TOPIC MODELING JUDUL PENELITIAN MENGGUNAKAN METODE *LATENT DIRICHLET ALLOCATION (LDA)* DAN FAKTORISASI MATRIKS *NON-NEGATIF (NMF)*

Saikin, Sofiansyah Fadli, Jihadul Akbar, Hairul Fahmi
Sistem Informasi, STMIK Lombok, Praya
Jln.Basuki Rahmat No.105 Praya, Lombok Tengah, Indonesia
eken.apache@gmail.com

ABSTRAK

Penelitian merupakan salah satu aspek penting dalam Tri Dharma Perguruan Tinggi yang berkontribusi dalam pengembangan ilmu pengetahuan melalui eksplorasi dan pemecahan masalah. Salah satu tantangan dalam analisis penelitian dosen adalah mengelompokkan tema penelitian berdasarkan judul yang tersedia agar dapat mengidentifikasi tren topik yang sedang berkembang. Permasalahan utama dalam penelitian ini adalah bagaimana mengelompokkan judul penelitian dosen ke dalam topik yang lebih terstruktur secara otomatis. Oleh karena itu, penelitian ini bertujuan untuk menerapkan metode *Latent Dirichlet Allocation (LDA)* dan *Non-Negative Matrix Factorization (NMF)* dalam topic modeling guna mengidentifikasi tren penelitian berdasarkan judul penelitian dosen. Metode yang digunakan dalam penelitian ini melibatkan tahap pengumpulan data, pra-pemrosesan teks, pemodelan topik menggunakan LDA dan NMF, serta evaluasi hasil. LDA dan NMF dipilih karena kemampuannya dalam mengidentifikasi distribusi kata dalam suatu korpus dan mengelompokkan dokumen berdasarkan pola kemunculan kata. Hasil penelitian menunjukkan bahwa model LDA menghasilkan lima topik utama dengan distribusi kata tertinggi pada Topik 0 (10.389 kata), diikuti oleh Topik 1 (2.517 kata), Topik 2 (2.432 kata), Topik 3 (2.286 kata), dan Topik 4 (2.247 kata). Sementara itu, model NMF menghasilkan distribusi data tertinggi pada Topik 0 (9.980 data), diikuti oleh Topik 2 (3.765 data), Topik 3 (2.275 data), Topik 1 (2.202 data), dan Topik 4 (1.649 data). Dari hasil tersebut, terlihat bahwa metode LDA dan NMF memiliki perbedaan dalam pemetaan topik, namun keduanya dapat digunakan untuk analisis tren penelitian dosen.

Kata kunci : *Topic Modeling, Latent Dirichlet Allocation (LDA), Faktorisasi Matriks Non-Negatif (NMF), Analisis Topik Penelitian, Python, Jupyter Notebook*

1. PENDAHULUAN

Penelitian dosen merupakan proses sistematis yang bertujuan menggali, mengeksplorasi, dan mengembangkan ilmu pengetahuan melalui pemecahan masalah penelitian. Kegiatan ini merupakan salah satu dari Tri Dharma Perguruan Tinggi, bersama dengan pengajaran dan pengabdian kepada Masyarakat. Pada tahun 2024, sebanyak 19 ribu lebih judul penelitian dosen lolos pendanaan dari Kementerian Riset dan Teknologi/Badan Riset dan Inovasi Nasional (Kemenristek/BRIN). Penelitian-penelitian tersebut mencakup berbagai kategori, seperti penelitian dasar (PD), penelitian terapan (PT), penelitian pengembangan (PP), penelitian dosen pemula (PDP), dan kategori bidang lainnya. Tema atau judul penelitian yang lolos seleksi umumnya merupakan topik-topik terkini yang relevan dengan isu-isu yang sedang berkembang saat ini. Mendapatkan topik penelitian yang bagus perlu melakukan analisis topik penelitian pada data judul penelitian yang lolos pada tahun 2024.

Pemodelan topik *(topic modeling)* merupkan suatu Teknik untuk mengelompokan kalimat atau dokumen berdasarkan hitungan probalistik penyusunan kata-kata *(word embedding)* kedalam subjek yang terkait [1] [2].

Topic modeling bertujuan mengeskstrak hubungan antar satu kata dengan kata lainnya dalam kumpulan dokumen. Metode *topic modeling* bisanya digunakan untuk memahami topik- topik pembicaran yang popular yang dapat diungkap berdasarkan korpus tertentu misalnya bidang kesehatan, pendidikan, penelitian, dan lain sebagainya [3].

Latent Dirichlet Allocation (LDA) yang dikembangkan pada tahun 2003 oleh peneliti David Blei, Andrew Ng dan Michael Jordan. Merupakan salah satu metode paling populer dalam pemodelan topik. LDA dapat digunakan untuk meringkas, klasterisasi, menghubungkan dan memproses data yang sangat besar karena LDA mampu menampilkan daftar topik dan membandingkan besar kemunculan topik [4].

Pada penelitian sebelumnya, analisis pemodelan topik untuk tugas akhir mahasiswa menggunakan metode *Latent Dirichlet Allocation (LDA)* telah berhasil meningkatkan nilai koherensi dari 0,53448 menjadi 0,617789 melalui penyetelan parameter. Dengan demikian, metode LDA mampu mereduksi setiap kata kedalam empat klaster utama yang bertemakan Tata Kelola pada topik 1, 2, 3, dan 4 [5].

