INTERNAL Information System Journal
INTERNAL (Information System Journal)
Volume 7 No. 2 | Desember 2024 : 167-177
DOI: 10.32627
http://jurnal.masoemuniveristy.ac.id/index.php/internal
p-ISSN: 2621-9433
e-ISSN: 2656-0259

Permodelan Topik pada Layanan Akademik Perguruan Tinggi dengan Menggunakan N-Gram

Aldy Rialdy Atmadja[1], Muhammad Naufal Rahman[2], Wildan Budiawan Zulfikar[3]
[1,2,3] Teknik Informatika, UIN Sunan Gunung Djati Bandung, Indonesia
aldyrialdy@uinsgd.ac.id

Info Artikel
Sejarah artikel:
Diterima Desember 2024
Direvisi Desember 2024
Disetujui Desember 2024
Diterbitkan Desember 2024

ABSTRACT

Automation of generating information in academic services are expected to provide convenience in providing academic services to students. Relevant topics can be extracted from social media by calculating the frequency of words asked by social media users. The research focuses on generating question topics on academic services at universities. Topic extraction are obtained through data taken from Instagram social media, so that relevant topics are obtained to get information that is most frequently asked by the public. The N-Gram and Term Frequency are approach to extract the topic. The initial stages in this study include conducting Web Scraping taken from Instagram social media. In this study, text preprocessing was carried out in several stages, namely cleansing, casefolding, stopwords removal and tokenizing, and stemming. The N-Gram approach is carried out by comparing three types, namely unigrams, bigrams and trigrams. The results obtained in this study prove that the bigram produces relevant word pairs in determining academic service topics on social media. This approach produces word pairs that are relevant to academic service topics including graduation list, paying UKT, independent admission, SPANPTKIN and independent test.

Keywords : *Topic Extraction; Academic Services; N-Gram; Preprocessing; Term Frequency.*

ABSTRAK

Otomatisasi penyajian informasi dalam layanan akademik diharapkan dapat memberikan kemudahan dalam memberikan pelayanan akademik kepada mahasiswa. Topik-topik yang relevan dapat dilakukan dengan melakukan ekstraksi pada media sosial dengan menghitung banyaknya frekuensi kata yang ditanyakan oleh pengguna media sosial. Penelitian ini berfokus untuk menghasilkan topik-topik pertanyaan pada layanan akademik Perguruan Tinggi. Ekstraksi topik tersebut diambil dari media sosial Instagram untuk mendapatkan topik yang relevan dan sering ditanyakan oleh publik. Metode yang dipergunakan dalam melakukan ekstraksi topik ini yakni dengan menggunakan pendekatan N-Gram dan Term Frequency. Tahapan awal pada penelitian ini diantaranya yaitu melakukan Web Scraping yang diambil dari media sosial Instagram. Pada penelitian ini praproses teks dilakukan dengan beberapa tahapan yaitu cleansing, casefolding, stopwords removal dan tokenizing, dan stemming. Pendekatan N-Gram dilakukan dengan membandingkan tiga tipe yakni unigram, bigram dan trigram. Hasil yang didapatkan pada penelitian ini membuktikan bahwa pendekatan bigram menghasilkan pasangan kata yang relevan dalam penentuan topik layanan akademik pada media sosial. Pendekatan ini menghasilkan pasangan kata yang relevan dengan topik layanan akademik diantaranya daftar wisuda, bayar ukt, jalur mandiri, jalur spanptkin dan uji mandiri.

Kata Kunci : Ekstraksi Topik; Layanan Akademik; N-Gram; Praproses; Term Frequency.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

PENDAHULUAN

Saat ini, perkembangan teknologi berkembang pesat dengan melalui adanya internet. Penggunaan internet saat ini memudahkan seseorang dalam mengakses halaman *website*. Salah satu fungsi *website* diantaranya sebagai media informasi. *Website* digunakan dalam menyajikan informasi untuk pembacanya, diantaranya informasi terkait perguruan tinggi[1]. Saat ini informasi layanan akademik sudah sangat lengkap diakses melalui *website* ataupun media sosial. Namun terkadang sebagian besar siswa kesulitan dalam menemukan informasi yang diinginkannya karena informasi tersebut tidak tersedia maupun kesalahan dalam melakukan pencarian informasi.

