LAPORAN HASIL PENELITIAN

PEMODELAN TOPIK MENGGUNAKAN LATENT DIRICHLET ALLOCATION DAN PACHINKO ALLOCATION MODEL UNTUK EKSTRAKSI BERITA SAHAM ONLINE

Oleh:

Ekka Pujo Ariesanto Akhmad
NIDN. 0724037402

Carlos L. Prawirosastro
NIDN. 0710078302

DIBIAYAI OLEH
UNIVERSITAS HANG TUAH

PROGRAM DIPLOMA PELAYARAN
UNIVERSITAS HANG TUAH
SURABAYA
2021

HALAMAN PENGESAHAN
HASIL PENELITIAN DOSEN

1. Diajukan kepada	:	Rektor c.,q. Ketua Lembaga Penelitian dan Pengabdian Masyarakat Universitas Hang Tuah Surabaya
a. Judul Penelitian	:	Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online
b. Kode/Nama Rumpun Ilmu	:	462/Teknologi Informasi
2. Ketua Peneliti:		
a. Nama lengkap dan gelar	:	Ekka Pujo Ariesanto Akhmad, S.E., M.Kom.
b. NIK/NIDN	:	01197/ 0724037402
c. Jabatan fungsional	:	Lektor
d. Fakultas/Jurusan	:	Program Diploma Pelayaran/KPN
3. Susunan Tim Peneliti		
Anggota	:	Carlos L. Prawirosastro, S.Pd.I., M.Pd.I.
4. Lokasi penelitian	:	kontan.co.id, Jakarta
5. Biaya Penelitian	:	Rp9.500.000,00

Surabaya, 25 Agustus 2021

Menyetujui,
Direktur PDP

Djamaludin Malik, S.E., M.AP.
NIK. 02581

Peneliti,

Ekka Pujo Ariesanto A., S.E., M.Kom.
NIK. 01197

Mengetahui,
Ka. LPPM

Dr. Ir. Ninis Trisyani, M.P.
NIK. 01071

PERNYATAAN BEBAS PLAGIAT

Yang bertanda tangan di bawah ini:

Nama :Ekka Pujo Ariesanto Akhmad

NIDN :0724037402

Jurusan/Fakultas :KPN/Program Diploma Pelayaran

Dengan ini menyatakan bahwa ~~proposal penelitian~~/hasil penelitianyang berjudul:

Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online

Adalah orisinil, bebas plagiat, semua sumber baik yang dikutip maupun yang dirujuk telah saya nyatakan dengan benar.

Apabila di kemudian hari terbukti terdapat plagiat dalam ~~proposal penelitian~~/hasil penelitian saya, maka saya bersedia dituntut dan diproses sesuai dengan ketentuan perundang-undangan yang berlaku.

Demikian pernyataan ini saya buat dengan sesungguhnya dan dengan sebenar-benarnya.

Surabaya, 25 Agustus 2021

Peneliti

Ekka Pujo Ariesanto Akhmad
NIDN.0724037402

ABSTRAK

Pada umumnya investor memperoleh informasi atau berita saham lewat situs resmi Bursa Efek Indonesia, yakni www.idx.co.id. Investor kadang mendapatkan informasi lain tentang analisis saham dan prediksi saham yang menguntungkan dari halaman web berita online. Namun, investor memerlukan waktu untuk menentukan topik yang paling sering muncul dan menjadi perbincangan hangat pada berita saham. Oleh karena itu, pemodelan topik diperlukan untuk mengekstrak berita saham online ke dalam topik-topik yang muncul dari hasil pemodelan. Tujuan penelitian ini adalah pemodelan topik dilakukan untuk menganalisis topik-topik yang sedang dibahas pada halaman web berita saham online dan PAM akan diterapkan untuk menggambarkan korelasi topik LDA yang berhubungan, sehingga topik mempunyai korelasi lebih sesuai. Data penelitian yang dikumpulkan sebanyak 181 berita saham selama bulan Februari hingga Juli 2021. Situs kontan.co.id dipilih karena link berita saham sudah jadi satu dengan ringkasan saham Google dari perusahaan yang ada di Bursa Efek Indonesia.Pemodelan topik dilakukan dengan metode Latent Dirichlet Allocation (LDA), sebuah metode text mining untuk menemukan pola tertentu pada sebuah dokumen dengan menghasilkan beberapa macam topik yang berbeda.Setelah luaran topik LDA diperoleh, langkah berikutnya mengerjakan model alokasi pachinko (Pachinko Allocation Model) untuk menunjukkan hubungan yang koheren antara topik yang dihasilkan.

