Jurnal Mantik Penusa

Volume 3, No. 3, Desember 2019

e-ISSN 2580-9741
p-ISSN 2088-3943

PRA-PEMROSESAN TEKS PADA GRUP WHATSAPP UNTUK PEMODELAN TOPIK

Maya Cendana[1] dan Silvester Dian Handy Permana[2]

[12]Prodi Teknik Informatika, Universitas Trilogi, Jl. TMP Kalibata Jakarta Selatan 12760, Indonesia

E-mail: [1]mcen2030@gmail.com, [2]handy@trilogi.ac.id

Abstrak

Fasilitas grup *chat* yang terdapat pada *WhatsApp* memungkinkan pengguna untuk saling berkomunikasi dalam kelompok yang diminatinya. Semakin banyak dan aktif sebuah grup *WhatsApp* maka akan meningkatkan jumlah pesan yang dikirimkan di dalam grup. Hal ini dapat memicu keinginan pengguna untuk mengetahui makna ataupun pola tersembunyi yang terdapat di dalam grup tersebut. Oleh karena itu, penelitian dibidang *text mining*, terutama *natural language programming (NLP)* menjadi semakin populer. Salah satunya adalah analisis data *chat* yang dapat menghasilkan informasi topik yang sering dibicarakan di dalam grup *WhatsApp*, terutama bagi pengguna yang tidak memiliki banyak waktu untuk membaca *chat* di dalam grup, ataupun jika grup tersebut merupakan grup diskusi, maka topik utama diskusi dapat diperoleh melalui teknik pemodelan topik. Tahapan penelitian yang dilalui adalah (1) pra-pemrosesan data, (2) pemrosesan teks menggunakan model atau algoritma tertentu, (3) analisis hasil, dan (4) evaluasi. Pada penelitian tahap awal ini akan dilakukan studi pendahuluan dan pra-pemrosesan data menggunakan bahasa pemrograman *python* yang mencakup proses tokenisasi, *filtering* dan *stemming*. Prototipe pra-pemrosesan data ini diharapkan dapat digunakan untuk kasus berbeda dengan data input *chat-log Whatsapp*. Aturan-aturan bahasa, terutama bahasa gaul atau bahasa *alay* yang nantinya ditemukan di dalam studi kasus *chat-log* grup *WhatsApp* diharapkan juga dapat memperkaya korpus Bahasa Indonesia.

Keywords: *Topic Modelling, Data Pre-processing, Python, WhatsApp (WA*

1. Pendahuluan

Aplikasi *WhatsApp (WA)* merupakan salah satu aplikasi *chatting* yang sangat popular di dunia. Berdasarkan Statista [1], pengguna WA pada bulan Juli 2018 telah mencapai 1,5 miliar, meningkat tiga kali lipat sejak 2014 [2]. Di Indonesia sendiri, terdapat 132,7 juta pengguna internet yang 40% diantaranya adalah pengguna WA, disusul oleh Line, BBM, Facebook Messenger, Skype dan WeChat [3]. Beberapa faktor yang mempengaruhi popularitas WA di kalangan anak muda adalah faktor ekonomi, komunikasi, hiburan, pengalihan rutinitas, kepedulian, tren, pemecahan masalah, dan kebutuhan sosial [4]. Pengguna mendapatkan kesenangan tersendiri ketika berbagi masalah dengan temannya di WA. Selain dapat digunakan untuk berkomunikasi antar-individu (*one to one communication*), WA juga memiliki fasilitas yang memungkinkan pengguna membuat grup sebagai media komunikasi kelompok. Grup yang dibuat pun beragam sesuai dengan tujuan, seperti grup keluarga, kantor, warga perumahan, reuni sekolah, dll. Belum terdapat penelitian jumlah rata-rata grup yang diikuti oleh orang Indonesia, namun melalui survei terhadap anak muda di Italia pada tahun 2015, diperoleh data sebanyak 39% anak muda rata-rata tergabung dalam 5-10 grup WA [5]. Interaksi masyarakat global ini pun menjadi semakin penting karena terdapat peleburan antara kehidupan dunia nyata dan dunia maya, bahkan Forbes merilis kiat-kiat sukses mengaktifkan grup WA [6], termasuk rekomendasi *tools* yang dapat menganalisis fitur-fitur WA maupun pesan WA, seperti WhatsAnalyzer [7] dan ChatVisualizer [8].