Media sosial merupakan salah satu media komunikasi untuk mencari informasi yang ingin diperoleh oleh seseorang[2]. Pertanyaan yang diajukan dapat berupa komentar yang dibuat pada unggahan akun resmi. Instagram merupakan salah satu *platform* yang digunakan sebagai wadah media sosial untuk sekedar bertanya dan mencari informasi berupa gambar dan teks. Namun terdapat kendala jika media sosial tersebut dijadikan alat komunikasi sehingga perlu dijawab satu per satu dan memakan waktu yang lama. Sehingga diperlukan otomatisasi dalam penyajian informasi layanan akademik.

Otomatisasi penyajian informasi dalam layanan akademik diharapkan dapat memberikan informasi secara otomatis sesuai keinginan atau pertanyaan mahasiswa tanpa perlu menjawabnya satu per satu. Hal ini dapat memberikan pelayanan yang lebih baik kepada mahasiswa dan meningkatkan efisiensi dalam proses administrasi akademik di perguruan tinggi. Otomatisasi ini dapat dilakukan dengan mencari topik-topik yang relevan yang sering ditanyakan oleh orang berdasarkan frekuensi banyaknya kata yang ditanyakan oleh orang.

Beberapa penelitian yang relevan yang kaitannya dengan melakukan ekstraksi informasi diantaranya yaitu penelitian yang menggunakan *N-Gram* untuk analisis sentimen pemilihan kepala daerah Jakarta menggunakan Algoritma *Naïve Bayes*. Penelitian ini dilakukan dengan menggunakan algoritma *Naïve Bayes* untuk mengklasifikasikan pendapat menjadi positif atau negatif dengan menggunakan seleksi fitur Chi Square yang telah dilakukan N-gram sebelumnya. Tujuan dari penelitian ini adalah untuk melihat tingkat akurasi klasifikasi menggunakan algoritma Naïve Bayes dengan melakukan penggunaan fitur N-gram[3]. Penelitian lainnya yaitu yang dilakukan oleh Nur Lickha dan Reisa yang melakukan analisis sentimen terhadap gangguan mental pada media sosial Twitter. Penelitian tersebut menghasilkan akurasi optimal sebesar 90,13% dengan menggunakan Multinomial Naïve Bayes[4]. Selain itu, penelitian lainnya membahas tentang penerapan algoritma cosine similarity dan pembobotan TF-IDF dalam penerimaan mahasiswa baru. Tujuan penelitian ini yaitu membangun sistem penjawab FAQ dengan menerapkan pembobotan TF-IDF *(Term Frequency-Inverse Document Frequency)* dan algoritma *cosine similarity*. Penelitian ini menggunakan 7 buah sampel data dari keseluruhan data FAQ yang didapat dari wawancara dengan pihak penerimaan mahasiswa baru Universitas Bumigora. Data sampel yang digunakan akan melalui praproses, pembobotan TF-IDF, dan metode *cosine similarity* untuk menentukan

p-ISSN: 2528-1682
e-ISSN: 2527-9165

tingkat keasaman. Penggunaan metode tersebut menghasilkan tingkat akurasi hingga mencapai 64,28%[5].

Penelitian ini merupakan penelitian awal yang dilakukan untuk melakukan ekstraksi informasi terhadap topik-topik pertanyaan yang didapatkan dari media sosial Instagram. Metode *N-Gram* digunakan untuk mengambil n buah kata dari rangkaian kata yang dibaca terus menerus mulai dari teks sumber hingga akhir dokumen[6], sedangkan *Term Frequency* adalah kuantitas istilah yang sering muncul dalam suatu dokumen[7]. Metode N-Gram dapat digunakan untuk mencari bahan yang digunakan dalam menghasilkan pertanyaan dan *Term Frequency* dapat digunakan untuk menghitung banyaknya data yang mengandung *N-Gram*[8]. Adapun penelitian ini akan berfokus pada membandingkan metode ekstraksi dengan menggunakan *Unigram, Bigram* dan *Trigram* sehingga dapat didapatkan topik yang relevan dengan pertanyaan yang ada. Selain itu, penelitian ini sebagai landasan terhadap penelitian utama yaitu menghasilkan pertanyaan dalam bentuk *Frequently Asked Questions*.

