dan minat pengguna direpresentasikan dalam bentuk vektor pembobotan *term*. Efektivitas pendekatan ini dapat lebih ditingkatkan dengan memperkenalkan vektor topik. Metode ini banyak kekurangan, dikarenakan masalah polisemi dan sinonim dan juga karena dari isi halaman web ikut terkait seperti navigasi link (home, page, contact dan lainnya). Beberapa penelitian yang mengunakan model pendekatan ini adalah [3], mengajukan metode yang dinamakan *Twopics*. *Twopics* dibagi kedalam dua langkah *high-level*. Pada langkah pertama, menemukan entitas di setiap *tweet*, *disambiguate tweet* tersebut, dan mengambil *sub-tree* kategori folksonomi yang berisi entitas ambigu. Karena output dari langkah ini adalah seperangkat kategori untuk *tweets*, maka langkah tersebut dinamakan langkah *Discover Categories*. Pada langkah kedua, dihasilkan profil topik bagi pengguna berdasarkan kategori yang ditemukan terkandung dalam *sub-tree* dan disebut sebagi langkah *Discover Profil*. Referensi [5], dalam penelitiannya tentang *twitterrank* yaitu pencarian *topic-sensitive* yang berpengaruh pada pengguna twitter. *Twitterrank* mengukur pengaruh dengan mengambil kedua kemiripan topik antara pengguna dan struktur link ke akun pengguna. Latar Belakang Riset tersebut adalah homofili tidak ada dalam konteks twitter. Hal ini membenarkan bahwa ada beberapa pengguna twitter yang secara serius "*follow*" seseorang karena minat topiknya daripada hanya bermain-main. Penelitian yang telah dikerjakan oleh [6] yaitu memecahkan masalah minat pengguna dengan memanfaatkan modifikasi model *author-topic* dinamakan Model *twitteruser*. Model *author-topic* adalah model generatif yang meneruskan metode *Laten Dirichlet Allocation* (*LDA*), untuk memasukkan informasi kepemilikan. Model ini dimodifikasi dengan memperkenalkan *Latent Variable* untuk menunjukkan apakah tweet berhubungan dengan minat dari penggunanya.

Pendekatan berbasis *Mechine Learning* mengembangkan model pengguna berdasarkan data latih. Model ini dalam formalisme internal dan seperti kotak hitam untuk pengguna. Pendekatan ini tidak banyak digunakan karena membutuhkan data yang ekstensif untuk pemodelan minat pengguna. Sifat dinamis dari minat pengguna membutuhkan proses pembelajaran yang berkelanjutan yang biasanya tidak terjadi pada metode pembelajaran dari pembelajaran minat pengguna.

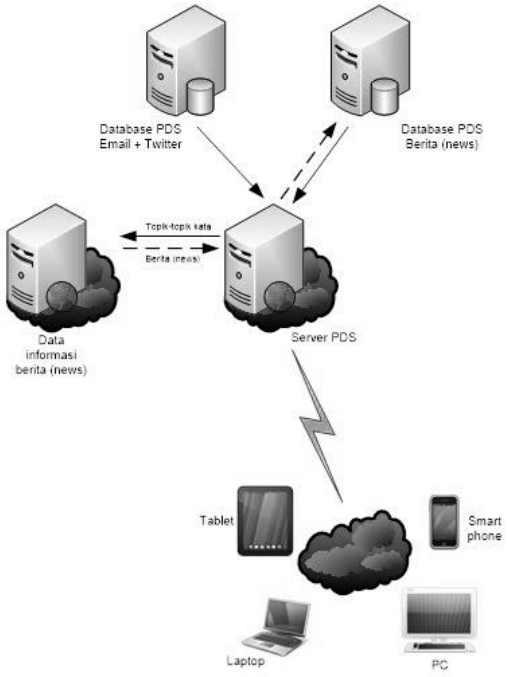
Dalam pendekatan *ontology based*, profil pengguna yang dihasilkan didasarkan pada beberapa referensi ontologi. Pemodelan minat pengguna untuk web sangat sulit karena tidak terdapat ontologi yang sepenuhnya dapat menutupi kepentingan pengguna. Banyak orang telah menggunakan *Open Directory Hierarchy* (*ODH*) sebagai ontologi untuk menemukan minat pengguna. *ODH* adalah sebuah hirarki yang sangat luas dan pengguna umum mungkin hanya memiliki beberapa kepentingan ketika dipetakan ke hirarki kepentingan ini.

