Analisis Trending Topik untuk Percakapan Media Sosial dengan Menggunakan Topic Modelling Berbasis Algoritme LDA

Ahmad Syaifuddin, *Teknologi Informasi iSTTS*, Reddy Alexandro Harianto, *Teknik Informatika iSTTS*, dan Joan Santoso, *Teknik Informatika iSTTS*

***Abstrak*— Aplikasi *WhatsApp* merupakan salah satu aplikasi *chatting* yang sangat populer terutama di Indonesia. *WhatsApp* mempunyai data unik karena memiliki pola pesan dan topik yang beragam dan sangat cepat berubah, sehingga untuk mengidentifikasi suatu topik dari kumpulan pesan tersebut sangat sulit dan menghabiskan banyak waktu jika dilakukan secara manual. Salah satu cara untuk mendapatkan informasi tersirat dari media sosial tersebut yaitu dengan melakukan pemodelan topik. Penelitian ini dilakukan untuk menganalisis penerapan metode LDA *(Latent Dirichlet Allocation)* dalam mengidentifikasi topik apa saja yang sedang dibahas pada grup *WhatsApp* di Universitas Islam Majapahit serta melakukan eksperimen pemodelan topik dengan menambahkan atribut waktu dalam penyusunan dokumen. Penelitian ini menghasilkan model topic dan nilai evaluasi *f-measure* dari model topik berdasarkan uji coba yang dilakukan. Metode LDA dipilih untuk melakukan pemodelan topik dengan memanfaatkan library LDA pada python serta menerapkan standar *text-preprocessing* dan menambahkan *slang words removal* untuk menangani kata tidak baku dan singkatan pada *chat logs*. Pengujian model topik dilakukan dengan uji *human in the loop* menggunakan *word instrusion task* kepada pakar Bahasa Indonesia. Hasil evaluasi LDA didapatkan hasil percobaan terbaik dengan mengubah dokumen menjadi 10 menit dan menggabungkan dengan *reply chat* pada percakapan grup *WhatsApp* merupakan salah satu cara dalam meningkatkan hasil pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA), didapatkan nilai *precision* sebesar 0.9294, nilai *recall* sebesar 0.7900 dan nilai *f-measure* sebesar 0.8541.**

***Kata Kunci*—Topic Modelling, Latent Dirichlet Allocation, LDA, Media Sosial.**

I. PENDAHULUAN

WhatsApp merupakan alat komunikasi yang bisa menimbulkan tantangan unik dalam pengelolaannya. Misalnya pada sebuah grup, banyak anggota yang cenderung mengabaikan kiriman atau tidak merespon pesan [1].

Ahmad Syaifuddin, Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: syaifuddin.skm@gmail.com)

Reddy Alexandro Harianto, Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: reddy@stts.edu)

Joan Santoso, Teknologi Informasi Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, Jawa Timur, Indonesia (e-mail: joan@stts.edu)

atau membalas postingan [2]. Karena sifat datanya yang cepat sekali berubah dan pengguna tidak sepenuhnya *online* atau mematau grup *WhatsApp* membuat pengguna kesulitan dalam mendapatkan informasi topik apa yang sedang dibahas pada grup *WhatsApp*. Padahal, kumpulan pesan tersebut merupakan sumber data yang sangat perpotensi untuk memberikan informasi yang penting bagi anggota lain.

Penelitian tentang pemodelan topik telah banyak dilakukan, banyak diantara penelitian menggunakan data dari *Twitter*, *Facebook*, *Online News*, *Blog*, jurnal dan lain sebagainya, namun untuk aplikasi *mobile chatting* khususnya *WhatsApp* belum banyak ditemukan topik penelitian di bidang *topic* modelling [3]. Padahal manfaat yang didapatkan dari data tersebut sangatlah banyak, misalnya untuk grup di Perguruan Tinggi. Salah satu fungsi utama grup di Perguruan Tinggi adalah untuk memudahkan komunikasi, penyampaian informasi yang penting dari pimpinan, keluhan, dan hal lain yang bersifat urgen.

Berdasarkan latar belakang tersebut, penelitian ini akan menawarkan solusi dalam melakukan analisis pemodelan topik menggunakan metode *Latent Dirichlet Allocation* pada grup *WhatsApp*. Analisis pemodelan topik digunakan untuk mengetahui tren topik yang muncul pada obrolan grup, sehingga memudahkan dalam mengetahui informasi apa yang sedang dibahas pada grup tersebut. Sebagai bagian dari *natural language programming*, maka langkah yang akan diimplementasikan mengikuti tahapan penelitian dalam ranah *text mining* secara umum, dengan menambahkan atribut reply chat yang dimiliki oleh WhatsApp dalam melakukan *pre-processing*.

Kontribusi paper ini adalah:

1. Analisis trending topik dengan LDA pada time frame tertentu yang dilakukan di Bahasa Indonesia.
2. Visualisasi trending topik dengan LDA untuk menggunakan atribut waktu dan reply bertingkat.

II. PENELITIAN PENJUNJANG

Guixian Xu [4] melakukan penelitian dalam melakukan ekstraksi topik dari teks berita menggunakan model Latent Dirichlet Allocation (LDA) dan metode Gibbs sampling untuk melakukan spekulasi parameter. Kemudian membandingkan dengan metode K-Mean dalam pendeteksian topik. Jumlah data yang digunakan berasal dari berita berlabel yang diterbitkan oleh Sogou Lab sebanyak 3.000 laporan berita. Hasilnya metode *LDA* lebih andal

daripada model *k-means clustering* untuk penentuan topik yang potensial.