Penelitian yang lainya analisis sosial media untuk analisis stunting di Indonesia. Metode yang digunakan ialah LDA dan hasil dari penelitian menunjukkan sentimen negatif mendominasi sebesar 60,6%, sentimen positif sebesar 31,5%, dan netral sebesar 7,9%. Selain itu, penelitian ini menunjukkan bahwa

'anak-anak', 'penurunan', 'angka', 'pencegahan', dan 'gizi' merupakan kata-kata yang sering muncul pada stunting [6].

Penelitian dengan judul Pemodelan Topik Menggunakan *n-Gram* dan *Non-negative Matrix Factorization.* Pada penelitian ini, pemodelan topic yang digunakan adalah *Non-Negative Matrix Factorization (NMF)* dengan mengimplementasikan n-gram hasil penelitian yang dilakukan diperoleh bahwa jumlah topik yang dapat dibentuk dari RMOL.id untuk unigram adalah 15 topik dengan nilai coherence value 0.812748 [7].

Penelitian selanjutnya ialah menggunakan metode LDA dikombinasikan dengan algoritma faktorisasi matriks *Non-Negatif (NMF)* untuk analisis topik penelitian yang lolos mendapatkan hibah pada tahun 2024. Tujuang dari penelitian ini untuk mengetahui topik penelitian apa saja yang lolos pada tahun tersebut sedang tools yang digunalan untuk analisi dan modeling data menggunakan *jupyter notebook* dengan Bahasa pemrograman *python*.

2. TINJAUAN PUSTAKA

Penelitian tentang analisis topik menggunakan metode *Latent Dirichlet Allocation (LDA)* dan Faktorisasi *Matriks Non-Negatif (NMF)* telah banyak dikembangkan untuk berbagai keperluan. *Topic Modeling*, sebuah teknik unsupervised machine learning, sering digunakan untuk menemukan struktur tersembunyi dalam kumpulan dokumen teks yang besar. Teknik ini berguna untuk mengidentifikasi tema-tema yang relevan dalam berbagai bidang, seperti pendidikan, kesehatan, dan penelitian ilmiah. *Latent Dirichlet Allocation (LDA)* yang diperkenalkan oleh Blei, Ng, dan Jordan (2003) adalah salah satu metode yang paling populer dalam analisis topik. LDA memungkinkan pengelompokan berdasarkan distribusi probabilistik kata-kata yang membentuk topik-topik laten dalam kumpulan data [1] [8].

Dalam pengembangan lebih lanjut, Faktorisasi *Matriks Non-Negatif (NMF)* menjadi alternatif metode analisis teks yang menawarkan pendekatan berbeda. NMF menguraikan matriks data menjadi dua *matriks non-negatif*, yang dapat digunakan untuk memahami pola laten dalam data teks. Metode ini unggul dalam pengelompokan data yang kompleks dan menghasilkan representasi yang lebih ringkas dibandingkan metode lain, seperti LDA [9] [10].

Kombinasi LDA dan NMF sering digunakan untuk meningkatkan akurasi analisis data teks. Penelitian sebelumnya menunjukkan bahwa pendekatan gabungan ini dapat mengidentifikasi tema utama dalam kumpulan dokumen besar dengan lebih efektif [11].

Penggunaan algoritma ini telah diterapkan pada analisis tema penelitian dosen, identifikasi sentimen pada media sosial, dan eksplorasi topik dalam berbagai domain aplikasi lainnya. Sebagai contoh, penelitian pada analisis media sosial di Indonesia menunjukkan bahwa kombinasi LDA dan NMF berhasil mengidentifikasi pola-pola yang relevan dengan isu stunting dan hubungannya dengan gizi masyarakat [12].

Selain itu, alat seperti *Jupyter Notebook* dan bahasa pemrograman *Python* mempermudah implementasi metode LDA dan NMF. Python memiliki berbagai pustaka, seperti *Scikit-Learn, Gensim,* dan *NumPy,* yang mendukung pengolahan data dalam analisis teks [13].

Penggunaan *Python* sebagai alat analisis membantu peneliti untuk mengembangkan model dengan efisiensi tinggi dan akurasi yang optimal [14].

Penelitian ini melanjutkan upaya tersebut dengan mengaplikasikan metode LDA dan NMF untuk menganalisis tema penelitian dosen yang mendapatkan hibah dari Kemenristek/BRIN pada tahun 2024. Studi ini diharapkan dapat memberikan wawasan baru mengenai tema penelitian yang menjadi tren dan relevansi metode analisis topik dalam skala besar.

2.1. Topic Modeling

Salah satu teknik *unsupervised machine learning* yang digunakan untuk menemukan tema tersembunyi dari kumpulan dokumen teks besar dan mengelompokkan tema-tema tersebut menjadi satu topik [15].

Topic modeling bertujuan untuk mengekstrak hubungan antar kata dalam satu kumpulan dokumen. Metode ini umumnya digunakan untuk mengidentifikasi topik pembicaraan populer yang dapat diungkap dari suatu korpus tertentu, misalnya bidang kesehatan, pendidikan, penelitian, dan sebagainya [6].