Kata kunci: berita saham, pemodelan topik, latent dirichlet allocation, pachinko allocation model

ABSTRACT

In general, investors obtain information or stock news through the official website of the Indonesia Stock Exchange, namely www.idx.co.id. Investors sometimes get other information about stock analysis and profitable stock predictions from online news web pages. However, investors need time to determine the topics that appear most often and become hot topics in stock news. Therefore, topic modeling is needed to extract online stock news into topics that emerge from the modeling results. The purpose of this study is that topic modeling is carried out to analyze the topics being discussed on the online stock news web page and PAM will be applied to describe the correlation of related LDA topics, so that the topics have a more appropriate correlation. The research data collected was 181 stock news during February to July 2021. The kontan.co.id site was chosen because the stock news link was integrated with Google stock summaries of companies listed on the Indonesia Stock Exchange. The topic modeling was carried out using the Latent Dirichlet Allocation method. (LDA), a text mining method to find certain patterns in a document by generating several different topics. After the LDA topic output is obtained, the next step is to work on a pachinko allocation model (Pachinko Allocation Model) to show a coherent relationship between the generated topics.

Keywords: stock news, topic modeling, latent dirichlet allocation, pachinko allocation model

KATA PENGANTAR

Puji syukur kepada Allah SWT karena hasil penelitian yang berjudul "Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online" telah selesai.

Tujuan penelitian ini adalah melakukan pemodelan topikuntuk menganalisis topik-topik yang sedang dibahas dan mengetahui korelasi antar topik yang menjadi luaran pemodelan topik halaman web berita saham online.

Bersama ini disampaikan ucapan terima kasih atas berbagai bantuan kepada:

1. Rektor Universitas Hang Tuah Surabaya.
2. Ketua Lembaga Penelitian dan Pengabdian Kepada Masyarakat Universitas Hang Tuah Surabaya.
3. Direktur Program Diploma Pelayaran Universitas Hang Tuah Surabaya.
4. Wakil Direktur Program Diploma Pelayaran Universitas Hang Tuah Surabaya..
5 Nurul Rosana, S.Pi., M.T. dan Muh. Taufiqurrohman, S.T., M.T. selaku reviewer penelitian ini.
6. Rekan dosen, pegawai, dan mahasiswa Program Diploma Pelayaran (PDP) Universitas Hang Tuah Surabaya.

Akhirnya semoga hasil penelitian ini bermanfaat.

Surabaya, 25 Agustus 2021

Peneliti,

Ekka Pujo A. A., S.E., M.Kom.

DAFTAR ISI

Halaman

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pergerakan harga di bursa saham penting bagi investor (Avellaneda, 2008). Satu dari sumber informasi saham adalah data berita tekstual (Mittermayer, 2004). Pada umumnya investor memperoleh informasi atau berita saham lewat situs resmi Bursa Efek Indonesia, yakni www.idx.co.id. Investor kadang mendapatkan informasi lain tentang analisis saham dan prediksi saham yang menguntungkan dari halaman web berita online.Data berita tekstual dikaitkan dengan data pasar saham adalah metode yang populer untuk mempelajari perilaku harga (Li, et. al, 2014). Namun, investor memerlukan waktu untuk menentukan topik yang paling sering muncul dan menjadi perbincangan hangat pada berita tentang saham.Karena itu, penelitian ini akan melakukan pemodelan topik artikel berita tentang saham yang diterbitkan oleh halaman web berita online.Pemodelan topik (Topic Modelling) merupakan model statistik untuk menentukan pokok dari topik dari sekumpulan dokumen (Bansal, 2020). Pemodelan topik digunakan untuk mengetahui tren topik saham yang menjadi perhatian di masyarakat, hingga informasi saham menjadi lebih ringkas.

Metodepemodelan topik yang diusulkan antara lain, klasifikasi teks berita dengan menggunakan metode Support Vector Machine (SVM) dan TF-IDF untuk ekstraksi fitur (Dadgar, dkk., 2016).SVM dapat memecahkan masalah over-fitting, lambatnya konvergensi, dan sedikitnya data training (Vapnik, 2005).Tetapi

SVM memiliki kelemahan pada sulitnya pemilihan parameter SVM yang optimal (Coussement, 2008).