YaJun Du [5] melakukan penelitian dalam melakukan mengestrak dan melacak topik pada *micro-blog* menggunakan improfikasi dari metode *Latent Dirichlet Allocation* (LDA). Penelitian ini menggunakan lima fitur unik blog mikro untuk mendorong distribusi probabilitas gabungan dari semua kata dan topik, dan meningkatkan LDA ke dalam model ekstraksi topik bernama MF-LDA (*Micro-blog features Latent Dirichlet Allocation*). Data yang digunakan dalam eksperimen model MF-LDA yaitu sekitar 500.000 data posting pada blog Mikro Sina yang dikumpulkan dari Mei 2016 hingga Desember 2016. Sedangkan pada pengujian hot topic mengambil sekitar 300.000 posting mikro-blog yang terkait dengan hampir 100 topik hangat dari platform Sina Micro-blog. Hasil percobaan menunjukkan bahwa model MF-LDA yang diusulkan berkinerja lebih baik dengan *perplexity* yang lebih rendah dan CR yang lebih tinggi daripada model LDA.

Dongjin Yu [6] melakukan penelitian tentang pemodelan topik hirarki data twitter untuk pemrosesan analitik online dengan mengusulkan model topik yang disebut twitter hierarchical latent Dirichlet alokasi (thLDA), dengan membandingkan dengan metode LDA biasa dan metode hLDA. Eksperimen dilakukan pada sejumlah besar data Twitter yang dikumpulkan melalui REST API Twitter. diperoleh 10.160.317 percakapan dari 6.907 tweet. Hasil penelitian didapatkan skor PMI pada model thLDA sedikit lebih tinggi daripada dua metode lainnya.

Ajai Gaur [7] melakukan penelitian tentang pemanfaatan LDA dalam *topic modelling* untuk identifikasi artikel yang diterbitkan oleh *Organization Research Methods* (ORM) sejak pertama kali terbit yaitu identifikasi kemiripan setiap topik pada seluruh artikel yang berasumsi bahwa topik yang muncul akan bervariasi tergantung pada latar belakang disiplin penulis. Hasil penelitiannya didapatkan sebanyak 15 topik yang sangat mirip dari 421 artikel ORM.

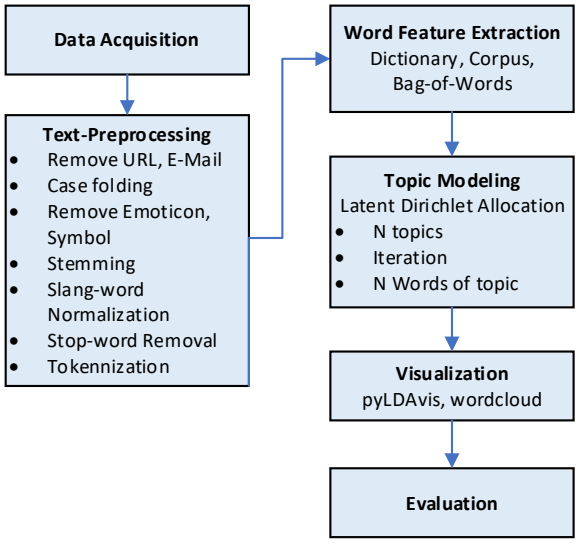
Terkait pelacakan topik, penelitian Jui-Feng Yeh [8] tentang pendeteksian dan pelacakan topik dalam percakapan lisan model *Conceptual Dynamic Latent Dirichlet Allocation* (CDLDA). Dataset berasal dari Chia Yi Chinese Conversation Dialogue Corpus (CYCCDC) yang diseleksi secara random didapatkan 769 dialog. Hasilnya didapatkan bahwa model CDLDA dapat secara dinamis mengekstraksi dan melacak topik dibandingkan dengan LDA tradisional. Basri [9] dalam penelitiannya yaitu klasterisasi dengan algoritma K-Means untuk identifikasi topik informasi publik media sosial di Kota Surabaya.

Khusus untuk grup WhatsApp, Cendana [10] dengan penelitiannya tentang pra-pemrosesan teks pada grup WhatsApp untuk pemodelan topik, pra-proses yang dilakukan yaitu tokenisasi, *filtering* dan *stemming*. Hasil penelitian didapatkan pada tahap pra-proses ditemukan kendala seperti, penggunakan bahasa tidak baku dalam chat, tidak dapat melakukan pemrosesan terhadap data yang besar. Penelitian ini tidak dilakukan normalisasi kata tidak baku pada tahap pra-proses. Penelitian Rosenfeld [11] tentang *WhatsApp Usage Patterns and Prediction Models*. Sumber datanya yaitu 4 juta pesan dari 100 pengguna dengan melihat kebiasaan dan perilaku pengiriman pesan berdasarkan *gender* dan usia. Model prediksi yang digunakan adalah algoritma C.45 dan Bayesian networks.

Penelitian Premalatha [12] dengan melakukan analisis sentimen dengan metode *Analytical Sandbox* terhadap pesan yang terdapat dalam grup WA dan visualisasinya sehingga didapatkan hasil opini positif dan negatif yang diilustrasikan dalam bentuk emotikon (*anger*, *fear*, *disgust*, *anticipation*, *joy*, *sadness*, *surprise*, *trust* dan *negative*). Sanchita Patil [13] melakukan implementasi bahasa pemrograman R untuk prediksi level kecanduan dari pengguna grup WA berdasarkan umur dan *gender*.

III. METODOLOGI

Alur sistem penelitian menggunakan langkah standar pada analisis *text mining* yaitu standar *pre-processing* ditambah dengan *remove URL-email*, *normalization slang word*, *stemming*, *remove stopword* dan *processing* menggunakan LDA, ditunjukan pada gambar 1.