Topic Modeling merupakan bentuk model statistik yang digunakan dalam machine learning dan *Natural Language Processing (NLP)* untuk mengidentifikasi struktur tersembunyi dari kumpulan data atau teks. Topic Modelling berfungsi untuk mengelompokkan kumpulan data atau teks menjadi beberapa topik [16].

Hal ini dapat dimanfaatkan untuk segmentasi produk dan pembuatan rekomendasi produk bagi stakeholder yang dapat memberikan keuntungan jangka panjang.

2.2. *Latent Dirichlet Allocation (LDA)*

Latent Dirichlet Allocation (LDA) merupakan salah satu metode yang populer dalam analisis topic modeling [1].

LDA juga merupakan model probabilistik yang digunakan untuk mengidentifikasi topik-topik laten (tersembunyi) dalam suatu kumpulan dokumen. Model ini didasarkan pada asumsi bahwa setiap dokumen merupakan campuran dari beberapa topik, dan setiap topik diwakili oleh distribusi kata-kata yang unik [4].

LDA sangat berguna untuk menemukan topik-topik yang mendasari kumpulan dokumen yang besar, seperti koleksi artikel berita, postingan blog, atau ulasan produk [17].

2.3. *Faktorisasi Matriks Non-Negatif (NMF)*

Non-Negative Matrix Factorization (NMF) merupakan model statistik yang dikembangkan untuk faktorisasi matriks pada data yang mempunyai struktur kompleks seperti data teks maupun dokumen [18].

NMF digunakan untuk mereduksi dimensi data dan pemodelan topik dengan cara membuat faktor sebuah matriks menjadi dua matriks non-negatif. Teknik pembelajaran Machine Learning pada metode *Non-Negative Matrix Factorization* difokuskan untuk menemukan topik yang tidak langsung terlihat di dalam data teks yang digabungkan *(hidden topic). Non-Negative Matrix Factorization* yang berhasil menemukan topik-topik laten, dapat menjelaskan hubungan kompleks antar kata dan memberikan akurasi yang lebih tinggi dan terperinci [19].

2.4. *Python*

Python salah satu Bahasa pemrograman yang sintaks yang digunakan sangat jelas dan mirip Bahasa inggris. *Python* digunakan untuk berbagai macam tujuan mulai dari pemrograman web, analisis data, hingga kecerdasan buatan [20].

Python terkenal akan bermacam-ragam kerangka kerjanya mulai dari Numpy adalah kerangka kerja terkemuka yang dirancang untuk perhitungan numerik dalam Python, *SciPy* adalah modul untuk sains, matematika, dan teknik, Scikit-Learn adalah kerangka kerja *machine learning Python* untuk penambangan data yang produktif, sehingga memungkinkan untuk melakukan proses regresi, pengelompokan, pemilihan model, *preprocessing*, dan klasifikasi [21].

2.5. *Jupyter Notebook*

Jupyter Notebook, sebuah IDE (Integrated Development Environment) interaktif, yang dapat menulis dan mendistribusikan dokumen dengan teks naratif, kode, visual, dan hasil eksekusi kode. *Jupyter notebook* sering digunakan untuk pembelajaran mesin, simulasi numerik, visualisasi data, pemrosesan data, dan eksplorasi data [22].

Jupyter notebook merupakan IDE menggunakan bahasa pemrograman python merupakan bahasa pemrograman tingkat tinggi yang sering digunakan dalam sejumlah pengembangan perangkat lunak dan domain aplikasi [23].

3. METODE PENELITIAN

Tahapan penelitian merupakan Langkah-langkah penelitian yang akan dilakukan pada penelitian ini.

Gambar 1. Tahapan Penelitian

3.1. Data *Collection*

Datasheet diambil dari judul penelitian dosen yang lolos mendapatkan hibah pada tahun 2024. Datasheet terdiri dari 19871 baris data dan 7 fitur datasheet, fitur-fitur tersebut ialah kategori merupakan kolom jenis perguruan tinggi, fitur institusi merupakan fitur nama-nama perguruan tinggi, NIDN fitur nomor induk dosen nasional, Nama fitur untuk nama-nama peneliti judul fitur untuk judul penelitian dan fitur lingku merupakan fitur untuk jenis penelitian.

```
df.head()
```

	Unnamed: 0	Kategori	institusi	NIDN	Nama	Judul	Lingkup
0	0	PTNBH	Institut PertanianInBogor	27046503.0	Achmad Farajallah	['pola', 'distribusi', 'macrobrachium', 'sinta...	PPS-PTM
1	1	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	['aplikasi', 'bio', 'ofertilizers', 'drone', '...	PFR
2	2	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	['aplikasi', 'seed', 'bomb', 'technology', 'cs...	PPS-PTM
3	3	PTNBH	Institut PertanianInBogor	2076607.0	Agus Buono	['optimasi', 'model', 'hybrid', 'convolutional...	PPS-PDD
4	4	PTNBH	Institut PertanianInBogor	18096208.0	Agus Hikmat	['pengembangan', 'sistem', 'agroforestri', 'pr...	PFR