Metode Latent Semantic Analysis (LSA) adalah model statistik untuk menentukan hubungan antara kumpulan dokumen dan istilah yang ada dalam dokumen dengan cara mendapatkan hubungan semantik antara kata-kata tersebut. LSA memiliki kelebihan, yakni mudah diimplementasikan, dipahami dan digunakan,kinerja: LSA mampu memastikan hasil yang layak, jauh lebih baik daripada model ruang vektor biasa. LSA dapat menangani masalah sinonim sampai batas tertentu (tergantung pada dataset), lebih cepat dibandingkan dengan model reduksi dimensi lainnya, tidak sensitif terhadap kondisi awal (seperti jaringan saraf) jadi konsisten (Balu, 2013).

Namun, LSA juga mempunyai kelemahan, karena LSA adalah model distribusi, jadi bukan representasi yang efisien, jika dibandingkan dengan metode canggih (katakanlah jaringan saraf yang dalam), representasi padat, sehingga sulit untuk diindeks berdasarkan dimensi indvidual. LSA adalah model linier, jadi bukan solusi terbaik untuk menangani dependensi non linier, dimensi topik laten tidak dapat dipilih untuk sembarang angka. Itu tergantung pada peringkat matriks, jadi tidak bisa lebih dari itu, model tidak dapat dibaca secara manusiawi,debug/evaluasi dimungkinkan dengan menemukan kata-kata yang mirip untuk setiap kata di ruang laten,tetapi sebaliknya tidak mudah untuk diartikan seperti kata LDA. Penentuan jumlah topik didasarkan pada heuristik dan membutuhkan keahlian (Balu, 2013).

Metode Latent Dirichlet Allocation (LDA) diusulkan oleh Blei(2012) untuk pemodelan topik. Algoritma LDA merupakan algoritma pemodelan topik dengan model probabilitas generatif pada koleksi dokumen. Tujuannya agar pemrosesan dokumen dalam koleksi data besar menjadi efisien. Metode ini membuat dokumen menghasilkan luaran berupa berbagai jenis topik yang berbeda, sehingga tidak secara spesifik mengelompokkan dokumen ke dalam sebuah topik tertentu.

LDA mempunyai kelemahan, yaitu ada batasan jumlah topik yang dapat dihasilkan, LDA tidak dapat menggambarkan korelasi yang menyebabkan terjadinya topik yang tidak berhubungan, tidak ada perkembangan topik dari waktu ke waktu, LDA mengasumsikan kata-kata dapat dipertukarkan, struktur kalimat tidak dimodelkan, dan tanpa pengawasan (terkadang diperlukan sedikit pengawasan , misalnya dalam analisis sentimen) (Bansal, 2020).

Pachinko Allocation Model (PAM) merupakan metode perbaikan dari model Latent Dirichlet Allocation. Model LDA memunculkan korelasi antar kata dengan mengidentifikasi topik berdasarkan hubungan tematik antar kata yang ada dalam korpus. Tapi PAM berimprovisasi dengan memodelkan korelasi antara topik yang dihasilkan. Model ini memiliki kekuatan yang lebih besar dalam menentukan hubungan semantik dengan tepat, karena model ini juga memperhitungkan hubungan antar topik. Model tersebut dinamai Pachinko, permainan populer di Jepang. Model ini menggunakan Grafik Asiklik Terarah (Directed Acyclic Graph/DAG) untuk memahami korelasi antar topik. DAG

adalah grafik berarah terbatas untuk menunjukkan bagaimana topik terkait(Manthiramoorthi, 2020).

Pada penelitian ini PAM akan diterapkan untuk menggambarkan korelasi topik LDA yang berhubungan, sehingga topik mempunyai korelasi lebih sesuai.

1.2. Rumusan Masalah

Uraian latar belakang masalah tersebut dapat diturunkan rumusan masalah yang akan dicari jawabannya melalui penelitian ini, yaitu

a.Bagaimana melakukan pemodelan topikuntuk menganalisis topik-topik yang sedang dibahas pada halaman web berita saham online ?

b. Bagaimana korelasiantar topik yang menjadi luaran pemodelan topik halaman web berita saham online ?

1.3. Tujuan Penelitian

Tujuan penelitian berikut ini merupakan uraian hasil yang akan dicapai melalui penelitian, yaitu

a. Pemodelan topikdilakukan untuk menganalisis topik-topik yang sedang dibahas pada halaman web berita saham online.

b.Korelasi antar topik diterapkan untuk luaran pemodelan topik halaman web berita saham online.

1.4 Urgensi (keutamaan) Penelitian

a. Memudahkan masyarakat atau investor menemukan topik berita saham, sehingga mengurangi waktu membaca dan memahami suatu dokumen.

b. Mengetahui kemampuan algoritma Pachinko Allocation Model dalam menggambarkan korelasi antar topik, sehingga proses pemilihan topik yang relevan menjadi lebih mudah.