GAMBAR 1.
ALUR SISTEM

Pada langkah pertama, akuisisi data yaitu dengan mengunduh *chat logs* pada *WhatsApp Web*, kemudian langkah kedua dilakukan *text-preprocessing*, selanjutnya melakukan ekstraksi fitur dengan membuat *dictionary*, *corpus* dan *bag-of-word*. Selanjutnya data diolah dengan algoritma LDA dan divisualisasi menggunakan library pyLDAvis dan wordcloud. Langkah terakhir yaitu pengujian menggunakan uji koherensi topik dan evaluasi *f-measure*.

A. Pengumpulan Data

Data didapatkan dari hasil *download chat* menggunakan fitur *WhatsApp Web* menggunakan aplikasi *Google Chrome*, kemudian ditambahkan *extension* atau plugin *Backup WhatsApp Chats*. Ekstensi *Backup WhatsApp Chats* dapat diunduh secara gratis melalui *Chrome Web Store*. Penulis menggunakan Ekstensi *Backup WhatsApp Chats* dikarenakan entitas yang dihasilkan lebih banyak bila dibandingkan dengan ekspor *chats* menggunakan fitur *export* yang terdapat pada *WhatsApp mobile*. Berikut merupakan tampilan dari percakapan yang ada pada grup *WhatsApp*. Sumber data yang dianalisa adalah data pesan yang dikirim oleh pengguna grup

WhatsApp Dosen Universitas Islam Majapahit pada bulan Desember tahun 2019. Data grup tersebut dipilih karena dianggap mempunyai kemiripan pada topik ataupun konten yang disampaikan, serta sudah mewakili civitas akademika di Universitas Islam Majapahit. Data yang digunakan adalah data teks saja, tidak termasuk gambar, video, emoticon dan file lainnya.

Input LDA berupa corpus yang terdiri dari beberapa dokumen, pada penelitian ini dokumen diambil berdasarkan pesan yang dikirim oleh user (1 pesan = 1 dokumen). Entitas MessageBody merupakan entitas utama dalam pembentukan dokumen, yaitu 1 isian pesan pada kolom MessageBody mewakili 1 dokumen, dalam hal ini UserName tidak dilibatkan dalam pembentukan dokumen. Pada eksperimen lebih lanjut, peneliti mengubah dokumen normal (1 pesan = 1 dokumen) berdasarkan reply chat dan kumpulan pesan dalam interval 10 menit.

B. *Preprocessing*

Langkah pra-proses data mencakup beberapa langkah utama pengerjaan yakni membersihkan data, *stemming* dan *stoword removal*.

1. Pembersihan data dilakukan untuk menghilangkan penulisan huruf besar menjadi huruf kecil, menghilangkan adanya karakter yang tidak diperlukan, menghapus token yang hanya terdiri dari angka. Selain itu, karena data yang diproses berasal dari pesan sosial media, maka perlu penanganan khusus seperti menghapus URL dan alamat email, menghapus *emoticon*, menghapus media gambar, video dan menghapus bilangan angka.
2. Normalisasi kata tidak baku, pesan pada whataspp sering kali diketik menggunakan kata slang atau singkatan, sehingga perlu melakukan pengumpulan kata tidak baku dan baku untuk digunakan sebagai *base dictionary*.
3. *Stemming* dilakukan untuk mengurangi kata-kata infleksi dalam Bahasa Indonesia menjadi bentuk dasarnya. Tahap *stemming* menggunakan library *sastrawi* dengan fungsi stem.
4. *Stopword removal* mengacu pada susunan *stopword* yang memanfaatkan *stopword removal indonesian* dari library nltk, dan *slang stopword* yang didapatkan peneliti dari temuan hasil ekspor data.
5. Menghapus dokumen yang mempunyai kata kurang dari dua. Dokumen yang hanya mempunyai satu kata penyusun dianggap tidak mempunyai arti dalam dokumen, sehingga data tersebut harus dihapus dan tidak dipakai sebagai data proses.

Pra-proses untuk mendapatkan dokumen reply chat, dilakukan dengan mengolah file csv hasil download dari WhatsApp Web menggunakan aplikasi *Libre Office* untuk menggabungkan isi pesan yang termasuk replay chat dan pesan yang dikirim dalam durasi 10 menit menjadi 1 buah dokumen. Urutan aturan dalam menggabungkan pesan menjadi dokumen sebagai berikut:

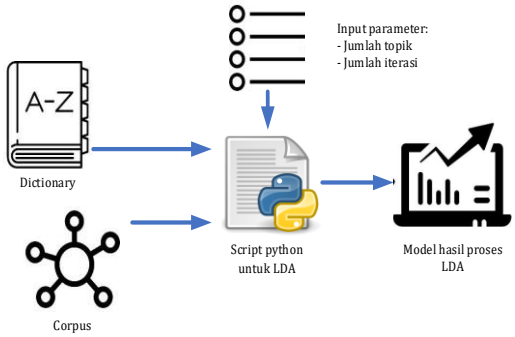
1. Pertama, pesan yang merupakan balasan dari pesan sebelumnya digabungkan terlebih dahulu
2. Kedua, pesan yang dikirim dalam interval waktu 10 menit digabungkan, dimulai dari menit ke 0-10, kemudian menit ke 11-20 dan seterusnya.
3. Ketiga, pesan yang sudah digabungkan pada tahap pertama dapat digabungkan lagi pada tahap kedua, jika pesan tersebut masuk dalam interval waktu tertentu.
4. Menyimpan hasil penggabungan pesan pada kolom bernama “MergedMessage”.

GAMBAR 2.
ALUR MEMPERSIAPKAN DATA

C. *Latent Dirichlet Allocation (LDA)*

LDA (*Latent Dirichlet Allocation*) adalah model probabilistik dalam *topic modelling* pada data teks untuk mendapatkan informasi berupa model topik. Model topik LDA dapat mewujudkan dimensi reduksi representasi teks dalam ruang semantik, dan memodelkan teks dengan probabilitas kosa kata, yang meringankan masalah data sampai batas tertentu [4].