Pada penelitian sebelumnya, [1] melakukan modifikasi terhadap algoritma *TF*IDF* menjadi *TF*IDF*DF* dengan tujuan untuk mendapatkan hasil ekstraksi topik yang lebih representatif dari informasi email dan media sosial. Dari hasil ekstraksi topik dengan menggunakan algoritma modifikasi *TF*IDF*DF* masih terdapat kekurangan yaitu kata (*term*) yang muncul terdapat kata berimbuhan, kata berimbuhan bukan merupakan kata yang murni dan bisa dikatakan sebagai *noisy text*. Kata yang berimbuhan juga kurang tepat digunakan sebagai topik dari suatu dokumen, oleh karena itu kata yang berimbuhan terlebih dulu harus dinormalkan untuk menjadi sebuah kata dasar. Untuk menormalkan suatu kata berimbuhan harus ditambahkan suatu proses yang dinamakan proses *stemming*. Proses *stemming* ini akan menghilangkan awalan dan akhiran, sehingga kata yang berimbuhan akan berubah menjadi bentuk kata dasar. Pada analisa hasil, akan dibandingkan kata hasil ekstraksi menggunakan metode *TF*IDF*DF* dengan penambahan proses *stemming* dan kata hasil ekstraksi tanpa penambahan proses *stemming*, dengan tujuan untuk mengetahui hasil terbaik informasi berita yang didapatkan menggunakan kata hasil ekstraksi topik.

II. METODE

A. Perancangan Sistem

Perancangan sistem dan topologi jaringan pada hardware tidak mengalami perubahan dari penelitian sebelumnya [1]. Perubahan yang terjadi pada penelitian ini terletak pada sisi *software*, dimana terdapat menambahan proses *stemming* pada proses *text preprocessing*. Perancangan sistem dan topologi jaringan untuk ekstraksi topik diperlihatkan pada Gambar 1.



Gambar 1. Desain sistem ekstraksi topik pengguna

Data informasi email dan twitter yang tersimpan dalam database *Personal Digital Secretary* (PDS) akan digali (*mining*) untuk mendapatkan topik pengguna dengan tahapan proses yaitu *text preprocessing* kemudian pembobotan kata. Setelah didapatkan topik pengguna, maka kata-kata yang mengandung topik ini kemudian akan digunakan untuk mengambil berita yang berkesesuaian dengan topik tersebut. Sehingga akan didapatkan berita-berita yang sesuai dengan topik pengguna. Setelah mendapatkan informasi berita dari internet maka data informasi berita tersebut disimpan ke dalam database yang selanjutkan bersama-sama data informasi email, dan social media twitter akan diproses lebih lanjut yang kemudian akan ditampilkan ke user melalui antar muka yang berbasis desktop dan mobile.

B. Dokumen Uji

Untuk pengujian ekstraksi topik pada dokumen digunakan dokumen uji sumber berupa dokumen uji email, dan dokumen uji tweet message. Sedangkan untuk pengambilan informasi digunakan dokumen uji berita yang akan diambil sesuai dengan topik yang didapatkan.

Dokumen uji email didapatkan dari email yang diambil dengan menggunakan aplikasi *email client pop3*. *Field* yang diperlukan pada *content* email adalah *message body*. Sedangkan untuk *field to, cc, bcc, from* dan *subject* sementara tidak digunakan dalam proses kali ini.

Dokumen uji *tweet message* didapatkan dari *twitter* yang diambil dengan menggunakan aplikasi API twitter. Data *tweet message* ini kemudian disimpan dalam database yang akan diproses lebih lanjut untuk didapatkan topiknya. Data *tweet message* ini berjumlah 20 buah. Pada dokumen *twitter, field* yang diambil dengan menggunakan API *twitter* hanya *field tweet text* saja.

Dokumen uji berita mengambil data dari berbagai macam sumber dan berbagai macam kategori Tabel I menunjukan data kategori berita yang akan diambil. Data informasi berita ini dalam bentuk *RSS feed* sehingga jika ingin mengakses keseluruhan berita, pengguna akan diarahkan ke *browser* untuk melihat informasi secara lebih lengkap. Sumber berita yang diambil berasal dari dalam dan luar negeri. Untuk informasi berita luar negeri dalam bahasa Inggris.