METODE

Metode yang dipergunakan dalam melakukan ekstraksi topik ini yakni dengan menggunakan pendekatan *N-Gram* dan *Term Frequency*. Tahapan yang dilakukan diantaranya yaitu melakukan *web scraping* yang bersumber dari media sosial Instagram. Tahapan praproses teks yang digunakan dalam penelitian ini diantaranya yaitu *cleansing, casefolding, stopwords removal* dan *tokenizing*, dan *stemming*[9]. Pada penelitian ini menggunakan pendekatan diantaranya menggunakan *unigram*, *bigram* dan *trigram*. Berikut ini merupakan gambaran dari langkah-langkah yang dilakukan pada penelitian ini yang tertera pada gambar 1.

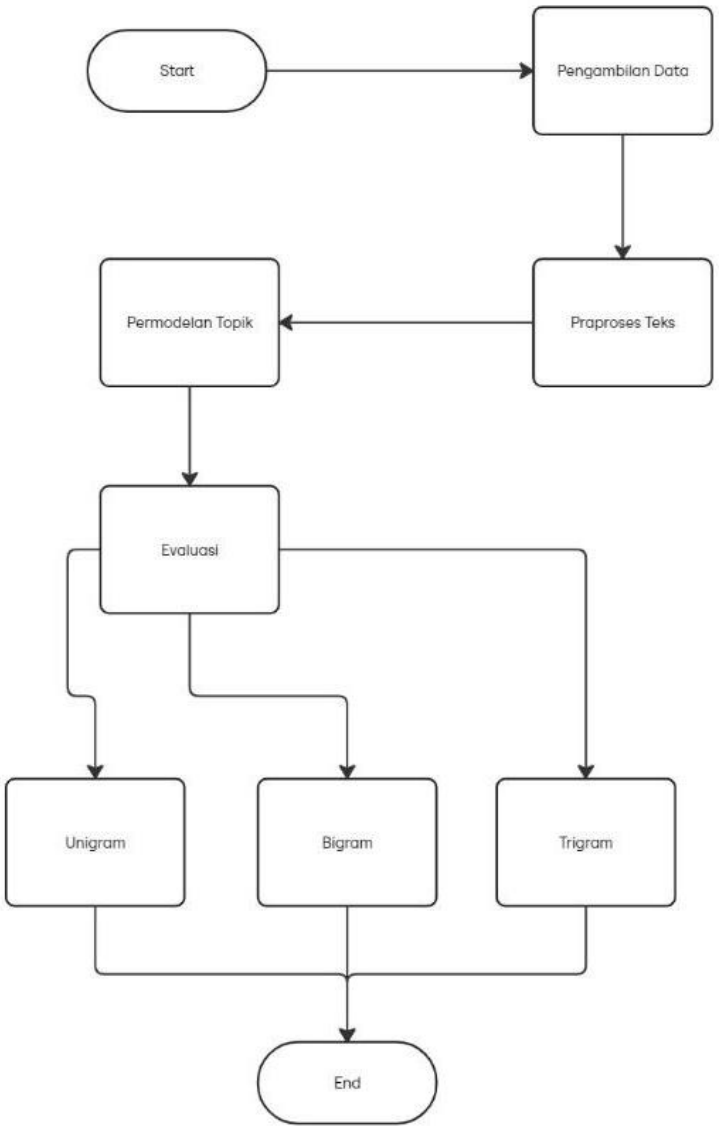
Pengambilan Data

Tahapan pertama yang dilakukan pada penelitian ini yakni melakukan pengambilan data dari kumpulan komentar Instagram. Data tersebut diambil dari akun uinsgd.official. *Instaloader* digunakan sebagai alat untuk melakukan pengumpulan data. Instaloader dipilih karena menyediakan beberapa kemudahan dalam mengambil data dari Instagram diantaranya dapat berupa video, foto, keterangan *posting*, dan komentar serta dapat mengambil data berdasarkan *hashtag* dan tanggal yang ditentukan. Data komentar yang diambil dari unggahan pada berjumlah 5244 komentar. Selain itu, data yang didapat dari proses *Web Scraping* merupakan data komentar berupa file JSON.