Keberadaan grup WA juga dapat membantu penanganan bencana alam, yaitu sebagai alat komunikasi yang interaktif untuk mengumpulkan data penting dari para sukarelawan yang bertugas di beberapa titik/wilayah bencana. Analisis chat-log dari WA digunakan untuk mengetahui informasi seperti (1) tempat yang dikunjungi oleh sukarelawan, (2) ketersediaan obat-obatan di lokasi, (3) laporan penting yang perlu tindak-lanjut cepat, juga (4) status pemulihan dan penyelamatan para korban di lokasi tertentu [9]. Dalam ruang lingkup penggunaan yang lebih sederhana, analisis dapat dilakukan untuk memprediksi apakah seseorang merupakan pencandu WA dengan memperhatikan lama penggunaan WA per hari dan seberapa sering pengguna membalas *chat*, dsb [10]. Selain itu juga terdapat penelitian lain untuk memprediksi kebiasaan dan perilaku sebuah grup WA

berdasarkan jenis kelamin dan umur pengguna jika diberikan tindakan tertentu [11].

Metode *sentiment analysis* juga dapat digunakan untuk melihat kecenderungan emosi pengguna dari grup WA tertentu [12][13]. Beberapa penelitian lainnya terhadap grup WA adalah untuk mengetahui hubungan antar-pengguna yang saling memberikan respon di dalam sebuah grup WA [14], maupun menganalisis jam *chatting* paling ramai, jumlah pesan yang dikirimkan, kata yang paling banyak muncul maupun pengguna yang paling aktif di grup WA [15]. Penelitian tersebut diimplementasikan dengan bahasa pemrograman *Python*, meskipun banyak analisis *chat-log* WA lainnya yang juga menggunakan R seperti pada penelitian [16].

Penelitian di bidang analisis teks terhadap data media sosial sudah dimulai sejak satu dekade lalu, namun untuk aplikasi *mobile chatting*, khususnya WA, belum ditemukan topik penelitian di bidang *natural language processing* (NLP) untuk pemodelan topik (*topic modelling*), padahal manfaat yang diperoleh sangat banyak, misalnya untuk grup warga Rusunami Bandar Kemayoran (RBK) Tower A4. Salah satu fungsi utama grup RBK tersebut adalah untuk menampung keluhan warga terhadap fasilitas RBK seperti kerusakan lift, pemadaman listrik, air mati, pembuangan sampah, kebocoran unit, dll. Jika mengimplementasikan pemodelan topik pada kasus ini, maka keluarannya bisa berupa visualisasi keluhan warga berdasarkan tingkat popularitas masalah. Contoh studi kasus lainnya yang akan digunakan dalam penelitian ini adalah untuk menganalisis tren antara topik yang popular pada waktu tertentu di dalam grup WA Dosen Universitas Trilogi. Penelitian yang menghasilkan model atau prototipe untuk aplikasi WA ini juga dapat digunakan untuk menganalisis topik dari *chat-log* komunikasi antar-individu, misalnya untuk mengetahui topik yang paling sering dibicarakan di dalam pesan WA.

Berdasarkan penelitian-penelitian sebelumya, belum terdapat penelitian terkait pemodelan topik (*topic modelling*) terhadap hasil *chat-log* WA. Penelitian lainnya seperti [17][18][19] sudah melakukan pemodelan topik, namun untuk kasus dan sumber data yang berbeda, seperti data yang diperoleh dari *Facebook, Twitter*, dan hasil pencariaan dari *Google Search Engine*. Oleh karena itu, penelitian yang akan dilakukan saat ini adalah pemodelan topik terhadap sumber data yang diperoleh dari grup WA. Sebagai bagian dari *natural language programming*, maka langkah-langkah yang akan diimplementasikan mengikuti tahapan penelitian dalam ranah *text mining* secara umum.

Tahap awal penelitian yang telah dilakukan adalah pra-pemrosesan terhadap data *raw chat-log WA* untuk grup warga RBK Tower A4 dan Grup WA Dosen Universits Trilogi. Tahap kedua yang akan dilakukan adalah tahap pemrosesan teks, yaitu melakukan pemodelan topik dengan pendekatan probabilistik menggunakan *Latent Dirichlet Allocation* (LDA) [20][21]. Proses selanjutnya adalah pemberian nama topik yang memiliki tingkat probabilitas yang tinggi dan membuat visualisasi klaster. Interpretasi hasil pengelompokan akan menggunakan metode intuitif, yaitu metode *human-in-the-loop*.