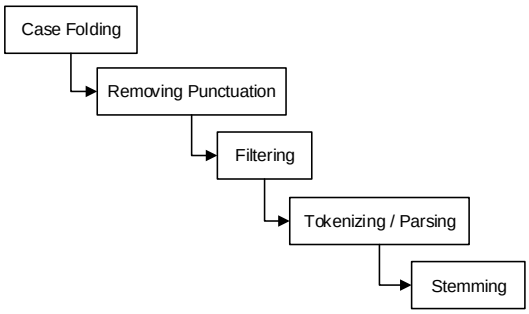
TABEL I.
KATEGORI BERITA YANG DIAMBIL

No.	Kategori
1.	General
2.	Movie & TV
3.	Cyberlife
4.	Business
5.	Finance
6.	Football
7.	MotoGP
8.	Otomotive
9.	Health

No.	Kategori
10.	Travel

C. Text Preprocessing

Text preprocessing merupakan proses awal pada teks yang digunakan untuk menghilangkan *noise*, memperbaiki struktur teks, serta memisahkan teks ke dalam bentuk kata. Beberapa proses dalam *text preprocessing* yaitu *case folding, removing punctuation, filtering, tokenizing/parsing,* dan *stemming,* yang diperlihatkan pada Gambar 2.



Gambar 2. Proses dalam *text preprocessing*

Case folding adalah suatu proses merubah *letter case* dalam bentuk yang sama dalam bentuk huruf kecil atapun huruf besar semua. Setelah suatu huruf dalam suatu kata berada pada bentuk yang sama, maka suatu kata tersebut akan dibandingkan untuk kebutuhan pengolahan teks tertentu. Dalam pengembangan sebuah perangkat lunak ataupun pengolahan teks seringkali dilakukan "normalisasi" teks dengan tujuan untuk perbandingan. Salah satu cara yang paling dasar untuk menormalkan teks yang digunakan untuk perbandingan adalah membandingkan dengan cara "*case sensitive*".

Removing punctuation adalah proses penghilangan tanda baca, sehingga karakter-karakter yang tidak digunakan untuk pengolahan teks akan dihilangkan. Tanda baca dalam kode program bisa diartikan sebagai semua karakter selain huruf a-z atau A-Z, dan untuk kasus ini angka bisa dihilangkan karena tidak dipakai dalam ekstraksi topik. Setelah tanda baca dihilangkan maka bentuk dokumen akan berupa data teks huruf saja dan akan siap diolah pada proses selanjutnya. Dalam suatu kalimat maupun paragraf, sering kali kita akan menjumpai tanda baca. Tanda baca ini kebanyakan tidak digunakan untuk pengolahan teks dan seringkali dihilangkan, kalau tanda baca ini tidak dihilangkan maka akan mengganggu dalam dalam proses *text retrieval*. Jika memang tanda baca ini diperlukan maka tidak akan dihapus tapi hal tersebut adalah sebagai pengecualian.

Filtering atau penyaringan adalah tahap mengambil kata-kata penting dari hasil *token*. Bisa menggunakan algoritma *stoplist* (membuang kata yang kurang penting) atau *wordlist* (menyimpan kata penting). *Stoplist/stopword* adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan *bag-of-words*. Contoh *stopwords* dalam bahasa indonesia adalah “yang”, “dan”, “di”, “dari” dan seterusnya.

Tokenization adalah proses memecah aliran teks menjadi kata-kata, frase, simbol, atau elemen lain yang disebut *token* bermakna. Daftar *token* menjadi masukan untuk diproses lebih lanjut seperti *parsing* atau *text mining*. *Tokenization* ini berguna baik dalam linguistik (di mana itu adalah bentuk segmentasi teks), dan dalam ilmu komputer, di mana ia merupakan bagian dari analisis leksikal. *Tokenization* terjadi di tingkat kata. Namun, kadang-kadang sulit untuk mendefinisikan apa yang dimaksud dengan "kata". Seringkali *tokenizer* bergantung pada heuristik sederhana, misalnya semua string bersebelahan karakter abjad adalah bagian dari satu *token*, demikian pula dengan angka. *Token* biasanya dipisahkan oleh karakter spasi, seperti spasi atau garis istirahat, atau dengan karakter tanda baca.

Stemming adalah tahap mencari *root* kata dari tiap kata hasil *filtering*. Pada tahap ini dilakukan proses pengembalian berbagai bentukan kata ke dalam suatu representasi yang sama. Tahap ini kebanyakan dipakai untuk teks berbahasa inggris dan lebih sulit diterapkan pada teks berbahasa Indonesia. Hal ini dikarenakan bahasa Indonesia tidak memiliki rumus bentuk baku yang permanen. Proses *stemming* pada teks berbahasa Indonesia lebih rumit/kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan *root word* dari sebuah kata. Algoritma *stemming* yang bisa digunakan adalah algoritma *stemming porter* untuk bahasa inggris dan algoritma *stemming nazief adriani* untuk bahasa indonesia.