1.5 Rencana Target Luaran Penelitian

Tabel 1.1 menjelaskan luaran yang ditargetkan dan lamanya penelitian yang akan dilakukan.

Tabel 1.1
Rencana Target Luaran Penelitian

No	Jenis Luaran			
	Kategori	**Sub Kategori**	**Wajib**	**Tambahan**
1	Artikelilmiahdimuat di jurnal	Internasional		
		NasionalTerakreditasi		
		Nasional ber OJS	✓	
2	Artikelilmiahdimuat di prosiding	Internasional		
		Nasional		✓
3	Buku	Monograf		
		Buku Ajar		
4	Lain-lain			

Sumber: Buku Panduan Pelaksanaan Penelitian Universitas Hang Tuah 2018

Rencana artikel akan dipublikasikan di Jurnal Manajemen Informatika, Universitas Komputer Indonesia, Bandung atau Jurnal Informatika, Universitas BSI, Jakarta.

Rencana tambahan artikel ilmiah akan dipublikasikan di prosiding Seminar Nasional Teknologi Informasi, UII, Yogyakarta.

BAB II

KAJIAN PUSTAKA

2.1 Penelitian Terdahulu

Tabel 2.1 memuat daftar penelitian terdahulu yang mendasari penelitian ini.

Penulis	Judul	Metode	Hasil
Setijohatmo, Urip T., d.k.k.	Analisis Metoda *Latent Dirichlet Allocation* untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik	*Latent Dirichlet Allocation*	Penelitian ini akan menggunakan Perluasan PLSA dari pendekatan lain yang disebut LDA (*Latent Dirichlet Allocation*), spesifiknya menggunakan algoritma *Gibbs Sampling*, dan dilakukan pada studi kasus pencarian dokumen laporan tugas akhir. Eksperimen menggunakan sekumpulan laporan tugas akhir yang telah diberi label. Selanjutnya hasil eksperimen akan diukur tingkat kerelevanannya jika dibandingkan dengan *judgement* manusia dalam bentuk laporan tugas akhir berlabel
Putra, I Made Kusnanta Bramantya	Analisis topik informasi publik media sosial di surabaya menggunakan pemodelan *latent dirichlet allocation (LDA)*	*Latent Dirichlet Allocation*	Eksperimen pemodelan topik dengan metode LDA menyimpulkan bahwa jumlah topik yang terdapat dalam pesan media sosial adalah 4 topik. Hasil eksperimen ini telah diuji secara mesin dengan nilai perplexity terbaik sebesar 213.41 dan diuji kemudahannya untuk diinterpretasi oleh manusia melalui uji koherensi topik yang terdiri dari Word Intrusion task dan Topic Intrusion Task. Kesimpulan dari uji koherensi topik menyatakan bahwa model yang dihasilkan dengan metode LDA pada studi kasus ini dapat diinterpretasi manusia dengan baik.
Zhou Tong dan Haiyi Zhang	*A Text mining Research Based on LDA Topic Modelling*	*Latent Dirichlet Allocation Topic Modelling*	Paper ini memperkenalkan *Text mining* dengan *LDA Topic Modelling*sebagai metodenya, dimana eksperimen dilakukan pada dua tipe dokumen, yaitu artikel *Wikipedia* dan *tweet* dari pengguna *Twitter*. Garis besar penelitian ini membahas tentang gambaran umum *text mining* dengan metode *LDA*, *pre-processing*, *model training* dan hasil analisis.
Ajai Gaur	*Topic Models as A Novel Approach To Identify Themes in Content Analysis: The Example of Organizational Research Methods*	*Latent Dirichlet Allocation (LDA)*	Penelitian ini mendemonstrasikan *LDA Topic Modelling*sebagai metode baru dalam analisis data dalam bentuk teks. Adapun data teks yang dianalisis adalah 421 artikel yang dipublikasikan pada *Organization Research Method (ORM)* yang berhasil mengungkapkan 15 topik, bahwa hasil analisis tersebut cukup sesuai dengan hasil kajian oleh manusia.
Jey Han	*On-line Trend*	*Online Variant of*	Penelitian ini mengusulkan pendekatan