Tahap pemrosesan teks bisa mencakup proses *Punctuation (P), Numbers (N), Lowercasing (L), Stemming (S), Stopword Removal (W), n-gram Inclusion (3), Infrequently Used Terms (I)*. Denny dalam penelitian [22] menjelaskan secara detail penggunaan beragam cara pemrosesan data tersebut, khususnya untuk jenis *unsupervised learning*. Pemilihan urutan proses tersebut juga memiliki beragam kriteria, misalkan yang paling mendekati dengan penelitian ini adalah tahapan P-L-S-W. Secara detail pra-pemrosesan dilakukan melalui (1) proses pembuangan karakter yang tidak digunakan (*punctuation*), seperti angka, markup/html/tag, tanda baca, *emoticon* dan spesial karakter ($, %, &, etc), (2) proses pembuangan angka (*numbering*) terutama untuk nomor telepon yang biasanya sering ada di dalam percakapan WA, (3) *stemming* untuk mencari akar kata atau kata dasar, serta membuang *stopword.*

Bahasa pemrograman yang digunakan adalah *python*, sedangkan *library* yang digunakan adalah NTLK dan Sastrawi.

Secara ringkas, beberapa tinjauan pustaka yang digunakan sebagai studi pendahuluan dalam penelitian ini dapat dilihat pada Tabel 1.

TABEL 1
TINJAUAN PUSTAKA

Tahun - Peneliti	Topik Penelitian - Judul	Kesimpulan
2016 - Pragna Debnath, Saniul Haque, Somprakash	WhatsApp - Post-disaster Situational Analysis from WhatsApp Group Chats of	Melakukan klastering terhadap informasi penting dari grup WA relawan dokter pada saat bencana alam untuk pengambilan keputusan. Implementasi menggunakan Python, dan TextBlob

Tahun - Peneliti	Topik Penelitian - Judul	Kesimpulan
Bandyopandhyay, Siuli Roy [9]	Emergency Response Providers	sebagai alat analisis sentimen serta kamus yang terdapat dalam WordNet. Model ini dikembangkan dengan *semantic similarity function*. Penelitian selanjutnya adalah mengembangkan sistem penanganan bencana secara *real-time* terhadap data *semi-supervised*.
2016 – R. Jayaparvathy [10]	WhatsApp - Analysis on Social Media Addiction using Data Mining	Mengklasifikasi perilaku pengguna untuk memprediksi tingkat kecanduan penggunaan WA menggunakan teknik *data mining*. Implementasi menggunakan RapidMiner dengan tingkat akurasi 90% (menggunakan opeartor x-validation)
2016 – Avi Rosenfeld, Sigan Sina, David Sarne, Or Avidov, Sarit Kraus [11]	WhatsApp - WhatsApp Usage Patterns and Prediction Models	Analisis dilakukan terhadap 4 juta pesan dari 100 pengguna untuk melihat kebiasaan dan perilaku pengiriman pesan berdasarkan *gender* dan usia. Model prediksi menggunakan algoritma C4.5 dan *Bayesian networks*, serta Weka sebagai *tools* pendukung.
2017 – C. Premalatha, S. Jansi Rani [12]	WhatsApp - Sentimental Analysis of WhatsApp Data Using Data Analytics Techniques	Melakukan analisis sentimen dengan *Analytical Sandbox* terhadap pesan yang terdapat dalam grup WA dan membuatkan visualisasinya sehingga dapat melihat berbagai opini positif dan negatif yang diilustrasikan dalam bentuk emotikon (*anger, fear, disgust, anticipation, joy, sadness, surprise, trust* dan *negative*)
2016 – Sanchita Patil [16]	WhatsApp - WhatsApp Group Data analysis with R	Implementasi bahasa pemrograman R untuk memprediksi level kecanduan dari pengguna grup WA berdasarkan umur dan *gender*. Tahapan penelitian yang dilalui adalah pengumpulan data, transformasi data, *loading* data, *exploratory data analysis* (EDA) dan visuaisasi.
2017 – I Made Kusnanta Bramantya Putra, Renny Pradina Kusumawardani [17]	Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation (LDA)	Pemodelan topik menggunakan metode LDA untuk menemukan pola tertentu pada sebuah dokumen (pesan media sosial) dan memperoleh jumlah topik terbaik berdasarkan nilai *perplexity*. Uji koherensi topik terdiri dari *word intrusion task* dan *topic instrusion task.*Perlu dilakukan normalisasi di awal proses sehingga hasil yang diperoleh lebih optimal.
2018 – Agung Priyanto, Muhammad Rifqi Ma'arif [18]	Implementasi Web Scraping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi Laman Web Tentang Hidroponik	Mengatasi masalah *informasion overload* dengan teknik *web scraping* yang dikombinasikan dengan *text mining* sehingga dapat mengelompokkan artikel-artikel terkait hidroponik ke dalam beberapa kategori berdasarkan topik artikel secara otomatis. Dalam seleksi fitur, pendekatan yang digunakan adalah TF.
2017 – Alfian Futuhul Hadi, Dimas Bagus C. W., Moh. Hasan [19]	Text Mining pada Media Sosial Twitter. Studi Kasus: Masa Tenang Pilkada Dki 017 Putaran 2	Penelitian ini membandingkan kinerja metode-metode *unsupervised learning*, yaitu antara K-Means dan LDA dalam melakukan pemodelan topik. Tahap pra-pemrosesan data menggunakan pendekatan TF-IDF. Hasil akhir yang diperoleh yaitu pengelompokan yang seragam pada topik yang dibentuk dan keanggotaan yang merata apabila menggunakan LDA, sedangkan kecenderungan K-Means adalah menghasilkan 1 kelompok dengan anggota yang sangat dominan.
2012 – David M. Blei [20]	Probabilistic Topic Models	Penelitian ini menyimpulkan bahwa pemodelan topik berdasarkan teori probabilitas adalah algoritma yang tepat untuk menyediakan solusi statistik, khususnya untuk dokumen yang besar. Algoritma ini juga mendukung data tidak terstruktur, serta bersifat fleksibel untuk melakukan pemodelan data.
2003 – David M. Blei [21]	Latent Dirichlet Allocation	Penelitian ini mengusulkan model LDA yang dapat digunakan untuk pemodelan topik. LDA sebenarnya adalah model hirarki tiga tingkat dari Bayesian. Penelitian ini juga membandingkan LDA dengan model lainnya seperti LSI dan pLSI.