Proses LDA diawali dengan konversi dokumen ke dalam bentuk *dictionary* dan mengkonversi *dictionary* ke dalam bentuk matriks dokumen atau *corpus*, kemudian membentuk model topik dengan menggunakan algoritma LDA. Proses algoritma LDA membutuhkan parameter input yang dalam hal penelitian ini yang digunakan adalah jumlah topik sebanyak 10, jumlah iterasi 50 dan jumlah kata dalam satu topik yaitu 10.



GAMBAR 3.
ALUR TOPIC MODELLING DENGAN LDA

Algoritma LDA terdiri dari proses inisialisasi, proses iterasi dan pengambilan sampel, proses membaca parameter akhir [14].

1. Tahap inisialisasi merupakan proses untuk menentukan frekuensi kemunculan kata dari setiap kata pada setiap file teks. Proses ini dilakukan pada teks hasil *pre-processing* data. Proses inisialisasi dilakukan dengan langkah:

a. Menentukan index dari setiap kata pada dokumen
b. Menghitung frekuensi kemunculan setiap kata pada setiap dokumen menggunakan *Bag-of-Words (BoW)*
c. Menentukan topik setiap kata dengan random berdasarkan nilai frekuensi kemunculan kata (*z0*)

LDA membutuhkan nilai topik yang ditentukan terlebih dahulu. Selain itu, dalam algoritma LDA, tidak ada nilai awal yang diberikan untuk setiap kata dalam dokumen, sehingga menghasilkan setiap kata memiliki nilai ketidakpastian. Dalam penelitian ini, nilai (*zo*) diambil secara acak yang menjadi nilai awal untuk setiap kata dalam dokumen.

d. Menentukan matriks kata-topik dan dokumen-topik

e. Menghitung jumlah total dari distribusi kata-topik dan dokumen-topik, dan menyimpan hasil matriks. Distribusi kata, Wi, pada tiap topik, Zi, NW dilihat pada persamaan berikut.

$$NW = \begin{Bmatrix} NW_{w1,z1} & NW_{w1,z2} & \dots & NW_{w1,zk} \\ NW_{w2,z1} & NW_{w2,z2} & \dots & NW_{w2,z1} \\ \dots & \dots & \dots & \dots \\ NW_{wn,z1} & NW_{wn,z2} & \dots & NW_{wn,zk} \end{Bmatrix} \quad (1)$$

Persamaan (1), *Wn* adalah *term* ke-n pada vocab dan *Zk* adalah Topik ke-k sedangkan $NW_{w_n z_k}$ banyak *term* ke-n yang berlabel topik ke-k

Kemudian membuat matriks distribusi topik, Zi, pada tiap dokumen, di, ND

$$ND = \begin{Bmatrix} ND_{d1,z1} & ND_{d1,z2} & \dots & ND_{d1,zk} \\ ND_{d2,z1} & ND_{d2,z2} & \dots & ND_{w2,z1} \\ \dots & \dots & \dots & \dots \\ ND_{dn,z1} & ND_{dn,z2} & \dots & Nd_{dn,zk} \end{Bmatrix} \quad (2)$$

Persamaan (2), dimana d_m adalah dokumen ke-m, Z_k adalah topik ke-k dan $ND_{d_n z_k}$ adalah banyak label topik ke-k pada Dokumen ke-n

Pada akhir proses inisialisasi, dilakukan proses perhitungan jumlah setiap topik dalam dokumen (*nd*) dan jumlah setiap kata dalam topik (*nw)* yang akan digunakan dalam proses iterasi dan pengambilan sampel topik. Nilai total semua *nd* juga dihitung sebagai *sumnd* dan nilai total semua *nw* sebagai *sumnw*. Nilai dari *sumnkm* dan *sumnw* digunakan untuk mengurangi dan menambahkan nilai *nd dan nw* di sembarang perubahan topik yang terjadi pada setiap kata. Jumlah distribusi NW dan ND sebagai berikut:

$$NWSum_{zi} = \sum_{j=1}^{m} NW_{wj,zi} \quad \dots\dots\dots \quad (3)$$

$$NDSum_{di} = \sum_{j=1}^{k} ND_{di,zj} \quad \dots\dots\dots \quad (4)$$

Persamaan (3), $NWSum_{zi}$ adalah jumlah seluruh kata dalam setiap topik, $NDSum_{di}$ adalah jumlah seluruh topik dalam setiap dokumen, $NW_{wj,zi}$ adalah jumlah setiap kata dalam topik dan $ND_{di,zj}$ adalah setiap topik dalam dokumen

2. Tahap sampling topik merupakan proses untuk menentukan topik baru dari setiap kata pada setiap dokumen. Proses ini dilakukan pada teks hasil *pre-processing*. Proses sampling topik dilakukan dengan langkah:

a. Menghitung probabilitas kata pada topik

$$\emptyset_{ij} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta} \quad (5)$$

Persamaan (5), ϕij adalah probabilitas dari kata i untuk topik j, C_{ij}^{WT} adalah Jumlah kata i pada topik j, WT adalah kata-topik, η adalah nilai beta (merupakan parameter dirichlet), $\sum_{k=1}^{W} C_{kj}^{WT}$ adalah jumlah seluruh kata pada topik j, *k* adalah indek topik dan *W dalah* Jumlah seluruh kata pada dokumen.

Menghitung probabilitas dokumen pada topik

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \quad (6)$$

Persamaan (6), θdj adalah proporsi topik j dalam dokumen d, C_{dj}^{DT} adalah jumlah topik j pada dokumen d, DT adalah dokumen-topik, α adalah nilai alpha, $\sum_{k=1}^{T} C_{dk}^{DT}$ adalah jumlah seluruh topik pada dokumen d T adalah jumlah seluruh topik yang sudah ditentukan

b. Menentukan topik baru dari setiap kata dengan distribusi multinomial (posterior) berdasarkan nilai probabilitas kata tertinggi.