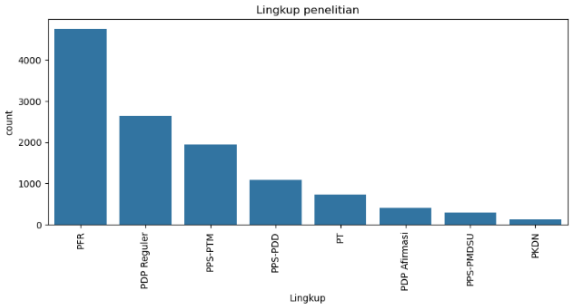
```
df.shape
```

```
(19871, 7)
```

Gambar 2. Datasheet judul penelitian tahun 2024

3.2. Analisis Lingkup Penelitian

Lingkup penelitian merupakan fitur untuk jenis penelitian yang diajukan disini ada beberapa jenis penelitian yang terdiri dari PFR, PDF regular, PPS-PTM, PPS-PDD, PT, PDP Afirmasi, PPS-PMDSU dan PKDN. Dari jenis penelitian tersebut PFR dan PDP Reguler yang lingkup penelitian yang banyak lolos.



Gambar 3. Analisi lingkup penelitian

3.3. Analisis *Wordcloud*

Analisis dengan melakkan visualisasi data dengan menggunakan Teknik *wordcloud* untuk menampilkan data kata pada judul penelitian kata yang paling banyak muncul atau digunakan pada judul penelitian. Pada visualisai dibawah menunjukan kata yang paling besar tampilan ialah berbasis, pengembangan, model, studi dan Indonesia. Diikuti oleh kata pendekatan, pengaruh dan metode.

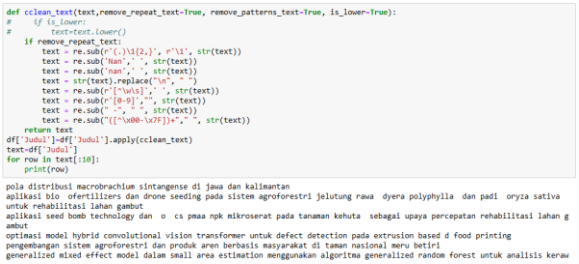
Gambar 4. Visualisasi *wordcloud*

3.4. *Text Cleaning*

Preprocessing text adalah serangkaian langkah yang dilakukan untuk membersihkan dan menyederhanakan teks sehingga lebih siap untuk diproses oleh model pembelajaran mesin. Tujuannya adalah untuk meningkatkan kualitas dan akurasi hasil analisis. Hasil visualisai menunjukan banyak *Nan value* yang ada pada setiap teks.

Gambar 5. Visualisasi teks

Tahap preprocessing dimulai dengan pembersihan teks tujuan untuk menghilangkan simbol-simbol atau karakter yang mengganggu pada tahap modeling.



```
def cclean_text(text,remove_repeat_text=True, remove_patterns_text=True, is_lower=True):
#     if is_lower:
#         text=text.lower()
    if remove_repeat_text:
        text = re.sub(r'(.)\1{2,}', r'\1', str(text))
        text = re.sub('Nan',' ', str(text))
        text = re.sub('nan',' ', str(text))
        text = str(text).replace("\n", " ")
        text = re.sub(r'[^\w\s].', ' ', str(text))
        text = re.sub(r'[0-9]',"", str(text))
        text = re.sub(" ", " ", str(text))
        text = re.sub("([^\x00-\x7F])+"," ", str(text))
    return text
df['Judul']=df['Judul'].apply(cclean_text)
text=df['Judul']
for row in text[:10]:
    print(row)
```

```
pola distribusi macrobrachium sintangense di jawa dan kalimantan
aplikasi bio  ofertilizers dan drone seeding pada sistem agroforestri jelutung rawa  dyera polyphylla  dan padi  oryza sativa
untuk rehabilitasi lahan gambut
aplikasi seed bomb technology dan  o  cs pmaa npk mikroserat pada tanaman kehuta  sebagai upaya percepatan rehabilitasi lahan g
ambut
optimasi model hybrid convolutional vision transformer untuk defect detection pada extrusion based d food printing
pengembangan sistem agroforestri dan produk aren berbasis masyarakat di taman nasional meru betiri
generalized mixed effect model dalam small area estimation menggunakan algoritma generalized random forest untuk analisis keraw
```

Gambar 6. Pembersian text

Hasil yang didapatkan setelah melakukan pembersihan teks dan value sudah dibersihkan, terlihat seperti gambar dibawah.