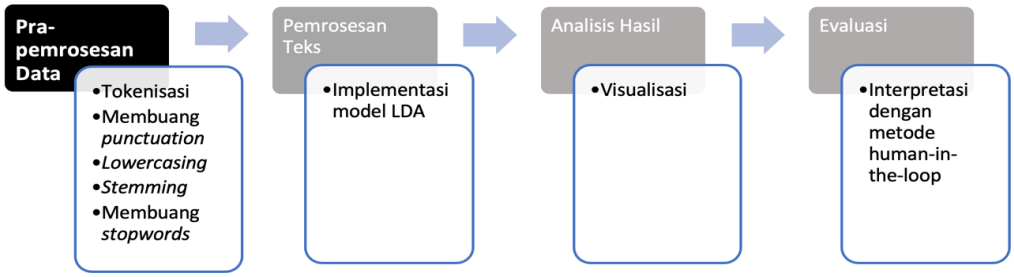
Tahun - Peneliti	Topik Penelitian - Judul	Kesimpulan
2017 - Matthew J. Denny, Arthur Spirling [22]	Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to Do About it	Menjelaskan secara detail penggunaan beragam cara pemrosesan data, khususnya untuk jenis unsupervised learning. Pemilihan urutan proses tersebut juga memiliki beragam kriteria, misalkan yang paling mendekati dengan penelitian ini adalah tahapan P-L-S-W.

2. Metodologi Penelitian

Sumber data yang digunakan seperti yang ditunjukkan pada Gambar 1 adalah data *chat-log* WA, yaitu jenis data *raw* yang tidak terstruktur dan harus melalui pra-pemrosesan terlebih dahulu. Data yang digunakan dalam Grup RBK adalah data pada bulan Februari 2017 s.d. April 2019, yaitu sebanyak +/- 17.000 data percakapan, sedangkan data dalam grup Dosen Trilogi adalah data pada bulan Agustus 2017 s.d April 2019, yaitu sebanyak +/- 6.000 data percakapan.