Porter stemmer secara de facto merupakan algoritma standar yang digunakan untuk English *stemming*. *Porter stemmer* ditulis oleh Martin Porter dan telah dipublikasikan pada bulan Juli 1980. *Porter stemming algorithm* banyak diimplementasikan dalam aplikasi, tersedia dalam berbagai bahasa pemrograman dan didistribusikan secara gratis. Berikut adalah algoritma dari *Porter Stemmer* untuk bahasa inggris [7]:

1. Menghilangkan kata jamak dan akhiran "-*ed*" atau "-*ing*".
2. Mengganti terminal "*y*" menjadi "*I*" ketika ada vokal didalam proses *stemming*.
3. Memetakan akhiran ganda menjadi akhiran tunggal: *-ization*, *-ationl*, dst.
4. Mengatur akhiran, *-full*, *-ness*, dst.
5. Melepas *-ant*, *-ence*, dst.
6. Menghilangkan sebuah akhiran *–e*.

Proses *stemming* pada teks berbahasa Indonesia dengan menghilangkan sufiks, prefiks, dan konfiks untuk mendapatkan kata dasar (*root word*). Algoritma Nazief & Adriani sebagai algoritma *stemming* untuk teks berbahasa Indonesia yang memiliki kemampuan prosentase keakuratan (presisi) yang baik dan banyak diimplementasikan dalam studi kasus. Berikut adalah Algoritma Nazief & Adriani untuk *stemming* bahasa Indonesia [8]:

1. Cari kata yang akan distem dalam kamus. Jika ditemukan maka diasumsikan bahwa kata tesebut adalah *root word*. Maka algoritma berhenti.
2. *Inflection Suffixes* ("-lah", "-kah", "-ku", "-mu", atau "-nya") dibuang. Jika berupa partikel ("-lah", "-kah", "-tah" atau "-pun") maka langkah ini diulangi lagi untuk menghapus *Possesive Pronouns* ("-ku", "-mu", atau "-nya"), jika ada.
3. Hapus *Derivation Suffixes* ("-i", "-an" atau "-kan"). Jika kata ditemukan di kamus, maka algoritma berhenti. Jika tidak maka ke langkah 3a
   a. Jika "-an" telah dihapus dan huruf terakhir dari kata tersebut adalah "-k", maka "-k" juga ikut dihapus. Jika kata tersebut ditemukan dalam kamus maka algoritma berhenti. Jika tidak ditemukan maka lakukan langkah 3b.
   b. Akhiran yang dihapus ("-i", "-an" atau "-kan") dikembalikan, lanjut ke langkah 4.
4. Hapus *Derivation Prefix*. Jika pada langkah 3 ada sufiks yang dihapus maka lanjutkan ke langkah 4a, jika tidak lanjutkan ke langkah 4b.
   a. Periksa tabel kombinasi awalan-akhiran yang tidak diijinkan. Jika ditemukan maka algoritma berhenti, jika tidak pergi ke langkah 4b.
   b. Untuk i = 1 to 3, tentukan tipe awalan kemudian hapus awalan. Jika root word belum juga ditemukan lakukan langkah 5, jika sudah maka algoritma berhenti. Catatan: jika awalan kedua sama dengan awalan pertama algoritma berhenti.
5. Melakukan *Recoding*.
6. Jika semua langkah telah selesai tetapi tidak juga berhasil maka kata awal diasumsikan sebagai *root word*. Proses selesai.

D. Pembobotan Kata

Dasar dari *TF*IDF* adalah dari teori pemodelan bahasa, bahwa suatu kata dalam sebuah dokumen bisa dipisahkan dengan atau tanpa sifat dari penggolongannya [9], dengan kata lain suatu kata apakah merupakan topik dari sebuah dokumen atau tidak. Penggolongan dari suatu kata dari sebuah dokumen dievaluasi dengan *TF* (*Term Frequency*), sedangkan *IDF* (*Inverse Document Frequency*) digunakan untuk mengukur seberapa penting kata tersebut dalam kumpulan dokumen. *dfi* adalah *Document Frequency* (*DF*), yaitu banyaknya dokumen yang mengandung *i*th *term* dalam koleksi [10]. Persamaan (1) menunjukkan formula dari *TF*IDF*. Dengan tujuan untuk mendapatkan hasil ekstraksi topik yang lebih representatif maka formula dari *TF*IDF* tersebut dimodifikasi menjadi *TF*IDF*DF* yang diperlihatkan pada persamaan (2).

$$wi, j = tfi, j \times \log(\frac{N}{dfi}) \quad (1)$$

$$wi, j = tfi, j \times \log(\frac{N}{dfi}) \times dfi \quad (2)$$

III. HASIL

Dalam hasil dan pembahasan dari penelitian ini akan ditunjukkan dua proses yang berbeda yang akan dievaluasi yaitu proses ekstrasi topik dengan menggunakan penambahan proses stemming dan yang kedua adalah proses ekstraksi topik tanpa proses stemming.