Praproses Teks

Tahapan selanjutnya yang dilakukan pada penelitian ini adalah melakukan praproses teks. Fase ini adalah persiapan data yang dilakukan dengan tujuan membersihkan data atau mengurangi atribut yang tidak diperlukan dan tidak mempengaruhi proses yang akan dilakukan. Seperti yang telah dijelaskan sebelumnya bahwa tahapan praproses yang digunakan diantaranya *cleansing, casefolding, stopwords removal*, *tokenizing*, dan *stemming*.

p-ISSN: 2528-1682
e-ISSN: 2527-9165



Gambar 1. Tahapan Penelitian

Cleansing

Tahapan ini dilakukan untuk menghapus karakter – karakter atau simbol - simbol yang tidak dipergunakan seperti atribut *mention* dan *hashtag* pada komentar Instagram (@, #), simbol-simbol (@#%^&*(){}[]<>.,/'"), huruf Arab seperti (ه, د, ج, ب, ا) dan kata berisi ('http'), *emoji*, dan *emoticon*[10]. Berikut ini merupakan contoh data yang diambil dari hasil *cleansing* yang tertera pada tabel 1 di bawah ini.

Tabel 1. Sampel tahapan *cleansing* pada Praproses Teks

No	*Input*	*Output*
1	Min, pendaftaran wisuda 90 kapan dibuka yah?? Hehehe	Min pendaftaran wisuda kapan dibuka yah Hehehe
2	Min pendaftaran wisuda maret kapan yaa	Min pendaftaran wisuda maret kapan yaa
3	min, kalau nau pembayaran ukt nya dicicil gimana??	min kalau nau pembayaran ukt nya dicicil gimana

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Casefolding

Proses ini memanfaatkan penggunaan fitur *transform cases* dengan tujuan penyeragaman seluruh teks ke dalam huruf kecil semua atau *lowercase*.

Tabel 2. Sampel Hasil *Casefolding*

No	*Input*	*Output*
1	Min pendaftaran wisuda kapan dibuka yah Hehehe	min pendaftaran wisuda kapan dibuka yah hehehe
2	Min pendaftaran wisuda maret kapan yaa	min pendaftaran wisuda maret kapan yaa
3	min kalau nau pembayaran ukt nya dicicil gimana	min kalau nau pembayaran ukt nya dicicil gimana

Stopwords Removal

Pada tahapan *Stopwords Removal* dilakukan dengan cara menggunakan daftar kata yang tercantum pada *stoplist* dengan membuang kata-kata yang kurang penting. Selain itu kata-kata yang penting tetap dipertahankan dalam daftar kata. Contoh *stopwords* adalah "yang", "dan", "di", "dari", "hashtag (#)", "url", tanda baca tertentu (*emoticon*), dan lainnya.

Tokenizing

Apabila karakter ke-i bukan sebagai tanda pemisah kata seperti titik (.), koma (,), dan spasi serta tanda pemisah lainnya, maka penggabungan dengan karakter selanjutnya akan dilakukan.

Tabel 4. Sampel hasil *Tokenizing*

No	*Input*	*Output*
1	min pendaftaran wisuda kapan dibuka yah hehehe	pendaftaran, wisuda, kapan, dibuka
2	min pendaftaran wisuda maret kapan yaa	pendaftaran, wisuda, maret, kapan
3	min kalau nau pembayaran ukt nya dicicil gimana	nau, pembayaran, ukt, dicicil, gimana

Stemming

Pada tahap *stemming*, dilakukan proses pengubahan kata yang berimbuhan menjadi kata dasar. Berikut sampel hasil proses *stemming*.