Pada penelitian ini akan dilakukan tahap yang pertama dari keseluruhan tahapan, yaitu (1) tahap pra-pemrosesan data yang mencakup pemilihan fitur/metode/langkah-langkah/tahapan pra-pemrosesan data, ekstraksi data raw chat-log WA, tokenisasi, membuang karakter yang tidak digunakan (*punctuation*), mengubah kata menjadi huruf kecil (*lowercasing*), *stemming*, membuang *stopwords*, dan visualisasi distribusi frekuensinya dengan menggunakan *bag-of-words*, kemudian di penelitian selanjutnya akan menuju ke (2) tahap pemrosesan teks, yaitu melakukan pemodelan topik dengan mengimplementasikan metode *Latent Dirichlet Allocation*, lalu (3) memberikan nama topik yang memiliki tingkat probabilitas yang tinggi dan membuat visualisasi klaster, dan pada tahap akhir akan dilakukan (4) interpretasi hasil pengelompokan topik. Metode interpretasi yang digunakan juga membutuhkan campur tangan manusia seperti membuat asumsi-asumsi seperti yang telah dilakukan oleh [23] melalui model evaluasi *human-in-the-loop*. Tahapan-tahapan yang dilalui atau metode penelitian mengacu pada [24] dengan melakukan perubahan sesuai dengan kebutuhan dalam penelitian ini, dan dapat dilihat pada Gambar 2.

Gambar 1. Data dari Grup *Whatsapp*



Gambar 2. Metodologi Penelitian

3. Hasil dan Pembahasan

Bagian ini merupakan implementasi pra-pemrosesan data dengan bahasa pemrograman *python*, yang mencakup (1) pemilihan fitur/metode/langkah-langkah/tahapan pra-pemrosesan data, (2) ekstraksi data *raw chat-log WA*, (3) tokenisasi, (4) membuang karakter yang tidak digunakan (*punctuation*), (5) *stemming*, (6) membuang *stopwords*, dan (7) visualisasi distribusi frekuensinya dengan menggunakan *bag-of-words*. Urutan proses mengikuti penelitian [b], namun proses *lowercasing* atau mengubah kata menjadi huruf kecil tidak dilakukan karena tidak berpengaruh terhadap proses pemodelan topik. Terdapat beberapa *dependecies* atau *library* di Python yang digunakan, yaitu nltk, numpy dan matplotlib, serta Sastrawi.

3.1 Tokenisasi

Tokenisasi digunakan untuk memecah kalimat menjadi potongan kata (*parsing*). Ada berbagai macam cara, misalnya menggunakan *library nltk* yang terdapat pada *python*. Kasus ini menggunakan *regular expression (regex)* untuk membuang informasi yang tidak dibutuhkan. Regex yang digunakan, yaitu:

```
(\[\d+\/\d+\/\d+)(\s)(\d+.\d+.\d+\])(\s\w+)*(:)
```

Gambar 3 dengan *highlight* berwarna biru merupakan hasil dari ekspresi *regrex*. Gambar 4 adalah kode yang digunakan dan gambar 5 adalah hasilnya.

Gambar 3. Regex WA (https://regexr.com)

```
import re

data_wa = "Selamat menunaikan ibadah haji ke tanah suci bu Irin,
semoga lancar dan dimudahkan dalam segala urusan dan selamat
kembali ke tanah air serta memperolah pahala Haji Mabrur....Aamiin
ya robbal alamin. Mohon didoakan juga kami yang dudini ya, wassalam
mualaikum.[12/08/17 06.30.14] Pak Dodi: aamiin... [12/08/17
06.31.24] Nurul Aisyah TRILOGI: Aamiin [12/08/17 06.57.26] Seveline
TRILOGI: Aaaaamiin [12/08/17 06.58.13] Heni TRILOGI: Amiin
[12/08/17 07.05.28] Yoni Atma TRILOGI: Aamiin... [12/08/17
07.30.28] Novita TRILOGI: Aamiin[12/08/17 08.19.20] Pak Kabul:
Aamiin [12/08/17 09.07.59] Bu Dhita: Hati hati bu irinn □"

hasil = re.sub(r'(\[\d+\/\d+\/\d+)(\s)(\d+.\d+.\d+\])(\s\w+)*(:)',
"", data_wa)
print ("Hasil token: ", hasil)
```