Gambar 7. Hasil permbersihan teks

3.5. *Tokenize*

Tokenize merupakan bagian awal pada pemrosesan teks dan sangat penting yang berfungsi untuk memecah kalimat didalam datasheet menjadi potongan kata-kata yang lebih kecil yang disebut token. Tujuannya membuat data lebih terstruktur dan mudah difahami oleh model yang aka dibangun.

```
def word_token(text):
    return word_tokenize(text)
df['Judul']=df['Judul'].apply(word_token)
text=df['Judul']
for row in text[:10]:
    print(row)
```

```
['pola', 'distribusi', 'macrobrachium', 'sintangense', 'di', 'jawa', 'dan', 'kalimantan']
['aplikasi', 'bio', 'ofertilizers', 'dan', 'drone', 'seeding', 'pada', 'sistem', 'agroforestri', 'jelutung', 'rawa', 'dyera',
'polyphylla', 'dan', 'padi', 'oryza', 'sativa', 'untuk', 'rehabilitasi', 'lahan', 'gambut']
['aplikasi', 'seed', 'bomb', 'technology', 'dan', 'o', 'cs', 'pmaa', 'npk', 'mikroserat', 'pada', 'tanaman', 'kehuta', 'sebaga
i', 'upaya', 'percepatan', 'rehabilitasi', 'lahan', 'gambut']
['optimasi', 'model', 'hybrid', 'convolutional', 'vision', 'transformer', 'untuk', 'defect', 'detection', 'pada', 'extrusion',
'based', 'd', 'food', 'printing']
['pengembangan', 'sistem', 'agroforestri', 'dan', 'produk', 'aren', 'berbasis', 'masyarakat', 'di', 'taman', 'nasional', 'mer
u', 'betiri']
['generalized', 'mixed', 'effect', 'model', 'dalam', 'small', 'area', 'estimation', 'menggunakan', 'algoritma', 'generalized',
'random', 'forest', 'untuk', 'analisis', 'kerawa', 'pangan']
['pengembangan', 'varietas', 'kentang', 'unggul', 'tipe', 'industri', 'berdaya', 'hasil', 'tinggi']
['isolasi', 'coxiella', 'burnetii', 'asal', 'indonesia', 'melalui', 'teknik', 'propagasi', 'in', 'vivo', 'sebagai', 'dasar', 'p
engembangan', 'alat', 'diagnostik', 'cepat', 'q', 'fever']
['karakteristik', 'genetik', 'chlamydia', 'sp', 'pada', 'burung', 'paruh', 'bengkok']
['studi', 'patogenesis', 'q', 'fever', 'dan', 'karakteristik', 'agen', 'penyebab', 'coxiella', 'burnetii', 'pada', 'sapi']
```

Gambar 8. Tokenize Text

3.6. *Stopword Removing*

Stopword adala kata-kata umum atau kata penghubung yang sering muncul pada suatu teks, namun kata-kata tersebut tidak memberikan informasi penting atau nilai tambah pada analisis teks. Pada analisis teks kata-kata tersebut sering diabaikan atau dihilangkan pada *preprocessing* teks. Teknik yang digunakan untuk *stopword removing* ialah membuat list kata-kata yang sering muncul atau kata penghubung yang ada pada *datasheet*.

```
list_stopwords = stopwords.words('indonesian')
list_stopwords.extend(["di", "dan", "atau", "pun", "ada", "dimana", 'pada',
                       'atau', 'antara', 'buat', 'apa', 'ini','saya',
                       'bukan', 'bila', 'karena', 'nya', 'itu', 'bagimana',
                       'suatu', 'tau', 'tahu', 'bukan', 'tiada', 'apapun',
                       'sudah', 'saja', 'akan', 'bagaimana', 'sesuatu'])
```

Gambar 9. List *stopword*

Langkah selanjutnya menghapus *stopword* berdasarkan list kata-kata sambung atau kata-kata yang sering muncul. Setiap ada kata yang ada pada list dan ditemukan ada pada datasheet otomatis akan dihapus.

```
list_stopwords=set(list_stopwords)
def stopwords_removal(words):
    return [word for word in words if word not in list_stopwords]
df['Judul'] = df['Judul'].apply(stopwords_removal)
text=df['Judul']
for row in text[:10]:
    print(row)
```

```
['pola', 'distribusi', 'macrobrachium', 'sintangense', 'jawa', 'kalimantan']
['aplikasi', 'bio', 'ofertilizers', 'drone', 'seeding', 'sistem', 'agroforestri', 'jelutung', 'rawa', 'dyera', 'polyphylla', 'p
adi', 'oryza', 'sativa', 'rehabilitasi', 'lahan', 'gambut']
['aplikasi', 'seed', 'bomb', 'technology', 'cs', 'pmaa', 'npk', 'mikroserat', 'tanaman', 'kehuta', 'upaya', 'percepatan', 'reha
bilitasi', 'lahan', 'gambut']
['optimasi', 'model', 'hybrid', 'convolutional', 'vision', 'transformer', 'defect', 'detection', 'extrusion', 'based', 'food',
'printing']
['pengembangan', 'sistem', 'agroforestri', 'produk', 'aren', 'berbasis', 'masyarakat', 'taman', 'nasional', 'meru', 'betiri']
['generalized', 'mixed', 'effect', 'model', 'small', 'area', 'estimation', 'algoritma', 'generalized', 'random', 'forest', 'ana
lisis', 'kerawa', 'pangan']
['pengembangan', 'varietas', 'kentang', 'unggul', 'tipe', 'industri', 'berdaya', 'hasil']
['isolasi', 'coxiella', 'burnetii', 'indonesia', 'teknik', 'propagasi', 'vivo', 'dasar', 'pengembangan', 'alat', 'diagnostik',
'cepat', 'fever']
['karakteristik', 'genetik', 'chlamydia', 'burung', 'paruh', 'bengkok']
['studi', 'patogenesis', 'fever', 'karakteristik', 'agen', 'penyebab', 'coxiella', 'burnetii', 'sapi']
```