Setelah dilakukan proses ekstraksi dengan menggunakan penambahan proses stemming terhadap modifikasi pembobotan algoritma *TF*IDF*DF*, terlihat ada perbedaan dengan hasil ekstraksi terhadap modifikasi pembobotan algoritma *TF*IDF*DF* tanpa menggunakan proses *stemming*. Hal ini dikarenakan proses *stemming* akan memurnikan kata-kata yang berimbuhan menjadi kata dasar Tabel II dan Tabel III memperlihatkan perbandingan 20 besar kata (*term*) hasil ekstraksi terhadap dua metode tersebut.

TABEL II.
KATA HASIL EKSTRAKSI TOPIK YANG DIAMBIL DARI SUMBER INFORMASI EMAIL TANPA MENGGUNAKAN PROSES *STEMMING*

*TF*IDF*DF*				
No	**Topik**	***tf***	***df***	***Weight***
1	android	16	1	16
2	nilai	10	2	13.9794
3	akakom	9	2	12.5815
4	dosen	8	3	12.5491
5	google	7	3	10.9805
6	mobil	10	1	10
7	praktikum	6	2	8.38764
8	rekor	8	1	8
9	ac	5	2	6.9897
10	id	5	2	6.9897
11	http	5	2	6.9897
12	com	2	4	6.36704
13	yue	6	1	6
14	han	6	1	6
15	matakuliah	4	2	5.59176
16	silabi	2	2	5.59176
17	panduan	4	2	5.59176
18	uas	5	1	5
19	device	5	1	5
20	manager	5	1	5

TABEL III.
KATA HASIL EKSTRAKSI TOPIK YANG DIAMBIL DARI SUMBER INFORMASI EMAIL DENGAN MENGGUNAKAN PROSES *STEMMING*

*TF*IDF*DF + stemming*				
No	**Topik**	***tf***	***df***	***Weight***
1	android	16	1	16
2	nilai	10	2	13.9794
3	akakom	9	2	12.5815
4	dosen	8	3	12.5491
5	googl	7	3	10.9805
6	mobil	10	1	10
7	praktikum	6	2	8.38764
8	rekor	8	1	8
9	id	5	2	6.9897
10	ac	5	2	6.9897
11	http	5	2	6.9897
12	com	2	4	6.36704
13	yue	6	1	6
14	h	6	1	6
15	matakuliah	4	2	5.59176
16	silabi	2	2	5.59176
17	pandu	4	2	5.59176
18	Ua	5	1	5
19	Devic	5	1	5
20	Manag	5	1	5

IV. PEMBAHASAN

Kata (*term*) yang dihasilkan dari proses ekstraksi akan mempengaruhi hasil yang didapatkan dari pencarian berita. Tabel IV memperlihatkan waktu komputasi yang diperlukan untuk melakukan proses ekstraksi pada masing-masing metode. Dengan ditambahkannya proses *stemming* maka waktu komputasi menjadi bertambah karena proses *stemming* sendiri membutuhkan waktu untuk bekerja.

TABEL IV.
ANALISA HASIL UJICOBA PERBADINGAN ALGORITMA TF*IDF*DF DAN ALGORITMA TF*IDF*DF + STEMMING

Algoritma	**Jumlah Dokumen**	**Waktu Komputasi**
*TF*IDF*DF*	10	24.4998002052
*TF*IDF*DF + stemming*	10	30.4970760345

*A. Pemilihan Berita Berdasarkan Topik dengan Algoritma TF*IDF*DF*

Setelah mendapatkan topik dari ekstraksi dokumen terhadap email dan twitter maka langkah selanjutnya adalah pencarian berita berdasarkan topik, dimana sumber-sumber berita diambil dari link rss dari kategori-kategori yang ditentukan pada Tabel I. Pencarian berita dalam bahasa Indonesia dan bahasa Inggris. Data informasi berita yang dihasilkan dari pencarian berdasarkan topik dari sumber email diperlihatkan pada Tabel V.