Gambar 4. Regex Menggunakan Python

```
Hasil token:  Selamat menunaikan ibadah haji ke tanah suci bu
Irin, semoga lancar dan dimudahkan dalam segala urusan dan
selamat kembali ke tanah air serta memperolah pahala Haji
Mabrur....Aamiin ya robbal alamin. Mohon didoakan juga kami yang
dudini ya, wassalam mualaikum. aamiin...  Aamiin  Aaaaamiin
Amiin  Aamiin...  Aamiin Aamiin  Hati hati bu irinn □
```

Gambar 5. Hasil Tokenisasi

3.2 Membuang Karakter yang Tidak Digunakan (*punctuation*)

Proses pembuangan karakter yang tidak digunakan (*punctuation*) mencakup angka, markup/html/tag, tanda baca, *emoticon* dan spesial karakter ($, %, &, etc). Pada proses ini juga dilakukan pembuangan angka (*numbering*) terutama untuk nomor telepon yang biasanya sering ada di dalam percakapan WA. Biasanya tag digunakan untuk analisis *trending topic*, tetapi untuk jenis pesan teks (*chat*), pengguna jarang menggunakan *tag*. Gambar 6 merupakan contoh dari penghapusan *emoticon* pada *whatsapp*, sedangkan gambar 7 merupakan hasil pemrosesannya.

```
import re

kalimat = u'Selamat menunaikan ibadah haji ke tanah suci bu Irin,
semoga lancar dan dimudahkan dalam segala urusan dan selamat
kembali ke tanah air serta memperolah pahala Haji
Mabrur....Aamiin ya robbal alamin. Mohon didoakan juga kami yang
dudini ya, wassalam mualaikum. aamiin...  Aamiin  Aaaaamiin
Amiin  Aamiin...  Aamiin Aamiin  Hati hati bu irinn □'
print ("Sebelum: ", kalimat)

emoji_pattern = re.compile("["u"\U0001F600-\U0001F64F"
                           "]+", flags=re.UNICODE)
print("Sesudah: ", emoji_pattern.sub(r'', kalimat))
```

Gambar 6. Kode Proses *Punctuation*

```
Sebelum: (diringkas) Hati hati bu irinn □
Sesudah: (diringkas) Hati hati bu irinn
```

Gambar 7. Hasil Proses *Punctuation*

3.3 *Stemming*

Proses *stemming* bertujuan untuk mencari akar kata atau kata dasar dengan cara menghilangkan imbuhan (*prefix/sufix*). Kode proses dan hasilnya dapat dilihat pada gambar 8 dan 9.

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory. create_stemmer()

kalimat = 'Selamat menunaikan ibadah haji ke tanah suci bu Irin,
semoga lancar dan dimudahkan dalam segala urusan dan selamat
kembali ke tanah air serta memperolah pahala Haji
Mabrur....Aamiin ya robbal alamin. Mohon didoakan juga kami yang
dudini ya, wassalam mualaikum. aamiin...  Aamiin  Aaaaamiin
Amiin  Aamiin...  Aamiin Aamiin  Hati hati bu irinn'

hasil = stemmer.stem(kalimat)

print(hasil)
```

Gambar 8. Kode Proses *Stemming*

```
selamat tunai ibadah haji ke tanah suci bu irin moga lancar
dan mudah dalam segala urus dan selamat kembali ke tanah air
serta olah pahala haji mabrur aamiin ya robbal alamin mohon
doa juga kami yang dudini ya wassalam mualaikum aamiin aamiin
aaaaamiin amiin aamiin aamiin aamiin hati hati bu irinn
```

Gambar 9. Hasil Proses *Stemming*

3.4 Membuang *Stopwords*

Proses pembuangan *stopwords* adalah membuang kata sambung dan beberapa kata yang tidak memiliki arti seperti yang ditunjukkan pada gambar 10 dan 11.