Parameter distribusi posterior dimana bobot tersebut didapatkan dari nilai distribusi probabilitas kata-topik dikali distribusi probabilitas topik-dokumen.

$$P(z_i = j | z_i, w_i, d_i) = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta} x \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha} \quad (7)$$

Persamaan (7), $P(z_i = j)$ adalah probabilitas posterior untuk kata *i* yang diberikan topik *j*, z_i adalah merupakan penugasan topik dari semua kata w_i adalah indeks kata ke *i* dan d_i adalah dokumen yang berisi kata

c. Menyimpan hasil distribusi posterior

d. Langkah-langkah ini dilakukan sebanyak n iterasi/pengulangan sampai mencapai kondisi konvergen.

3. Tahap perhitungan parameter final merupakan proses untuk menghitung jumlah dokumen untuk setiap topik dan jumlah kata untuk setiap topik berdasarkan matriks kata-topik dan dokumen-topik yang telah konvergen.

D. Evaluasi LDA

Evaluasi hasil model topik dilakukan dengan menggunakan perhitungan *precision*, *recall* dan *f-measure* yang ditampilan dalam matrik evaluasi [15]. *Precision* (P) merupakan nilai pembagian dari jumlah item relevan yang diperoleh terhadap jumlah seluruh item yang diperoleh, rumus mencari *precision* dapat dilihat dalam persamaan (1).

$$P(relevant\ |retrieved) = \frac{relevant\ items\ retrieved}{retrieved\ items} \quad (8)$$

Recall (R) merupakan nilai pembagian dari jumlah item relevan yang diperoleh terhadap jumlah seluruh item yang relevan, rumus mencari *recall* dapat dilihat dalam persamaan (2)

$$R(retrieved|relevant) = \frac{relevant\ items\ retrieved}{relevant\ items} \quad (9)$$

F-Measure (F) merupakan bobot rata-rata dari nilai P dan R, rumus mencari recall dapat dilihat dalam persamaan (3).

$$F = \frac{2PR}{P+R} \quad (10)$$

IV. Uji Coba Sistem

Pembentukan model LDA dilakukan dengan mengiplementasikan 2 kali uji coba, yaitu model dengan dokumen sesuai hasil unduh data *WhatsApp* (uji coba tahap 1), model dengan dokumen hasil unduh data *WhatsApp* dalam durasi 10 menit serta hasil gabungan reply chat (uji coba tahap 2). Uji coba LDA dilakukan dengan bahasa python menggunakan library gensim dan *word cloud* untuk visualisasi model topik.

A. *Uji Coba pada Dokumen Normal*

Jumlah data mentah yang dimuat sebagai sumber data masukan dapat dilihat pada tabel berikut. Data mentah yang dimuat selanjutnya dijadikan sumber data input. Tahap pra-proses cukup efektif dalam menyeleksi data, terlihat sebanyak 443 data yang terhapus setelah dilakukan pra-proses. Data yang telah melalui tahap pra-proses dapat dilihat pada tabel sebagai berikut:

Tabel 1.
Data uji coba pada dokumen normal

	Jumlah dokumen	Jumlah token
Sebelum pra-proses	773	5702
Setelah pra-proses	330	1614

Hasil uji coba tahap pertama yang dilakukan dengan menggunakan dokumen sesuai hasil ekspor data *WhatsApp* dapat ditampilkan bawa tabel dibawah, yaitu berupa topik, kata penyusun topik dan nilai probabilitas dari masing-masing kata, berikut cuplikan model topik percobaan pertama dapat dilihat pada tabel 2.

Tabel 2.
Hasil Model Topik pada Dokumen Normal

Topik	Model Topik
T1	pmb, 0.0419 + tim, 0.0196 + thp, 0.0189 + personil, 0.0113 + prodi, 0.0113 + sistem, 0.0113 + unim, 0.0086 + promosi, 0.0086 + bijak, 0.0086 + review, 0.0058
T2	data, 0.0242 + prodi, 0.0212 + dosen, 0.0152 + mahasiswa, 0.0122 + bimbing, 0.0122 + guru, 0.0122 + didik, 0.0092 + teliti, 0.0092 + seminar, 0.0062 + riset, 0.0062
T3	wisata, 0.0451 + mojokerto, 0.0307 + air, 0.0162 + panas, 0.0130 + kuasa, 0.0114 + padusan, 0.0098 + pipa, 0.0098 + banjir, 0.0098 + sungai, 0.0095 + cinta, 0.0081
T4	kirim, 0.0343 + mahasiswa, 0.0311 + pkm, 0.0186 + tulis, 0.0186 + proposal, 0.0160 + email, 0.0160 + terima, 0.0148 + dosen, 0.0106 + kampus, 0.0081 + judul, 0.0081
T5	ajar, 0.0258 + didik, 0.0221 + mendikbud, 0.0203 + bijak, 0.0148 + sekolah, 0.0148 + uji, 0.0129 + guru, 0.0129 + kualitas, 0.0111 + data, 0.0093 + presiden, 0.0093
T6	ipk, 0.0136 + indonesia, 0.0119 + kejar, 0.0119 + kerja, 0.0117 + didik, 0.0102 + riset, 0.0102 + bangsa, 0.0102 + sekolah, 0.0102 + peringkat, 0.0102 + urut, 0.0102
T7	allah, 0.1415 + dosa, 0.0864 + hati, 0.0589 + maha, 0.0425 + rezeki, 0.0320 + sembuh, 0.0209 + guru, 0.0209 + terima, 0.0150 + keluarga, 0.0149 + urus, 0.0142
T8	khotimah, 0.0296 + husnul, 0.0259 + rojiun, 0.0182 + lillahi, 0.0163 + amal, 0.0130 + almarhum, 0.0112 + kanker, 0.0112 + sembuh, 0.0094 + jantung, 0.0094 + allah, 0.0094
T9	listrik, 0.0174 + pln, 0.0161 + didik, 0.0134 + mati, 0.0134 + dosen, 0.0107 + mobil, 0.0094 + parkir, 0.0094 + lebaran, 0.0094 + usaha, 0.0094 + sepeda, 0.0081
T10	dosen, 0.0229 + insulin, 0.0216 + glukosa, 0.0173 + unim, 0.0155 + gula, 0.0145 + mahasiswa, 0.0141 + diabetes, 0.0116 + parkir, 0.0111 + manusia, 0.0101 + karbo, 0.0101