TABEL V. PEMILIHAN BERITA BERDASARKAN TOPIK DARI SUMBER EMAIL DENGAN ALGORITMA TF*IDF*DF

Topik	Judul Informasi Berita
mini	Toyota. Rivals Tackle Hacking (10/22/14)
mini	Feds Order Full Takata Airbag Recall (11/26/14)
google	Barra Reflects on First Year (1/15/15)
manager	Franklin Square Provides New Unitranche Term Loan to Dent Wizard International Corp.
teknologi	Aneka Aplikasi Untuk Mencari Jodoh Saat Traveling
mobil	Arek Suroboyo Cetak Konsumsi BBM 22.1 Km/Liter Pakai Mirage
mobil	Apakah Tidak Basi Baru Bicara Mobnas Sekarang?
mobil	3 Teknologi Penyelamat Nyawa
mobil	Gaikindo Tepis Pabrikan Jepang Pelit Teknologi
ac	El Shaarawy Tegaskan Inzaghi Masih Didukung Penuh Para Pemain Milan
manager	Adi Rusli Menyebrang ke VMware
google	Google Segera Matikan Gtalk
google	Google Beli Aplikasi Backup Foto
teknologi	Heboh Bonus Vendor IT China: Semalam Bersama Bintang Porno
mobil	Menanti Robot Garapan Samsung
teknologi	Santri Harus Melek IT. Ngaji Bisa Pakai iPad
google	Kenalan dengan Wanita di Balik Kehebatan Google Maps
mobil	Toyota Fortuner Terjebak Banjir di Underpass Kemayoran
mobil	Kejar Buronan di Jalanan. Polisi Lepaskan Tembakan
mobil	Kerjasama Indonesia-Malaysia Soal Mobil Nasional Timbulkan Pertanyaan
mobil	Kontroversi Mobil Nasional
cm	Bayi Raksasa dengan Berat .4 Kg Lahir di Probolinggo
Mobil	Minim Lokasi Pengungsian. Korban Banjir 'Ngungsi' di Angkot

TABEL VI. PEMILIHAN BERITA BERDASARKAN TOPIK DARI SUMBER TWITTER DENGAN ALGORITMA TF*IDF*DF

Topik	Judul Informasi Berita
honda	Used-Car Dealer Crackdown (9/22/14)
honda	How Honda Fended Off Nissan (12/15/14)
honda	Audiâ€™s SUV â€˜opportunityâ€™ (1/26/15)
indonesia	Mengenal Cantiknya Seragam Batik Pramugari Garuda
indonesia	Pria Ini Koleksi 1.000 Seragam Pramugari. Garuda Indonesia Pun Ada
indonesia	Kisah Fobia Naik Pesawat Sang Desainer Ternama. Anne Avantie
indonesia	Tenaga Kesehatan Minim. Pasien Gangguan Jiwa Banyak yang Tak Tertangani
orang	Musim Hujan Rawan Banjir. Ini Trik Agar Kebersihan Kulit Anak Tetap Terjaga
indonesia	Kantor Urusan Agama Turut Didorong Sosialisasikan Kesehatan Reproduksi
motor	Hak Paten Nama IIMS Bukan Milik Gaikindo
motor	Setelah NMAX. Yamaha Siapkan WR250R
indonesia	Perusahaan AM Hendropriyono Belum Ajukan Diri Jadi Anggota Gaikindo
motor	Banjir. Pengiriman Motor Yamaha Terganggu
motor	Gara-gara Nyalip dari Kiri. Pemotor Didenda Rp 17 Juta
honda	Rossi: Honda Bisa Menang Kalau Balapannya Hari Ini. tapi...
motor	Honda Sudah di Level yang Sama Seperti Tahun Lalu
honda	Untuk Jadi Juara. Pedrosa Harus Lebih Agresif
honda	Pedrosa Puas dengan Tes Pramusim di Sepang
motor	Donington Park Urung Gelar Balapan di MotoGP 2015
orang	Usai Posting. Blogger Ini Bunuh Diri
motor	Hindari Motor. Minibus Terbalik di Jalan MT Haryono
amp	Underpass Kemayoran Terendam Banjir
orang	Preman Pukuli Wartawan Televisi
amp	Banjir Terjadi di Mana-mana
amp	Keluarkan Sprindik. Bareskrim Segera Periksa Saksi Kasus Bambang Widjojanto
motor	Banjir. Jalan Tol Dibuka untuk Sepeda Motor
motor	Razia Atribut TNI. Warga Dipaksa Melepas Celana Lorong oleh Petugas
orang	Bentrokan Warga di Maluku Tenggara. 3 Orang Terluka
indonesia	Aktivis Perempuan di Yogya Demo Desak Jokowi Tegas Sikapi KPK Vs Polri

Sedangkan data informasi berita yang dihasilkan dari pencarian berdasarkan topik dari sumber twitter diperlihatkan pada Tabel 6. Data informasi ini kemudian disimpan kedalam database yang berelasi terhadap pengguna. Sehingga setiap pengguna memiliki data informasi berita yang berbeda-beda berdasarkan topik yang didapatkan dari hasil ekstraksi dari data informasi email dan data informasi twitter masing-masing pengguna.