Tabel 5. Sampel Hasil *Stemming*

No	*Input*	*Output*
1	pendaftaran, wisuda, kapan, dibuka	daftar wisuda kapan buka
2	pendaftaran, wisuda, maret, kapan	daftar wisuda maret kapan
3	nau, pembayaran, ukt, dicicil, gimana	nau bayar ukt cicil gimana

Setelah melalui tahapan praproses teks, dataset berupa data mentah yang mengandung atribut-atribut yang tidak diperlukan saat proses klasifikasi diubah menjadi *dataset* yang bersih. *Dataset* bersih adalah kumpulan data komentar yang siap digunakan pada tahapan selanjutnya.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

N-Gram

N-Gram adalah serangkaian n-item yang berurutan dari sebuah data, biasanya berupa teks[11]. Nilai N tersebut bisa bervariasi tergantung dari kebutuhan, mulai dari satu hingga sepanjang teks yang ada[12]. Sedangkan menurut Shoffan Azizurahman, N-Gram merupakan pemecahan dan kombinasi berdasarkan jumlah N karakter dari sebuah kata atau string yang kemudian dibandingkan dengan kata yang telah disimpan pada koleksi kata (*corpus*)[13]. Dalam melakukan ekstraksi topik yang digunakan untuk mengetahui topik-topik yang dihasilkan dari data yang diambil berkaitan dengan Layanan Akademik dipergunakan *N-Gram*. Di dalam penelitian ini metode yang akan dibandingkan dengan cara membandingkan antara menggunakan *Unigram, Bigram* dan *Trigram*. Sedangkan *Term Frequency* akan dipergunakan untuk menghitung banyaknya kemunculan dari setiap kombinasi kata baik pada *unigram, bigram* maupun *trigram*.

Term Frequency

Term Frequency (TF) menghitung frekuensi jumlah kemunculan kata pada sebuah dokumen[14], [15]. Karena panjang dari setiap dokumen bisa berbeda-beda, maka umumnya nilai *Term Frequency* ini dibagi dengan panjang dokumen (jumlah seluruh kata pada dokumen).

$$tf_{t,d} = \frac{n_{t,d}}{total\ number\ of\ terms\ in\ document} \quad (1)$$

Keterangan:
tf = *Term Frequency* (jumlah kemunculan suatu kata dalam satu dokumen)

HASIL DAN PEMBAHASAN

Data yang diambil dalam penelitian ini dilakukan dengan cara melakukan penarikan data. Adapun *library* yang dipergunakan yaitu menggunakan Instaloader sebagai alat *Web Scraping*. Pengambilan data dengan Instaloader dapat dilakukan sesuai dengan apa yang dibutuhkan oleh peneliti, dengan memberikan rentang waktu tanggal 01 September 2022 sampai dengan 31 Mei 2023. Data ini diambil dengan melihat representasi data yang mencakup pada semester ganjil dan genap pada tahun Akademik 2022/2023. Selain itu, data ini diambil untuk melihat fenomena terhadap pola data yang dihasilkan pada setiap awal semester genap dan ganjil.

Hasil Praproses Teks

Pada bagian ini terdiri dari hasil tahapan praproses teks. Adapun tahapan ini dimulai dari tahapan *cleansing*, dimana tahapan ini menghapus karakter-karakter tidak dipergunakan dan tidak bermakna dalam memproses sebuah teks. Pada tahapan ini pula dilakukan *casefolding* dimana setiap kata diubah kedalam bentuk *lowercase*. Sehingga tidak terdapat huruf yang ditulis dalam bentuk huruf kapital. Berikut ini merupakan hasil dari proses *cleansing* seperti pada gambar 1 dibawah ini.

p-ISSN: 2528-1682
e-ISSN: 2527-9165

```
                                                  text
0     kalo rumahnya ngontrak surat kepemilikan dan p...
1     kalo yang baru mau masuk pendaftarannya kapan ...
2                       halo bolehkah di baca dm ny kak
3     min kalau pemberkasan itu yang di kumpulin di ...
4     saya salah terus didata ayahapa ada kontak bag...
...                                                 ...
5239                          tuu liat komenan buaya teh
5240       bukannya km ya yg deketin adek adek tingkat
5241                                  hallo aa fahriz
5242                                     naon sih bon
5243                                            oalah

[5244 rows x 1 columns]
```

Gambar 2. Hasil *Cleansing* pada tahapan Praproses Teks

Stopwords removal merupakan tahapan dimana penghapusan kata dilakukan terhadap kata yang tidak memiliki makna berarti. Kata tersebut diantaranya yaitu kata penghubung, kata pengganti orang. Implementasi pada tahapan ini dilakukan dengan menggunakan *library* NLTK dan Sastrawi. Selain itu, pada tahapan ini daftar kata *stopwords* ditambahkan seperti yang ditunjukkan pada tabel 6.