B. *Uji Coba Pada Dokumen Reply Chat*

Distribusi probabilitas kata dalam 10 topik yang tersusun dari 10 kata dalam masing-masing topik dari model dokumen sesuai hasil ekspor data *WhatsApp* serta gabungan dari reply chat, dapat dilihat pada tabel hasil pembentukan model LDA dengan dokumen hasil ekspor data *WhatsApp* dalam durasi 10 menit dan gabungan reply chat. Jumlah data mentah yang dimuat sebagai sumber data masukan dapat dilihat pada tabel berikut. Data mentah yang dimuat selanjutnya dijadikan sumber data input berikut cuplikan data yang digunakan pada uji coba kedua, dapat dilihat pada tabel 3.

Tabel 3.
Data Uji Coba pada Dokumen Reply Chat

	Jumlah dokumen	Jumlah token
Sebelum pra-proses	340	5702
Setelah pra-proses	205	1664

Hasil uji coba tahap kedua yang dilakukan dengan menggunakan dokumen hasil ekspor data *WhatsApp* berdasarkan interval waktu 10 menit dan reply chat dapat ditampilkan bawa tabel dibawah, yaitu berupa topik, kata penyusun topik dan nilai probabilitas dari masing-masing kata, berikut cuplikan model topik percobaan pertama dapat dilihat pada tabel 4.

TABEL 4.
HASIL MODEL TOPIK PADA DOKUMEN REPLY CHAT

Topik	Model Topik
T1	ajar, 0.0168 + didik, 0.0168 + mendikbud, 0.0168 + bijak, 0.0141 + sekolah, 0.0113 + menteri, 0.0099 + uji, 0.0099 + guru, 0.0099 + program, 0.0085 + kualitas, 0.0085
T2	data, 0.0168 + dosen, 0.0165 + update, 0.0155 + kerja, 0.0117 + sdm, 0.0116 + teliti, 0.0091 + bimbing, 0.0091 + riset, 0.0091 + prodi, 0.0078 + sosialisasi, 0.0078
T3	insulin, 0.0237 + glukosa, 0.0190 + didik, 0.0127 + badan, 0.0127 + manusia, 0.0111 + jenderal, 0.0111 + liver, 0.0111 + rutin, 0.0111 + karbo, 0.0111 + diabetes, 0.0096
T4	khotimah, 0.0311 + husnul, 0.0272 + lillahi, 0.0253 + rojiun, 0.0234 + allah, 0.0229 + inna, 0.0195 + umroh, 0.0118 + amal, 0.0118 + belasungkawa, 0.0118 + almarhum, 0.0118
T5	pmb, 0.0186 + tim, 0.0145 + kerja, 0.0145 + praktik, 0.0124 + mahasiswa, 0.0111 + thp, 0.0104 + unim, 0.0090 + dosen, 0.0088 + personil, 0.0083 + promosi, 0.0083
T6	klik, 0.0243 + unim, 0.0157 + nilai, 0.0147 + mahasiswa, 0.0138 + fe, 0.013 + jurnal, 0.0098 investasi, 0.0098 + ahli, 0.0098 + dosen, 0.0098 + ilmiah, 0.0092
T7	wisata, 0.0430 + mojokerto, 0.0292 + air, 0.0154 + panas, 0.0108 + kuasa, 0.0108 + banjir, 0.0093 + cinta, 0.0093 + kirim, 0.0093 + padusan, 0.0093 + pipa, 0.0093
T8	parkir, 0.0312 + mobil, 0.0240 + dosen, 0.0166 + sepeda, 0.0140 + pln, 0.0140 + fisip, 0.0106 + listrik, 0.0072 + padam, 0.0072 + nasib, 0.00720 + mahal, 0.0072
T9	ipk, 0.0142 + indonesia, 0.0106 + didik, 0.0106 + bangsa, 0.0106 + riset, 0.0106 + sekolah, 0.0106 + peringkat, 0.0106 + urut, 0.0106 + kerja, 0.0071 + nilai, 0.0071
T10	allah, 0.1442 + dosa, 0.0891 + hati, 0.0613 + maha, 0.0462 + rezeki, 0.0334 + guru, 0.0224 sembuh, 0.0223 + terima, 0.0144 + berkah, 0.0142 + keluarga, 0.0132

Word cloud dapat memberikan gambaran model topik yang terbentuk melalui proses LDA, kata dengan ukuran yang lebih besar menggambarkan kata dengan frekuensi kemunculan tertinggi pada model topik. Pada uji coba pertama, terlihat bahwa kata dosen, didik, mahasiswa dan guru menjadi kata dominan dengan frekuensi kemunculan tertinggi. Berikut *word cloud* pada uji coba pertama dapat dilihat pada gambar 4.