```
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory, StopWordRemover, ArrayDictionary

kalimat = 'selamat tunai ibadah haji ke tanah suci bu irin moga
lancar dan mudah dalam segala urus dan selamat kembali ke tanah
air serta olah pahala haji mabrur aamiin ya robbal alamin mohon
doa juga kami yang dudini ya wassalam mualaikum aamiin aamiin
aaaaamiin amiin aamiin aamiin aamiin hati hati bu irinn'

stop_factory = StopWordRemoverFactory().get_stop_words()
more_stopword = ['bu']

data = stop_factory + more_stopword

dictionary = ArrayDictionary(data)
str = StopWordRemover(dictionary)

print(str.remove(kalimat))
```

Gambar 10. Hasil Proses Membuang *Stopwords*

```
selamat tunai ibadah haji tanah suci irin moga lancar mudah
segala urus selamat ke tanah air olah pahala haji mabrur
aamiin robbal alamin mohon doa kami dudini wassalam mualaikum
aamiin aamiin aaaaamiin amiin aamiin aamiin aamiin hati hati
irinn
```

Gambar 11. Hasil Pembuangan *Stopwords*

3.5 Visualisasi Distribusi Frekuensi

Hasil distribusi kemunculan kata ditunjukkan pada gambar 12, 13, 14 dan grafik pada gambar 15.

```
from nltk.tokenize import word_tokenize
from nltk.probability import FreqDist
kalimat = "selamat tunai ibadah haji tanah suci irin moga lancar
mudah segala urus selamat ke tanah air olah pahala haji mabrur
aamiin robbal alamin mohon doa kami dudini wassalam mualaikum
aamiin aamiin aaaaamiin amiin aamiin aamiin aamiin hati hati
irinn"
kalimat =
kalimat.translate(str.maketrans('','',string.punctuation)).lower()

tokens = nltk.tokenize.word_tokenize(kalimat)
hasil = nltk.FreqDist(tokens)
print(hasil.most_common())
```

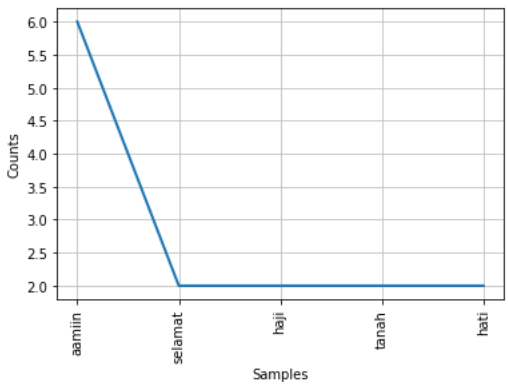
Gambar 12. Kode Frekuensi Kata yang Muncul

```
[('aamiin', 6), ('selamat', 2), ('haji', 2), ('tanah', 2),
('hati', 2), ('tunai', 1), ('ibadah', 1), ('suci', 1),
('irin', 1), ('moga', 1), ('lancar', 1), ('mudah', 1),
('segala', 1), ('urus', 1), ('ke', 1), ('air', 1), ('olah',
1), ('pahala', 1), ('mabrur', 1), ('robbal', 1), ('alamin',
1), ('mohon', 1), ('doa', 1), ('kami', 1), ('dudini', 1),
('wassalam', 1), ('mualaikum', 1), ('aaaaamiin', 1),
```

Gambar 13. Hasil Frekuensi Kata yang Muncul