Tabel 6. Daftar Kata *Stopwords* Tambahan

Stopwords Removal
'uin', 'bandung', 'sgd', 'innalillahi', 'wa', 'inna', 'sunan', 'gunung', 'djati', 'innailaihi', 'kalo', 'ilaihi', 'min', 'masya', 'allah', 'rojiun', 'ga', 'hatur', 'nuhun', 'nya', 'kasih', 'husnul', 'khotimah', 'wainna', 'syaa', 'selamat', 'keren', 'banget', 'jalan', 'na', 'waafihi', 'wafuanhu', 'iya', 'nih', 'dm', 'assalamualaikum', 'kak', 'warhamhu', 'century', 'girl', 'proud', 'of'

Pada tahapan selanjutnya dilakukan penyederhanaan terhadap bentuk kata (*stemming*). *Stemming* ini digunakan untuk mendapatkan kata dasar dari setiap kata yang mengandung imbuhan. Dari hasil *stemming* yang dilakukan terdapat 5.113 data. Berikut ini merupakan hasil dari *stemming* dari penelitian ini.

```
                                                  text
0                 rumah ngontrak surat milik pbb gimana
1                                           masuk daftar
2                                           halo baca ny
3                 berkas kumpulin gedung al jamiah beda
4     salah data ayahapa kontak bagiaan mahasiswa ko...
...                                                 ...
5108                          tuu liat komenan buaya teh
5109                km yg deketin adek adek tingkat
5110                                  hallo aa fahriz
5111                                     naon sih bon
5112                                            oalah

[5113 rows x 1 columns]
```

Gambar 3. Hasil *Stemming* pada Tahapan Praproses Teks

Hasil Permodelan Topik dengan N-Gram

Setelah data melalui rangkaian tahapan praproses teks, maka data tersebut siap untuk dilakukan permodelan topik dengan menggunakan *N-Gram*. Proses perhitungan *N-Gram* dilakukan dengan menggunakan data hasil *stemming* yang berjumlah 5.113 data. Hasil penelitian ini akan berfokus pada perbandingan antara

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Unigram, *Bigram* dan *Trigram*. Penelitian ini menggunakan *library* sklearn yaitu *CountVectorizer* untuk melakukan perhitungan *N-Gram*.

Hasil Permodelan dengan *Unigram*

N-gram adalah pemotongan kalimat atau string menjadi lebih kecil sesuai dengan N karakter. Tahapan pada permodelan dengan *Unigram* ini dilakukan dengan melakukan perhitungan frekuensi kata dengan formula *Term Frequency*. Setelah frekuensi pada setiap kata dihitung, selanjutnya dilakukan pengurutan dari frekuensi terbesar hingga terkecil. Pada bagian pertama, perhitungan kata yang dilakukan dihitung hanya satu suku kata saja (*unigram*). Dari hasil perhitungan tersebut, kata-kata dengan topik wisuda, daftar dan kampus merupakan topik yang paling sering disebutkan pada Instagram terkait tema layanan akademik. Disamping itu, topik lainnya seperti jurus, mahasiswa, masuk dan jalur merupakan topik selanjutnya yang menghasilkan frekuensi terbanyak. Pada gambar 4 dibawah ini hasil yang ditunjukkan bahwa kata-kata yang mengandung frekuensi terbanyak merupakan kata yang relevan dengan topik layanan akademik, namun kata tersebut mengandung makna yang masih luas.

	frekuensi	unigram
0	282	wisuda
2	255	daftar
3	191	kampus
6	157	jurus
8	142	mahasiswa
9	137	masuk
10	131	jalur

Gambar 4. Hasil Permodelan Topik dengan *Unigram*

Hasil Permodelan dengan *Bigram*

Selanjutnya untuk melihat perbandingan dari permodelan topik yang dihasilkan pada dilakukan perhitungan dengan menggunakan Bi-Gram. Perhitungan dilakukan dengan menggunakan *library* yang sama yaitu *CountVectorizer*. Hasilnya menunjukkan bahwa pasangan kata (*Bi-Gram*) memperlihatkan frekuensi tertingginya yaitu "daftar wisuda" sebanyak 31 dan paling rendah oleh "jalur umptkin" sebanyak 11 dari 5.113 total data yang diproses.