Gambar 10. *Stopword* kata umum yang sering muncul

3.7. *Stemming*

Stemming merupakan bagian dari *text processing* atau pembersihan data, bagian ini merupakan bagian dari normalisasi *text* yang bertujuan untuk mengubah teks mentah menjadi format yang bisa dibaca untuk tugas pemroresan Bahasa alami. Inti dari *stemming* ialah mengubah kata yang menggunakan imbuhan

menjadi kata dasar. Berikut tampilan teks setelah melakukan *stemming*.

```
pola distribusi macrobrachium sintangens jawa kalimantan
aplikasi bio nanofertil drone seed sistem agroforestri jelutung rawa dyera polyphylla padi oryza sativa rehabilitasi lahan gambut
aplikasi seed bomb technolog nano cs pmaa npk mikroserat tanaman kehutanan upaya percepatan rehabilitasi lahan gambut
optimasi model hybrid convolut vision transform defect detect extrus base food print
pengembangan sistem agroforestri produk berbasi masyarakat tanan nasion meru betiri
general mix effect model small area estim algoritma general random forest analisi kerawanan pangan
pengembangan varieta kentang unggul tipe industri berdaya hasil
isolasi coxiella burnetii indonesia teknik propagasi vivo dasar pengembangan alat diagnostik cepat q fever
karakteristik genetik chlamydia sp burung paruh bengkok
studi patogenesi q fever karakteristik agen penyebab coxiella burnetii sapi
```

Gambar 11. *Stemming* teks

4. HASIL DAN PEMBAHASAN

Modeling data menggunakan dua Teknik yakni menggunakan metode LDA dan Teknik yang kedua menggunakan *Non-Negative Matrix Factorization (NMF)* yang bertujuan untuk mereduksi data.Model ini dilatih menggunakan 12.000 sampel data. Data tersebut diubah menjadi matriks vektor dengan pembobotan frekuensi kata. Kata-kata yang muncul lebih dari 80% dalam seluruh dokumen akan diabaikan untuk mengurangi *noise*. Selain itu, kata-kata yang muncul kurang dari 2 kali juga akan diabaikan.

4.1. Modeling LDA

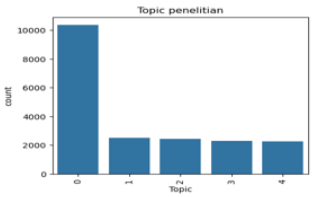
Pemodelan menggunakan LDA menggunakan 5 parameter topik utama yang ada didalam dokumen. Sedang jumlah seed generator yang digunakan sebanyak 42 untuk memastikan hasil yang dihasilkan selama proses iterasi akan sama. Dari hasil model tersebut menghasilkan 5 cluster topik mulai dari topi 0 sampai topik 4. Kata-kata yang ada pada topik tersebut seperti analisis, kinerja, umkm ada pada topik 0, teknologi, sentesis, *learning*, *mechine* ada pada topik 1, pengembangan , berbasis dan model ada pada topik 2, senyawa, aktivitas dan ekstrak ada pada topik 3 dan kata pada topik 4 seperti pengembangan , kemampuan dan pembelajaran. Berikut tampilan 10 top kata yang ada pada masing-masing topik.

```
Top 10 words for topic with LDA #0:
['kota', 'analisis', 'indonesia', 'pengaruh', 'kinerja', 'green', 'berbasis', 'umkm', 'digital', 'model']

Top 10 words for topic with LDA #1:
['teknologi', 'sintesis', 'deteksi', 'learning', 'optimasi', 'metode', 'limbah', 'pengembangan', 'sistem', 'berbasis']

Top 10 words for topic with LDA #2:
['indonesia', 'wisata', 'desa', 'masyarakat', 'berkelanjutan', 'berbasis', 'pengembangan', 'analisis', 'kabupaten', 'model']

Top 10 words for topic with LDA #3:
['uji', 'obat', 'bakteri', 'aktivitas', 'tanaman', 'senyawa', 'ikan', 'daun', 'potensi', 'ekstrak']

Top 10 words for topic with LDA #4:
['literasi', 'kemampuan', 'learning', 'sekolah', 'model', 'siswa', 'meningkatkan', 'pembelajaran', 'berbasis', 'pengembangan']
```

Gambar 12. Sepuluh kata teratas pada setiap topik

Dari jumlah topik yang dihasilkan topic 0 menghasilkan jumlah data yang paling tinggi sebanyak 10389 kata, topik 1 menghasilkan 2517, topik 2 menghasilkan 2432, topik 3 menghasilkan 2286 dan topik 4 menghasilkan 2247. Tampailkan hasil seperti gambar 10.

	Topic	count
0	0	10389
1	1	2517
2	2	2432
3	3	2286
4	4	2247



Gambar 13. Topic LDA

Hasil modeling dengan LDA dibuat menjadi sebuah dataframe menunjukan judul indek nomor 1 menunjukan pada topic 2, indek nomor 2 masuk kedalam topic 3 indek nomor 3 masuk kedalam topik 2 indek nomor 4 masuk kedalam topik 4 dan indek nomor 5 masuk kedalam topik 2. Hasil dalam *datafreame* seperti gambar dibawah.