Lau, et. al	*Analysis with Topik Models: #Twitter trends detection topik model online*	*Latent Dirichlet Allocation (LDA) Topic modeling*	berbasis topik model *on-line* untuk menganalisis tren. Secara umum, pada setiap *update*, metode ini menghitung evolusi topik untuk mendeteksi topik baru yang muncul dalam koleksi dokumen. Metodologi ini dapat menunjukkan kekuatan dari model dalam mendeteksi dokumen-dokumen individual dengan menggambarkan tren saat ini, dan terus bergerak.
Radim Rehurek dan Petr Sojka	*Software Framework for Topic Modelling with Large Corpora*	*Vector Space Model (VSM)* yaitu *Latent Semantic Analysis (LSA)* dan *Latent Dirichlet Allocation (LDA)* dengan *framework Gensim* dan Bahasa *Python*	*Vector Space Model (VSM)* adalah paradigma dalam *modelling* yang terbukti dan ampuh dalam *Natural Language Processing*, dimana dokumen direpresentasikan sebagai vektor dalam ruang berdimensi tinggi. Metode yang termasuk dalam *VSM* cukup beragam, sehingga algoritma yang diterapkan juga beragam. Dengan demikian, diperlukan suatu *framework* untuk memberikan arahan sebagai solusi dari adanya *practical gap* antara model matematis, algoritma dan *Source code*. Penelitian ini mengajukan adanya *framework* yang mencakup aspek *Corpus size independence, Intuitive API, Easy deployment, Cover popular algorithms* dengan menggunakan bahasa *Python*

Sumber: Setijohatmo (2020), Putra (2017), Tong dan Zhang (2016), Gaur (2015), Lau, et.al (2012), Rehurek dan Sojka (2010)

Persamaan peneltian terdahulu dengan penelitian yang dilakukan penulis sekarang, yaitu sama-sama menggunakan metode Latent Dirichlet Allocation untuk pemodelan topik.

Perbedaan penelitian terdahulu dengan penelitian yang dilakukan penulis sekarang adalah menggunakan Pachinko Allocation Model untuk menunjukkan hubungan yang koheren antara topik yang dihasilkan LDA.

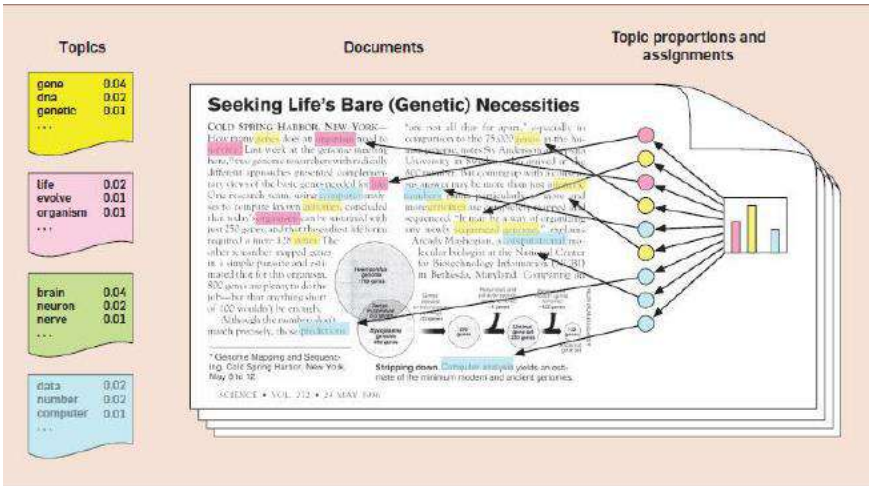
2.2 Dasar Teori

2.2.1. Topic Modelling

Konsep *Topic Modelling*menurut Blei terdiri dari entitas-entitas, yaitu "kata","dokumen", dan "*corpora*". "Kata" dianggap sebagai unit dasar dari data diskrit dalam dokumen, didefinisikan sebagai item dari kosa kata yang diberi indeks untuk setiap kata unik pada dokumen. "Dokumen" adalah susunan N kata-kata.

Sebuah *corpus* adalah kumpulan M dokumen dan *corpora* merupakan bentuk jamak dari *corpus*. Sementara "topik" adalah distribusi dari beberapa kosakata yang bersifat tetap. Secara sederhana, setiap dokumen dalam *corpus* mengandung proporsi tersendiri dari topik-topik yang dibahas sesuai kata-kata yang terkandung di dalamnya (Blei, 2003).