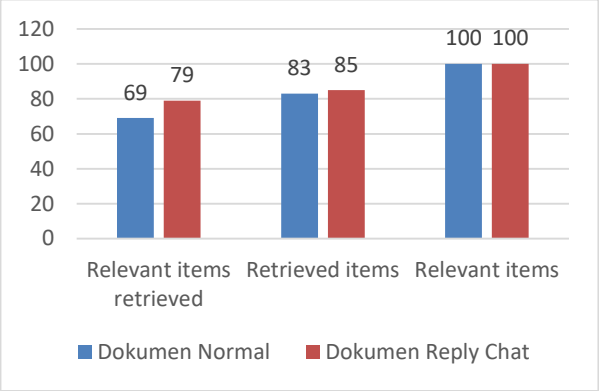
GAMBAR 4
WORD CLOUD MODEL LDA PADA DOKUMEN NORMAL

Word cloud pada uji coba kedua terlihat bahwa kata dosen, didik dan kerja menjadi kata dominan dengan frekuensi kemunculan tertinggi.

GAMBAR 5.
WORD CLOUD MODEL LDA PADA DOKUMEN REPLY CHAT

Evaluasi untuk menilai apakah topik-topik yang dihasilkan baik pada uji coba pertama dan uji coba kedua sudah sesuai dengan persepsi manusia, dilakukan uji koherensi topik dengan menggunakan metode *human in the loop*. Yaitu dengan memberikan kuesioner kepada seorang pakar untuk menulis model topik berdasarkan dataset yang sama. Kuesioner uji koherensi topik dilakukan sebanyak satu kali, dilakukan pada tanggal 30 Oktober 2020 secara *online* dengan seorang pakar yaitu dosen Program Studi Pendidikan Bahasa Indonesia. Pakar diminta untuk menuliskan 10 topik dan 10 kata-kata penyusun topik.

Evaluasi model topik dengan metode LDA menggunakan kuesioner *word instrusion task* terhadap seorang pakaer menghasilkan tiga parameter yang akan menjadi acuan dalam evaluasi. Tiga parameter yaitu, *relevant items*, *retrieved items* dan *relevant items retrieved*. *Relevant items* merupakan jumlah kata dalam model topik yang terbentuk melalui proses LDA yang diambil 10 kata dengan nilai probabilitas tertinggi. *Retrieved items* merupakan jumlah kata yang diambil dari pakar sebagai pembentuk model topik melalui kuesioner *word instrusion task*. Pakar disini adalah Dosen Program Studi Pendidikan dan Sastra Indonesia di Universitas Islam Majapahit. Sedangkan *relevant items retrieved* merupakan jumlah kata yang diseleksi oleh pakar dan sesuai dengan model topik yang didapatkan dari sistem. Berikut ini cuplikan dari hasi kuesioner yang sudah dilakukan rekapitulasi dalam gambar dibawah ini.



GAMBAR 6.
HASIL MODEL TOPIK PADA DOKUMEN REPLY CHAT

Evaluasi dengan matriks pengukuran *precision, recall dan f-measure*. Berikut matrik hasil evaluasi yang dilakukan.

TABEL 5.
MATRIK HASIL EVALUASI

Uji Coba	Dokumen normal	Dokumen Reply Chat
Precision	0.8313	0.9294
Recall	0.6900	0.7900
F-Measure	0.7541	0.8541

Berdasarkan hasil evaluasi diatas, didapatkan bahwa nilai *precision, recall* dan *f-measure* pada uji coba kedua (dengan mengubah dokumen menjadi 10 menit dan gabungan *reply chat*) memiliki nilai yang lebih tinggi bila dibandingkan dengan uji coba pertama yang telah dilakukan. Didapatkan nilai *precision* sebesar 0.9294, nilai *recall* sebesar 0.7900 dan nilai *f-measure* sebesar 0.8541. Hal ini menunjukkan bahwa mengubah dokumen menjadi 10 menit dan menggabungkan *reply chat* pada percakapan grup *WhatsApp* merupakan salah satu cara dalam meningkatkan hasil pemodelan topik menggunakan algoritma *Latent Dirichlet Allocation* (LDA). Hal ini karena anggota grup pada *WhatsApp* sebagian besar merespon pesan yang dikirimkan oleh anggota lain dalam jangka waktu yang cukup lama, sehingga meningkatkan dokumen dengan interval 10 menit dapat mengumpulkan percakapan dengan topik diskusi yang sama selain itu hasil *reply chat* juga sangat berpengaruh besar dalam pembentukan dokumen. Penelitian topik modelling pada grup *WhatsApp* menggunakan *Latent Dirichlet Allocation* (LDA) dapat dengan baik mengidentifikasi topik-topik yang sering dibicarakan anggota grup *WhatsApp* dengan kondisi mengubah dokumen menjadi 10 menit dan menggabungkan *reply chat*.

Grup *WhatsApp* cukup efektif guna mendukung kinerja dari civitas akademika di perguruan tinggi. Sesuai penelitian Andjani [16] penggunaan media komunikasi *WhatsApp* di instansi sangat baik dan dapat membantu dalam peningkatan efektivitas kinerja karyawan. Penelitian Rahartri [17] juga menjelaskan *WhatsApp* merupakan pengganti SMS yang praktis dan tepat waktu dalam mengirimkan pesan karena lebih simpel dan mudah digunakan pada layanan jasa informasi ilmiah di kawasan Puspiptek.

Uji coba pertama didapatkan hasil evaluasi *precision, recall* dan *f-measure* paling kecil dengan nilai *precision* sebesar 0.8313, nilai *recall* sebesar 0.6900 dan nilai *f-measure* sebesar 0.7541. Hal ini disebabkan karena respon anggota dalam membalas pesan dari anggota lain dengan durasi yang lama, sehingga topik yang dibicarakan tidak terkumpul menjadi satu dokumen. Selain itu, karena karakter dari pesan *WhatsApp* yang kebanyakan ditulis dengan kalimat yang pendek, sehingga pada saat tahap *pre-processing* kata penyusun dokumen ternormalisasi menjadi sedikit atau mungkin hilang, karena penulis tidak membuang dokumen yang hanya berisi satu kata.