	frekuensi	bigram
0	31	daftar wisuda
1	31	bayar ukt
2	29	jalur mandiri
3	23	jalur spanptkin
4	15	uji mandiri
6	13	jalur snbp
11	11	jalur umptkin

Gambar 5. Hasil Permodelan Topik dengan *Bigram*

p-ISSN: 2528-1682
e-ISSN: 2527-9165

Hasil Permodelan dengan *Trigram*

Setelah mendapatkan hasil *Bi-Gram*, maka tahapan selanjutnya perbandingan dilakukan dengan menghitung pasangan tiga kata (*Trigram*). Dari hasil perhitungan yang dilakukan menunjukkan bahwa pasangan kata yang dihasilkan sudah cukup merepresentasikan topik yang dimaksud, namun jika dilihat frekuensinya menghasilkan angka yang sangat sedikit sehingga kurang tepat jika dikatakan kata tersebut merupakan representasi dari pertanyaan yang sering ditanyakan. Berikut ini merupakan hasil permodelan topik dengan menggunakan *Trigram* yang dapat dilihat pada gambar 5.

	frekuensi	trigram
0	6	jalur mandiri buka
1	5	daftar wisuda buka
6	4	info daftar wisuda
10	4	daftar wisuda maret
11	4	daftar jalur mandiri
13	3	umptkin uji mandiri
16	3	telat bayar ukt

Gambar 6. Hasil Permodelan Topik dengan Trigram

Dari hasil perhitungan terhadap ketiga model *Unigram, Bigram* dan *Trigram* bahwa hasil yang lebih relevan dapat ditunjukkan dengan permodelan topik pada model *Bigram*. Pada pasangan kata tersebut menunjukkan bahwa kata-kata tersebut relevan terhadap topik layanan akademik dan mudah dipahami dibandingan dengan pasangan kata pada *Unigram*. Dari keseluruhan data tersebut juga menunjukkan bahwa terdapat topik seputar penerimaan mahasiswa baru seperti jalur mandiri, jalur spanptkin, uji mandiri, jalur snbp dan jalur umptkin. Sedangkan pasangan kata terbesar yang disebutkan dalam data yakni terkait daftar wisuda dan bayar ukt.

PENUTUP

Penelitian ini menunjukkan bahwa permodelan topik pada layanan akademik perguruan tinggi menghasilkan topik yang sangat relevan bila menggunakan *Bigram*. Sedangkan bila menggunakan *Unigram*, topik tersebut dapat dihasilkan namun kata tersebut masih mencakup makna yang luas. Selain itu bila permodelan topik dilakukan dengan menggunakan *Trigram* menghasilkan frekuensi yang relatif dan jumlahnya yang kecil membuat pasangan kata yang dihasilkan mempunyai relevansi yang kurang tepat dari data yang diambil dari akun media sosial. Penelitian ini merupakan bagian dari penelitian yang masih dikembangkan dalam rangka menghasilkan generator pertanyaan dari topik-topik yang sering ditanyakan. Selain itu, jika dikembangkan lebih lanjut, penelitian ini akan menjadi sangat berharga dan dapat digunakan secara umum dalam rangka mendapatkan topik dan pertanyaan atau komentar yang paling sering ditanyakan pada suatu data set. Sehingga diharapkan penelitian ini berkembang menjadi sesuatu yang lebih kompleks dan bermanfaat seperti pembuatan *FAQ Generator*, *Chatbot* dan rekomendasi jawaban dengan menggunakan kecerdasan buatan.

DAFTAR PUSTAKA

[1] M. Jamil, S. F. Saputra, M. I. Wahid, and D. Riana, "Evaluasi Metode ISO/IEC 9126 Pada Kinerja Website Sistem Informasi Akademik Perguruan Tinggi," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 16, no. 1, Art. no. 1, Mar. 2021, doi: 10.30872/jim.v16i1.5209.

[2] Afnibar and D. F. N, "Pemanfaatan Whatsapp Sebagai Media Komunikasi Antara Dosen dan Mahasiswa dalam Menunjang Kegiatan Belajar (Studi Terhadap Mahasiswa UIN Imam Bonjol Padang)," *AL MUNIR J. Komun. Dan Penyiaran Islam*, vol. 0, no. 0, pp. 70–83, Jun. 2020, doi: 10.15548/AMJ-KPI.V0I0.1501.