Ide dasar dari *Topic Modelling*adalah bahwa sebuah topik terdiri dari kata-kata tertentu yang menyusun topik tersebut, dan dalam satu dokumen memiliki kemungkinan terdiri dari beberapa topik dengan probabilitas masing-masing. Namun secara pemahaman manusia, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen, dan penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi, maka dari itu *Topic Modelling*bertujuan untuk menemukan topik dan kata-kata yang terdapat pada topik tersebut (Blei, 2012). Konsep *Topic Modelling*menurut Blei, ditunjukkan pada Gambar 2.1.



Gambar 2.1 Konsep Topic Modelling (Blei, 2012)

Menurut Blei (2012), *Topic Modelling*merupakan rangkaian algoritma yang bertujuan untuk menemukan dan memberikan keterangan pada suatu arsip besar dokumen dengan informasi tematik, yaitu pembelajaran tepadu yang menggunakan tema untuk mengaitkan beberapa tema dengan entitas yang dikaitkan. Secara sederhana, *Topic Modelling*merupakan algoritma yang bertujuan untuk menemukan topik yang tersembunyi dari rangkaian kata dalam dokumen yang tidak terstruktur. Algoritma *Topic Modelling*menganalisis kata-kata dari teks asli untuk menemukan topik yang berada diantara teks tersebut, bagaimana topik topik saling terhubung satu sama lain, dan bagaimana tema-tema tersebut dapat berubah dari waktu ke waktu, sehingga dapat dikembangkan untuk pencarian, ataupun meringkas teks yang terdapat dalam dokumen. (Blei, 2003).

Pendapat lain terkait *Topic Modelling*disampaikan oleh Megan R. Brett (2012), yang menyatakan bahwa *Topic Modelling*merupakan bentuk *text mining*, sebagai salah satu metode untuk mengidentifikasi pola dalam sebuah *corpus*. *Topic Modelling*dapat pula dikatakan sebagai sebuah tool untuk mengubah *corpus* yang berbentuk kumpulan kata, menjadi topik yang dapat menggambarkan *corpus* tersebut.

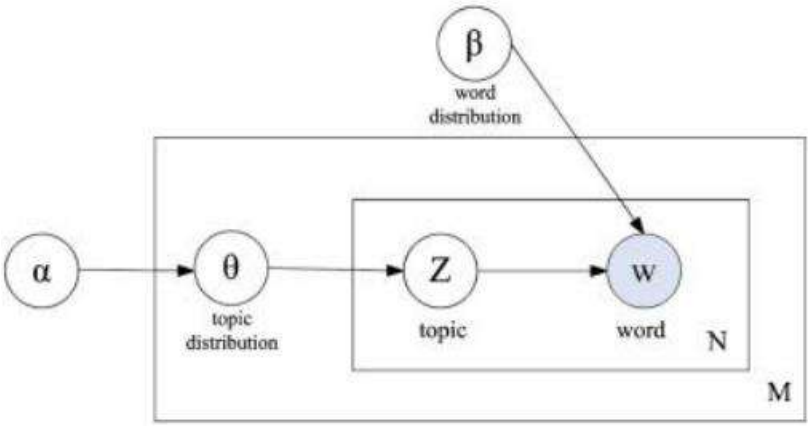
Algoritma pemodelan topik tidak memerlukan penjelasan sebelumnya atau pelabelan karena topik dokumen muncul dari analisis teks asli. Pada umumnya, *Topic Modelling*diaplikasikan pada jumlah dokumen yang sangat besar dan dapat mengelola dokumen-dokumen individual sesuai topik yang ditemukan, sehingga *Topic Modelling*memungkinkan kita untuk mengatur dan meringkas arsip elektronik pada skala yang tidak mungkin dengan penjelasan manusia secara

manual (Blei, 2012).

2.2.2. *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) merupakan metode Topic Modelling dan topik analisis yang paling populer saat ini. *LDA* muncul sebagai salah satu metode yang dipilih dalam melakukan analisis pada dokumen yang berukuran sangat besar. *LDA* dapat digunakan untuk meringkas, melakukan klasterisasi, menghubungkan maupun memproses data yang sangat besar karena *LDA* menghasilkan daftar topik yang diberi bobot untuk masing-masing dokumen (Campbell, et.al, 2014). Adapun distribusi yang digunakan untuk mendapatkan distribusi topik per-dokumen disebut distribusi Dirichlet, kemudian dalam proses generatif untuk *LDA*, hasil dari Dirichlet digunakan untuk mengalokasikan kata-kata pada dokumen untuk topik yang berbeda. Dalam *LDA*, dokumen-dokumen merupakan objek yang dapat diamati, sedangkan topik, distribusi topik per-dokumen, penggolongan setiap kata pada topik per-dokumen merupakan struktur tersembunyi, maka dari itu, Algoritma ini dinamakan *Latent Dirichlet Allocation* (*LDA*) (Blei, 2012).