V. KESIMPULAN

Latent Dirichlet Allocation (LDA) dapat dengan baik mengidentifikasi topik-topik yang sering dibicarakan anggota grup *WhatsApp*, dibuktikan dengan topik yang dihimpun dari pakar, menyatakan lebih dari 80% topik yang didapatkan dari sistem sesuai dengan model/kata penyusun topik. Eksperimen dengan menambahkan atribut waktu dalam penyusunan dokumen, didapatkan bahwa mengubah dokumen menjadi 10 menit dan menggabungkan *reply chat* didapatkan hasil yang lebih baik bila dibandingkan dengan dokumen normal yaitu sebanyak 79 kata relevan yang diterima oleh pakar. Hasil uji *f-measure* didapatkan bahwa dokumen dengan interval 10 menit dan menggabungkan *reply chat* mempunyai nilai yang tinggi bila dibandingkan dengan dokumen normal yaitu sebesar 0.8541.

Berdasarkan pengamatan penulis masih ada beberapa kata yang tidak bisa dilakukan *stemming* menggunakan *sastrawi python* yakni sebesar 4 dari 100 random kata, bagi peneliti selanjutnya dapat meningkatkan performa dalam *pre-processing* pada proses *stemming*, sehingga dapat meningkatkan nilai probabilitas kata dan efisiensi iterasi pada perhitungan LDA. Selain itu penggunaan LDA pada library gensim dalam pembentukan topik kurang konsisten, maka untuk penelitian lebih lanjut dapat menambahkan langkah khusus dalam penentuan z0 atau membuat dictionary dan corpus yang dapat meningkatkan stabilitas dalam membuat model topik, misalnya menggunakan n-gram atau lainnya.

DAFTAR PUSTAKA

[1] B. T. Shambhu, “WhatsApp: Unlocking The Goldmine.” New Delhi, India: Educreation Publishing, 2016.

[2] A. O. Afolaranmi, “Towards the Possibility of Internet Ministry as an Alternative Pastoral Ministry in Nigeria during the COVID-19 Pandemic,” *Int. J. Inf. Technol. Lang. Stud.*, vol. 4, no. 2, 2020.

[3] Z. Tong and H. Zhang, “A text mining research based on LDA topic modelling,” in *International Conference on Computer Science, Engineering and Information Technology*, 2016, pp. 201–210.

[4] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, “Research on topic detection and tracking for online news texts,” *IEEE access*, vol. 7, pp. 58407–58418, 2019.

[5] Y. Du, Y. Yi, X. Li, X. Chen, Y. Fan, and F. Su, “Extracting and tracking hot topics of micro-blogs based on improved Latent Dirichlet Allocation,” *Eng. Appl. Artif. Intell.*, vol. 87, p. 103279, 2020.

[6] D. Yu, D. Xu, D. Wang, and Z. Ni, “Hierarchical topic modeling of Twitter data for online analytical processing,” *IEEE Access*, vol. 7, pp. 12373–12385, 2019.

[7] A. Piepenbrink and A. S. Gaur, “Topic models as a novel approach to identify themes in content analysis,” in *Academy of Management Proceedings*, 2017, vol. 2017, no. 1, p. 11335.

[8] J.-F. Yeh, Y.-S. Tan, and C.-H. Lee, “Topic detection and tracking for conversational content by using conceptual dynamic latent Dirichlet allocation,” *Neurocomputing*, vol. 216, pp. 310–

318, 2016.
[9] M. H. Basri, “Identifikasi Topik Informasi Publik Media Sosial Di Kota Surabaya Berdasarkan Klasterisasi Teks Pada Twitter Dengan Menggunakan Algoritma K-means,” Institut Teknologi Sepuluh Nopember, 2015.
[10] M. Cendana and S. D. H. Permana, “Pra-Pemrosesan Teks Pada Grup Whatsapp Untuk Pemodelan Topik,” *J. Mantik Penusa*, vol. 3, no. 3, 2019.
[11] A. Rosenfeld, S. Sina, D. Sarne, O. Avidov, and S. Kraus, “WhatsApp usage patterns and prediction models,” 2016.
[12] C. Premalatha and S. J. Rani, “SENTIMENTAL ANALYSIS OF WHATSAPP DATA USING DATA ANALYTICS TECHNIQUES,” *J. Data Min. Manag. (e-ISSN 2456-9437)*, vol. 2, 2017, Accessed: Jul. 05, 2021. [Online]. Available: http://matjournals.in/index.php/JoDMM/article/view/1906.
[13] S. Patil, “WhatsApp Group Data Analysis with R,” *Int. J. Comput. Appl.*, vol. 154, no. 4, 2016, doi: 10.5120/ijca2016912116.
[14] P. M. Prihatini, I. Putra, I. A. D. Giriantari, and M. Sudarma, “Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents,” *Contemp. Eng. Sci.*, vol. 10, pp. 403–421, 2017.
[15] P. M. Prihatini, I. K. Suryawan, and I. N. Mandia, “Feature extraction for document text using Latent Dirichlet Allocation,” in *Journal of Physics: Conference Series*, 2018, vol. 953, no. 1, p. 12047.
[16] A. Anjani, I. A. Ratnamulyani, and A. A. Kusumadinata, “Penggunaan Media Komunikasi Whatsapp terhadap Efektivitas Kinerja Karyawan,” *J. Komun.*, vol. 4, no. 1, 2018.
[17] L. Rahartri, “‘ WHATSAPP’ MEDIA KOMUNIKASI EFEKTIF MASA KINI (STUDI KASUS PADA LAYANAN JASA INFORMASI ILMIAH DI KAWASAN PUSPIPTEK,” *VISI PUSTAKA Bul. Jar. Inf. Antar Perpust.*, vol. 21, no. 2, pp. 147–156, 2019.