No	KategoriInInstitusi	Nama Institusi	NIDN	Nama	Judul	RuangInLingkup	Topic
1.0	PTNBH	Institut PertanianInBogor	27046503.0	Achmad Farajallah	pola distribusi macrobrachium sintangens jawa ...	PPS-PTM	2
2.0	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	aplikasi bio ofertil drone seed sistem agrofo...	PFR	3
3.0	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	aplikasi seed bomb technolog o cs pmaa npk mi...	PPS-PTM	2
4.0	PTNBH	Institut PertanianInBogor	2076607.0	Agus Buono	optimasi model hybrid convolut vision transfor...	PPS-PDD	4
5.0	PTNBH	Institut PertanianInBogor	18096208.0	Agus Hikmat	pengembangan sistem agroforestri produk berbas...	PFR	2

Gambar 14. Topik dengan LDA

4.2. Modeling LDA dan NMF

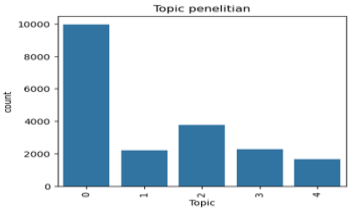
Faktorisasi *Matriks Non-Negatif (NMF)* adalah suatu 3443eknik pembelajaran tanpa pengawasan yang digunakan untuk pengelompokan data dan reduksi dimensi. NMF dikombinasikan dengan LDA untuk membangun model topik data menjadi lima topik dan teks yang dihasilkan atau yang masuk dalam lima topik tersebut seperti gambar 15.

```
Top 10 words for topic #0:
['metode', 'pendekatan', 'deep', 'deteksi', 'machine', 'sistem', 'learning', 'pengembangan', 'berbasis', 'model']

Top 10 words for topic #1:
['green', 'peningkatan', 'peran', 'kota', 'keuangan', 'indonesia', 'analisis', 'kinerja', 'umkm', 'digital']

Top 10 words for topic #2:
['tanaman', 'uji', 'produksi', 'aktivitas', 'senyawa', 'bahan', 'potensi', 'daun', 'limbah', 'ekstrak']

Top 10 words for topic #3:
['strategi', 'ekonomi', 'pariwisata', 'wisata', 'berkelanjutan', 'masyarakat', 'desa', 'kearifan', 'kabupaten', 'lokal']

Top 10 words for topic #4:
['berpikir', 'berbasis', 'dasar', 'literasi', 'kemampuan', 'sekolah', 'pengembangan', 'siswa', 'meningkatkan', 'pembelajaran']
```

Gambar 15. Sepuluh kata tertinggi dengan Teknik NMF

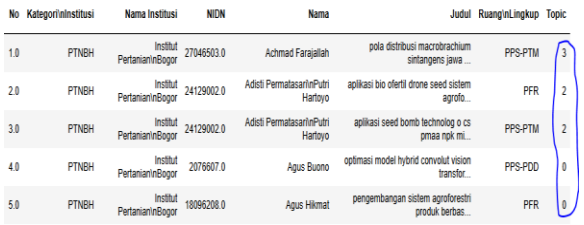
Dari hasil modeling menggunakan Teknik NMF hasil topik 0 menghasilkan data tertinggi sebanyak 9980 data, kemudian topik 2 jumlah data sebanyak 3765, topik 3 menghasilkan 2275, topik 1 menghasilkan 2202 dan topik 4 menghasilkan 1649.

	Topic	count
0	0	9980
2	2	3765
3	3	2275
1	1	2202
4	4	1649



Gambar 16. Topic penelitian dengan NMF

Hasil modeling menggunakan Teknik NMF menunjukan nomor indek 1 masuk kedalam topik 3, indek nomor 2 masuk kedalam topik 2, indek nomor 3 masuk kedalam topik 2, indek nomor 4 masuk kedalam topik 0 dan indek nomor 5 masuk kedalam topik 0.



No	KategoriInInstitusi	Nama Institusi	NIDN	Nama	Judul	RuangInLingkup	Topic
1.0	PTNBH	Institut PertanianInBogor	27046503.0	Achmad Farajallah	pola distribusi macrobrachium sintangens jawa ...	PPS-PTM	3
2.0	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	aplikasi bio ofertil drone seed sistem agrofo...	PFR	2
3.0	PTNBH	Institut PertanianInBogor	24129002.0	Adisti PermatasariInPutri Hartoyo	aplikasi seed bomb technolog o cs pmaa npk mi...	PPS-PTM	2
4.0	PTNBH	Institut PertanianInBogor	2076607.0	Agus Buono	optimasi model hybrid convolut vision transfor...	PPS-PDD	0
5.0	PTNBH	Institut PertanianInBogor	18096208.0	Agus Hikmat	pengembangan sistem agroforestri produk berbas...	PFR	0

Gambar 17. Topic dengan NMF

5. KESIMPULAN DAN SARAN

Berdasarkan hasil pemaparan pada tahapan sebelumnya dari pembahasan menyimpulkan model topik yang menggunakan LDA menghasilkan lima topik. Topik 0 memiliki jumlah kata terbanyak yaitu 10.389 kata, diikuti oleh topik 1 (2.517 kata), topik 2 (2.432 kata), topik 3 (2.286 kata), dan topik 4 (2.247 kata). Model topik yang menggunakan kombinasi LDA dan NMF hasil topik 0 menghasilkan data tertinggi sebanyak 9980 data, kemudian topic 2 jumlah data sebanyak 3765, topik 3 menghasilkan 2275, topik 1 menghasilkan 2202 dan topik 4 menghasilkan 1649. Kedua model yang digunakan menghasilkan jumlah topik yang terlalu banyak pada topik 0, dengan rentang nilai topik yang terlalu tinggi. Untuk penelitian selanjutnya, disarankan untuk menggunakan metode BERTopic dalam membangun model topik.