Menurut Blei (2003), *LDA* merupakan model probabilistik generatif dari kumpulan tulisan yang disebut *corpus*. Ide dasar yang diusulkan metode *LDA* adalah setiap dokumen direpresentasikan sebagai campuran acak atas topik yang tersembunyi, yang mana setiap topik memiliki karakter yang ditentukan berdasarkan distribusi kata-kata yang terdapat di dalamnya. Blei merepresentasikan metode *LDA* sebagai model *probabilistic* secara visual seperti pada Gambar 2.2



Gambar 2.2 Metode *LDA* (Blei, 2012)

Sesuai visualisasi model di atas, terdapat tiga tingkatan pada *LDA* Modelling. Parameter α dan β merupakan parameter distribusi topik yang berada pada tingkatan *corpus*, yaitu kumpulan dari M dokumen. Parameter α digunakan dalam menentukan distribusi topik dalam dokumen, semakin besar nilai alpha dalam suatu dokumen, menandakan campuran topik yang dibahas dalam dokumen semakin banyak. Parameter β digunakan untuk menentukan distribusi kata dalam topik. Semakin tinggi nilai beta, maka semakin banyak kata-kata yang ada di dalam topik, sedangkan semakin kecil nilai beta, maka semakin sedikit kata-kata yang ada di dalam topik sehingga topik tersebut mengandung kata-kata yang lebih spesifik. Variabel θm adalah variabel yang berada di tingkat dokumen (M). Variabel θ merepresentasikan distribusi topik untuk dokumen tertentu. Semakin tinggi nilai θ, maka semakin banyak topik yang ada di dalam dokumen, sedangkan semakin kecil nilai θ, maka dapat dikatakan dokumen tersebut semakin spesifik pada topik tertentu. Variabel Zn dan Wn adalah variabel tingkat kata (N). Variabel Z dan merepresentasikan topik dari kata tertentu pada sebuah dokumen sdangkan

variabel Wmerepresentasikan kata yang berkaitan dengan topik tertentu yang terdapat dalam dokumen (Blei, 2003)

Secara umum, *LDA* bekerja dengan masukan dokumen-dokumen individual dan beberapa parameter, untuk menghasilkan luaran berupa model yang terdiri dari bobot yang dapat dinormalisasi sesuai probabilitas. Probabilitas ini mengacu pada dua jenis, yaitu jenis (a) probabilitas bahwa suatu dokumen spesifik tertentu menghasilkan topik yang spesifik pula dan jenis (b) probabilitas bahwa topik spesifik tertentu menghasilkan kata-kata spesifik dari sebuah kumpulan kosakata. Probabilitas jenis (a), dokumen yang sudah diberi label dengan daftar topik seringkali dilanjutkan hingga menghasilkan probabilitas jenis (b), yang menghasilkan kata-kata spesifik tertentu (Blei, 2012).

LDA dapat digunakan dalam bidang analisis trend pada media sosial (Jey Han Lau, 2012), melakukan identifikasi tema pada 421 artikel ilmiah pada Organizational Research Methods (Gaur, 2015), mendeteksi topik untuk pelacakan konten percakapan (Yeh, et.al, 2016) dan telah terbukti mampu bekerja dengan baik untuk dokumen panjang seperti artikel *Wikipedia* maupun dokumen pendek seperti *tweet* (Tong, 2016)

2.2.3. Validasi Topik dengan *Perplexity* dan *Topic Coherence*

*Topic Modelling*mempelajari kumpulan dari kata-kata dari sebuah dokumen maupun *corpus* tanpa supervisi manusia (*unsupervised*). Berdasarkan kata-kata yang digunakan yang terdapat dalam dokumen, penggalian relasi topik dilakukan dengan asumsi bahwa satu dokumen mencakup suatu set kecil dari

topik yang ringkas, yang mana topik-topik ini perlu dikorelasikan dengan interpretasi manusia (Stevens, 2012).