[3] W. C. Indhiarta, "Penggunaan N-Gram pada Analisa Sentimen Pemilihan Kepala Daerah Jakarta Menggunakan Algoritma Naïve Bayes."

[4] N. L. Lavenia and R. Permatasari, "Sentiment Analysis on Twitter Social Media Regarding Depression Disorder Using the Naive Bayes Method," *CoreID J.*, vol. 1, no. 2, pp. 66–74, Jul. 2023, doi: 10.60005/COREID.V1I2.14.

[5] A. Apriani, H. Zakiyudin, and K. Marzuki, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF System Penerimaan Mahasiswa Baru pada Kampus Swasta," *J. Bumigora Inf. Technol. BITe*, vol. 3, no. 1, pp. 19–27, Jul. 2021, doi: 10.30812/BITE.V3I1.1110.

[6] S. A. Sugianto, Liliana, and S. Rostianingsih, "Pembuatan Aplikasi Predictive Text Menggunakan Metode N-Gram-Based," *J. Infra*, vol. 1, no. 2, p. 125-p.130, Jul. 2013.

[7] E. L. Amalia, A. J. Jumadi, I. A. Mashudi, and W. Wibowo, "Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA)," *J. Teknol. Inf. Dan Ilmu Komput.*, vol. 8, no. 2, pp. 343–348, Mar. 2021, doi: 10.25126/JTIIK.2021824356.

[8] "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter | Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)." Accessed: Nov. 28, 2024. [Online]. Available: https://jurnal.iaii.or.id/index.php/RESTI/article/view/3146

[9] M. F. Juna and M. Hayaty, "The observed preprocessing strategies for doing automatic text summarizing," *Comput. Sci. Inf. Technol.*, vol. 4, no. 2, pp. 119–126, Jul. 2023, doi: 10.11591/CSIT.V4I2.P119-126.

[10] M. K. K. Insan, U. Hayati, and O. Nurdiawan, "Analisis Sentimen Aplikasi Brimo pada Ulasan Pengguna di Google Play Menggunakan Algoritma Naive Bayes ANALISIS SENTIMEN APLIKASI BRIMO PADA ULASAN," *JATI J. Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 478–483, Mar. 2023, doi: 10.36040/JATI.V7I1.6373.

[11] Y. Zhang and Z. Rao, "N-BiLSTM: BiLSTM with n-gram Features for Text Classification," *Proc. 2020 IEEE 5th Inf. Technol. Mechatron. Eng. Conf. ITOEC 2020*, pp. 1056–1059, Jun. 2020, doi: 10.1109/ITOEC49072.2020.9141692.

[12] C. Supriadi, H. D. Purnomo, I. Sembiring, and J. O. N. B. Sidorejo, "Sensitivitas Sistem Pencarian Artikel Bahasa Indonesia Menggunakan Metode n-gram Dan Tanimoto Cosine," *J. Transform.*, vol. 18, no. 1, pp. 63–70, Jul. 2020, doi:

10.26623/TRANSFORMATIKA.V18I1.2184.

[13] S. Azizurahman, Y. Firdaus, and A. A. Suryani, "Analisis Dan Implementasi Metode N-Gram Pada Information Retrieval," *Tugas Akhir Fak. Tek. Inform. Univ. Telkom*, 2011.

[14] E. K. Susanto, M. B. Subkhi, A. Z. Arifin, M. Maryamah, R. W. Sholikah, and R. Indraswari, "Metode Pembobotan Hibrida untuk Ekstraksi Frasa Kunci Bahasa Arab," *INSYST J. Intell. Syst. Comput.*, vol. 4, no. 2, pp. 93–101, Oct. 2022, doi: 10.52985/INSYST.V4I2.255.

[15] M. Irfan, Jumadi, W. B. Zulfikar, and Erik, "Implementation of Fuzzy C-Means algorithm and TF-IDF on English journal summary," in *2017 Second International Conference on Informatics and Computing (ICIC)*, Nov. 2017, pp. 1–5. doi: 10.1109/IAC.2017.8280646.