DAFTAR PUSTAKA

[1] U. Chauhan and A. Shah, "Topic modeling using latent Dirichlet allocation: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 7, pp. 1-35, 2021.

[2] G. G. Priyatna, "Pemodelan Topik Terkait Ulasan Video Game dengan Genre Battle Royale Menggunakan Metode Bertopic dengan Fitur Guided Topic Modelling," Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta, 2023.

[3] D. M. Diannzah, "Analisis Interaksi Pengguna Twitter Menggunakan Social Network Analysis Dan Topic Modelling Terkait Strategi Pemasaran E-Commerce (Studi Kasus: Tokopedia, Shopee, Dan Bukalapak)," 2021.

[4] C. L. Prawirosastro and E. P. A. Akhmad, "Pemodelan Topik Menggunakan Latent Dirichlet Allocation Dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online," Program Diploma Pelayaran, Universitas Hang Tuah, 2022.

[5] N. Novarian, S. Khomsah, and A. B. Arifa, "Topic Modeling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Menggunakan Metode Latent Dirichlet Allocation," *LEDGER: Journal Informatic and Information Technology*, vol. 2, no. 1, pp. 14-27, 2023.

[6] A. Muhaimin *et al.*, "Social Media Analysis and Topic Modeling: Case Study of Stunting in Indonesia," *Telematika: Jurnal Informatika dan Teknologi Informasi*, vol. 20, no. 3, pp. 406-415, 2023.

[7] R. P. F. Afidh, "Pemodelan Topik Menggunakan n-Gram dan Non-negative Matrix Factorization," *Jurnal Informasi dan Teknologi*, pp. 265-275, 2023.

[8] K. Sohn *et al.*, "Visual prompt tuning for generative transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19840-19851.

[9] A. Jahanian, X. Puig, Y. Tian, and P. Isola, "Generative models as a data source for multiview representation learning," *arXiv preprint arXiv:2106.05258*, 2021.

[10] G. A. Khan, J. Hu, T. Li, B. Diallo, and H. Wang, "Multi-view data clustering via non-negative matrix factorization with manifold regularization," *International Journal of Machine Learning and Cybernetics*, pp. 1-13, 2022.

[11] T. Aonishi, R. Maruyama, T. Ito, H. Miyakawa, M. Murayama, and K. Ota, "Imaging data analysis using non-negative matrix factorization," *Neuroscience Research*, vol. 179, pp. 51-56, 2022.

[12] K. Luong, R. Nayak, T. Balasubramaniam, and M. A. Bashar, "Multi-layer manifold learning for deep non-negative matrix factorization-based multi-view clustering," *Pattern Recognition*, vol. 131, p. 108815, 2022.

[13] S. Chandel, C. B. Clement, G. Serrato, and N. Sundaresan, "Training and evaluating a jupyter notebook data science assistant," *arXiv preprint arXiv:2201.12901*, 2022.

[14] W. McKinney, *Python for data analysis*. " O'Reilly Media, Inc.", 2022.

[15] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc): Topic Modeling on Online News Media Using Latent Diriclet Allocation (Case Study Somethinc Brand)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 481-489, 2024.

[16] A. Anisatuzzumara, "Implementasi Latent Dirichlet Allocation (LDA) dan K-Nearest Neighbors (KNN) pada Sistem Rekomendasi Jurnal Terindeks Garuda," Universitas Islam Sultan Agung Semarang, 2024.

[17] M. Jelita, "Text Mining dengan Topic Modelling LDA dari Pertanyaan Gelar Wicara Literasi Perpustakaan Nasional RI," *Media Pustakawan*, vol. 31, no. 3, pp. 253-265, 2024.

[18] Y. N. Bisoumi, J. Munandar, S. Amrullah, M. T. Pandiriyan, K. R. Akmalia, and F. Fauzi, "Papua dalam Perspektif Komentar Youtube: Studi Pemodelan Topik dan Analisis Sentimen dengan Pendekatan Text Mining," in *PROSIDING SEMINAR NASIONAL SAINS DATA*, 2024, vol. 4, no. 1, pp. 270-281.

[19] J. Zimmermann, L. E. Champagne, J. M. Dickens, and B. T. Hazen, "Approaches to improve preprocessing for Latent Dirichlet Allocation topic modeling," *Decision Support Systems*, vol. 185, p. 114310, 2024.

[20] J. Supriyanto, D. Alita, and A. R. Isnain, "Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring," *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74-80, 2023.

[21] L. Legito *et al.*, *Buku Ajar Pengantar Ilmu Komputer*. PT. Sonpedia Publishing Indonesia, 2023.

[22] A. Nur and P. R. Cahyani, "Evaluasi Pengguna Jupyter Notebook Pada Python Dalam Pembelajaran Data Science (Studi Kasus: Kapal Titanic)," *Kohesi: Jurnal Sains dan Teknologi*, vol. 4, no. 10, pp. 21-30, 2024.

[23] T. M. Fahrudin and M. S ST, *Algoritma dan Pemrograman Dasar dalam Bahasa Pemrograman Python*. Thalibul Ilmi Publishing & Education, 2023.