*B. Pemilihan Berita Berdasarkan Topik dengan penambahan proses stemming Algoritma TF*IDF*DF*

Setelah mendapatkan topik dari ekstraksi dokumen terhadap email dan twitter dengan penambahan proses *stemming* pada algoritma *TF*IDF*DF* maka langkah selanjutnya adalah pencarian berita berdasarkan topik, dimana sumber-sumber berita diambil dari kategori yang ada pada Tabel I. Data informasi berita yang dihasilkan dari pencarian berdasarkan topik dengan penambahan proses *stemming* pada algoritma *TF*IDF*DF* dari sumber email diperlihatkan pada Tabel VII, sedangkan dari sumber data informasi berita yang dihasilkan dari pencarian berdasarkan topik dari sumber twitter diperlihatkan pada Tabel VIII.

TABEL VII.
PEMILIHAN BERITA BERDASARKAN TOPIK DARI SUMBER EMAIL DENGAN PENAMBAHAN PROSES *STEMMING* PADA ALGORITMA *TF*IDF*DF*

Topik	Judul Informasi Berita
mini	Toyota. Rivals Tackle Hacking (10/22/14)
mini	Feds Order Full Takata Airbag Recall (11/26/14)
mobil	Arek Suroboyo Cetak Konsumsi BBM 22.1 Km/Liter Pakai Mirage
mobil	Apakah Tidak Basi Baru Bicara Mobnas Sekarang?
mobil	3 Teknologi Penyelamat Nyawa
buat	Gara-gara Nyalip dari Kiri. Pemotor Didenda Rp 17 Juta
mobil	Gaikindo Tepis Pabrikan Jepang Pelit Teknologi
ac	El Shaarawy Tegaskan Inzaghi Masih Didukung Penuh Para Pemain Milan
mobil	Menanti Robot Garapan Samsung
mobil	Toyota Fortuner Terjebak Banjir di Underpass Kemayoran
mobil	Kejar Buronan di Jalanan. Polisi Lepaskan Tembakan
mobil	Kerjasama Indonesia-Malaysia Soal Mobil Nasional Timbulkan Pertanyaan
mobil	Kontroversi Mobil Nasional
cm	Bayi Raksasa dengan Berat .4 Kg Lahir di Probolinggo
mobil	Minim Lokasi Pengungsian. Korban Banjir 'Ngungsi' di Angkot

TABEL VIII.
PEMILIHAN BERITA BERDASARKAN TOPIK DARI SUMBER TWITTER DENGAN PENAMBAHAN PROSES *STEMMING* PADA ALGORITMA *TF*IDF*DF*

Topik	Judul Informasi Berita
honda	Used-Car Dealer Crackdown (9/22/14)
honda	How Honda Fended Off Nissan (12/15/14)
honda	Audiâ€™s SUV â€˜opportunityâ€™ (1/26/15)
indonesia	Mengenal Cantiknya Seragam Batik Pramugari Garuda
indonesia	Pria Ini Koleksi 1.000 Seragam Pramugari. Garuda Indonesia Pun Ada
indonesia	Kisah Fobia Naik Pesawat Sang Desainer Ternama. Anne Avantie
indonesia	Tenaga Kesehatan Minim. Pasien Gangguan Jiwa Banyak yang Tak Tertangani
orang	Musim Hujan Rawan Banjir. Ini Trik Agar Kebersihan Kulit Anak Tetap Terjaga
indonesia	Kantor Urusan Agama Turut Didorong Sosialisasikan Kesehatan Reproduksi
motor	Hak Paten Nama IIMS Bukan Milik Gaikindo
motor	Setelah NMAX. Yamaha Siapkan WR250R
indonesia	Perusahaan AM Hendropriyono Belum Ajukan Diri Jadi Anggota Gaikindo
motor	Banjir. Pengiriman Motor Yamaha Terganggu
honda	Rossi: Honda Bisa Menang Kalau Balapannya Hari Ini. tapi...
motor	Honda Sudah di Level yang Sama Seperti Tahun Lalu
honda	Untuk Jadi Juara. Pedrosa Harus Lebih Agresif
honda	Pedrosa Puas dengan Tes Pramusim di Sepang
motor	Donington Park Urung Gelar Balapan di MotoGP 2015
indonesia	Adi Rusli Menyebrang ke VMware
orang	Google Beli Aplikasi Backup Foto
orang	Usai Posting. Blogger Ini Bunuh Diri
motor	Hindari Motor. Minibus Terbalik di Jalan MT Haryono
orang	Preman Pukuli Wartawan Televisi
motor	Banjir. Jalan Tol Dibuka untuk Sepeda Motor
motor	Razia Atribut TNI. Warga Dipaksa Melepas Celana Lorong oleh Petugas
orang	Bentrokan Warga di Maluku Tenggara. 3 Orang Terluka
indonesia	Aktivis Perempuan di Yogya Demo Desak Jokowi Tegas Sikapi KPK Vs Polri