```
import matplotlib.pyplot as plt
hasil.plot(5,cumulative=False)
plt.show()
```

Gambar 14. Matplotlib



Gambar 15. Grafik 5 Kata yang Paling Banyak Muncul

4. Kesimpulan

Pada penelitian awal ini, telah dilakukan studi pendahuluan dan analisis terkait penggunaan LDA untuk pemodelan topik pada kasus Grup WhatsApp. Tahap pra-pemrosesan yang telah dilakukan memiliki beberapa kendala, seperti penggunaan bahasa yang tidak baku dalam *chat*, *regex* yang harus dilakukan lebih dari 1 kali terutama untuk memisahkan *date-sender* dengan isi pesan, khususnya jika *sender* berupa nomor telepon (bukan nama). Namun hal tersebut tidak menjadi masalah karena proses *parsing* dapat dilakukan dengan cara membuang angka (*numbering*). Selain itu, kendala yang ditemui adalah Anaconda/Jupyter tidak dapat melakukan pemrosesan terhadap data yang besar. Tahap penelitian selanjutnya adalah mengimplementasikan algoritma LDA dan tampilan antar-muka dengan *input csv* WhatsApp dan keluaran berupa topik yang paling banyak muncul dalam grup tersebut.

5. References

[1] Statista. Most Popular Mobile Messaging Apps Worldwide As Of July 2018, Based On Number Of Monthly Active Users (In Millions). Sumber: https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/

[2] Statista. WhatsApp Has 600m Users Now. Sumber: https://www.statista.com/chart/2614/monthly-active-whatsapp-users-worldwide/

[3] We are Social. Global Digital Report 2018. Sumber: https://digitalreport.wearesocial.com

[4] Tanjum Kamboj dan Manoj Dayal. 2015. Usage of Instant Messaging Application on Smartphones among Youths: A study of Uses and Gratification of WhatsApp. URL: https://www.academia.edu/28203254/WhatsApp_Research_Paper.pdf

[5] Statista. To how many WhatsApp groups do you think to belong to and to how many do you actually belong to?. Sumber: https://www.statista.com/statistics/664440/number-of-whatsapp-groups-among-millennials-italy/

[6] Forbes. How To Run A Successful WhatsApp Group. Sumber: https://www.forbes.com/sites/paularmstrongtech/2018/04/29/how-to-run-a-successful-whatsapp-group/#66e050476364

[7] WhatsAnalyzer. Sumber: https://whatsanalyzer.informatik.uni-wuerzburg.de

[8] ChatVisualizer. Sumber: https://chatvisualizer.com

[9] Pragna Debnath, Saniul Haque, Somprakash Bandyopandhyay, Siuli Roy. 2016. Post-disaster Situational Analysis from WhatsApp Group Chats of Emergency Response Providers. Proeedings of the ISCRAM 2016 Conference. Brazil: Rio de Janeiro.

[10] R. Jayaparvathy . 2016. Analysis on Social Media Addiction using Data Mining Technique. International Journal of Computer Applications. Vol. 139. No. 7. Hal: 23-26.

[11] Avi Rosenfeld, Sigan Sina, David Sarne, Or Avidov, Sarit Kraus. 2016. WhatsApp Usage Patterns and Prediction Models. ICWSM/IUSSP Workshop on Social Media and Demographic Research.

[12] C. Premalatha, S. Jansi Rani. 2017. Sentimental Analysis of WhatsApp Data Using Data Analytics Techniques. Journal of Data Mining and Management. Vol 2. Issue 3. Hal: 1-6.

[13] Abhishek Soni. Introduction to Text Mining in WhatsApp Chats using Python-Part 1, https://www.zeolearn.com/magazine/introduction-to-text-mining-in-whatsapp-chats-using-python-part-1

[14] Amjad Karim. Whatsapp-(ening!) Text Analytics with a WhatsApp message Log. https://d-science.com/articles/whatsapp-ening-text-analytics-with-a-whatsapp-message-log/

[15] Yashodhan Joglekar. Analysing WhatsApp messages with Python and Tableau, https://databulary.net/2016/07/04/analysing-whatsapp-messages/

[16] Sanchita Patil. 2016. WhatsApp Group Data analysis with R. International Journal of Computer Applications. Vol. 154. No. 4. Hal: 31-36.

[17] I Made Kusnanta Bramantya Putra, Renny Pradina Kusumawardani . 2017. Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation* (*LDA*). Jurnal Teknik ITS. Vol. 6. No. 2. Hal: 11-16.

[18] Agung Priyanto, Muhammad Rifqi Ma'arif. 2018. Implementasi Web Scraping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi Laman Web Tentang Hidroponik. Indonesia Journal of Information Systems. Vol. 1. No. 1. Hal: 25-33.

[19] Alfian Futuhul Hadi, Dimas Bagus C. W., Moh. Hasan. 2017. Text Mining pada Media Sosial Twitter. Studi Kasus: Masa Tenang Pilkada Dki 017 Putaran 2. Seminar Nasional Matematika dan Aplikasinya. Surabaya: Universitas Airlangga.

[20] David M. Blei. 2012. Probabilistic Topic Model. Communications of the ACM. Vol. 55. No. 4. Hal: 77-84.

[21] David M. Blei. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. Vol. 3. Hal: 993-1022.

[22] Matthew J. Denny dan Arthur Spirling. 2017. Text Preprocessing for Unsupervised Learning: Why it Matters, When it Misleads, and What to Do About it. https://doi.org/10.7910/DVN/XRR0HM, Harvard Dataverse, V1.

[23] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, dan David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. Proceedings of the 22nd

International Conference on Neural Information Processing Systems. Hal: 288-296

[24] Zhou Tong dan Haiyi Zhang. 2016. A Text Mining Research Based on LDA Topic Modelling. Computer Science and Information Technologi. Hal: 201-210.