V. SIMPULAN DAN SARAN

Tujuan dari serangkaian penelitian yang dilakukan adalah untuk membandingkan hasil ekstraksi topik pengguna dengan menggunakan modifikasi algoritma *TF*IDF* yaitu *TF*IDF*DF* dan penambahan proses stemming pada algoritma *TF*IDF*DF*. Masing-masing hasil ekstraksi topik tersebut kemudian akan digunakan untuk proses pengambilan data informasi berita dari luar, berdasarkan topik yang didapatkan. Serangkaian uji coba telah dilaksanakan dan didokumentasikan, yang menghasilkan beberapa topik dari hasil ekstraksi dokumen informasi email dan twitter. Dari proses pembuktian tersebut maka dapat diambil kesimpulan sebagai berikut:

1. Hasil dari ekstraksi topik dengan algoritma *TF*IDF*DF* menunjukkan bahwa kata (*term*) yang didapatkan masih mengandung kata berimbuhan, hal ini akan mempengaruhi hasil pencarian berita berdasarkan topik pengguna karena berita yang diambil akan mengandung kata yang identik dengan kata dari topik-topik yang didapatkan.
2. Dengan penambahan proses *stemming* pada algoritma *TF*IDF*DF* menunjukkan bahwa kata (*term*) hasil dari proses ekstraksi yang didapatkan telah menjadi kata dasar karena adanya proses *stemming*. Kata dasar disini merupakan bentuk dasar yang merupakan indikasi sebuah topik. Dibandingkan dengan kata yang berimbuhan, kata dasar merupakan indikasi yang tepat untuk sebuah topik.

REFERENSI

[1] Pramono, L. H., A.S. Rohman and H. Hindersah, "Modified weighting method in TF*IDF algorithm for extracting user topic based on email and social media in Integrated Digital Assistant," 2013 Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT & ICeV-T). Bandung: IEEE, 2013, pp. 1-6.

[2] Alsmadi, I. and Alhami, I., "Clustering and classification of email contents," Journal of King Saud University - Computer and Information Sciences, vol. 1, no. 27, pp. 46–57, January 2015.

[3] Michelson, M., and S. Macskassy, "Discovering users' topics of interest on twitter: a first look," AND '10 Proceedings of the fourth workshop on Analytics for noisy unstructured text data. New York: ACM, 2010, pp. 73-80.

[4] Wasim, M., I. Shahzadi, Q. Ahmad and W. Mahmood, "Extracting and modeling user interests based on social media," Multitopic Conference (INMIC), 2011 IEEE 14th International, Karachi: IEEE, 2011, pp. 284-289.

[5] Weng, J., E.P. Lim, J.Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential twitterers". Proceedings of the third ACM international conference on Web search and data mining. New York: ACM, 2010, pp. 261-270.

[6] Xu, Z., L. Ru, L. Xiang, and Q. Yang, "Discovering User Interest on Twitter with a Modified Author-Topic Model," IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Lyon: IEEE/WIC/ACM, 2011, pp. 422-429.

[7] Waegel, D, "Porter's stemmer," Applications of Natural Language Processing, 17 May 2011, pp. 1-29.

[8] R. Setiawan, A. Kurniawan, W. Budiharto, I. H. Kartowisastro and H. Prabowo, "Flexible affix classification for stemming Indonesian Language," 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Mai, 2016, pp. 1-6.

[9] Azam, N. and Yao, J., "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Systems with Applications, 39(5), 2012, pp. 4760–4768.

[10] Ko, Y., "A study of term weighting schemes using class information for text classification," SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, New York: ACM, 2012, pp. 1029-1030.