Dalam penelitian ini, akan diterapkan dua metode untuk melakukan validasi topik, yaitu *Perplexity* dan *Topic Coherence.*

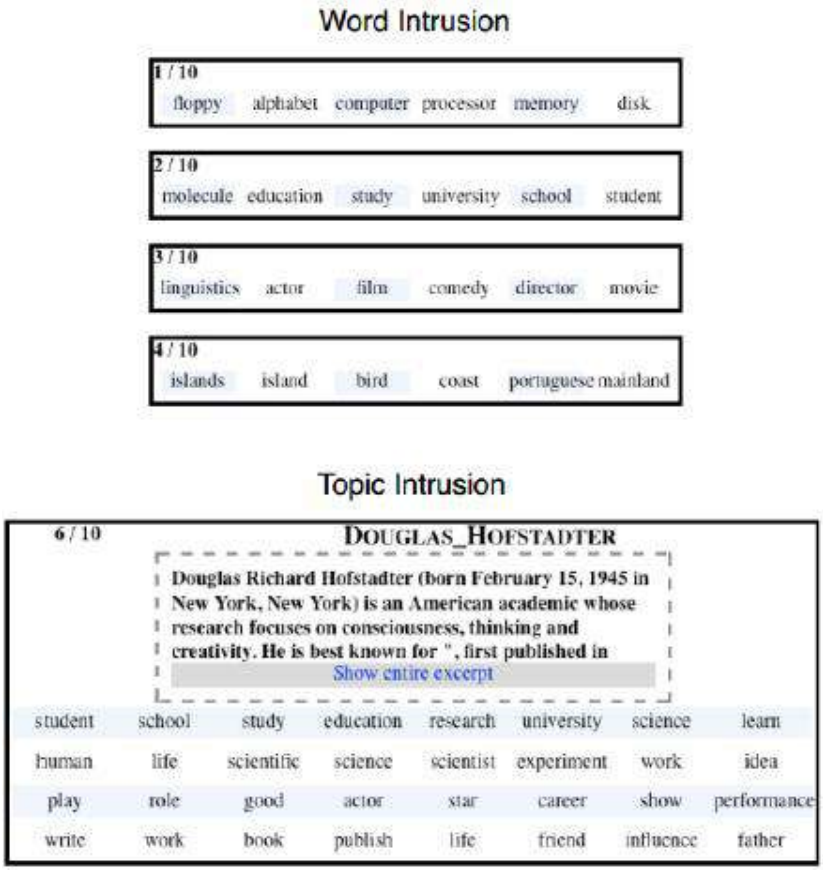
Perplexity menjadi ukuran kualitas standar untuk model topik. *Perplexity* mengukur kemampuan model topik untuk menggeneralisasi dokumen setelah memperkirakan model menggunakan dokumen pelatihan. *Perplexity* yang lebih rendah berarti kemampuan generalisasi yang lebih baik. Penelitian lain menyebutkan *Perplexity* merupakan ukuran kinerja pemodelan bahasa berdasarkan probabilitas rata-rata yang dikembangkan dalam bidang teori informasi (Muslich, 2010).

Untuk menghitung *Perplexity*, diperlukan pemodelan bahasa dan teks yang akan diujikan. *Perplexity* dapat pula digunakan untuk membandingkan berbagai jenis pemodelan Bahasa, namun untuk perbandingan ini, harus menggunakan teks yang sama sebagai teks yang diujikan. Ukuran *vocabulary* dapat dengan mudah terlihat sebagai relevansi dengan *Perplexity* karena dengan berkurangnya kardinalitas dapat secara langsung menurunkan jumlah kemungkinan kata (Muslich, 2010).

Chang, dkk mengusulkan bentuk evaluasi topik yang lebih menonjolkan sisi kemudahan dalam interpretasi oleh manusia. Adalah *Topic Coherence*, dimana satu set kata-kata yang dihasilkan oleh topik model dinilai berdasarkan tingkat koherensi atau tingkat kemudahannya dalam diinterpretasi manusia (Newman, 2010). *Topic Coherence* mengukur nilai suatu topik dengan mengukur

tingkat kesamaan semantik antara kata-kata yang terdapat dalam topik. Pengukuran ini membantu membedakan antara topik yang dapat diinterpretasi secara semantik dengan topik yang memiliki keterkaitan secara statistik (Stevens, 2010). Chang, dkk, dalam penelitiannya tahun 2009 menawarkan task-task untuk mengukur kesuksesan interpretasi topik hasil *Topic Modelling*, yaitu *Word Intrusion Task* dan *Topic Intrusion Task* (Chang, 2009). Untuk memudahkan dalam memahami *word intrusion* dan *topic intrusion* dapat dilihat pada Gambar 2.3.

Berdasarkan topik-topik yang sudah dihasilkan melalui Topic Modelling, Word Intrusion Task bekerja dengan memilih salah satu dari beberapa topik yang terdapat dalam dokumen dipilih secara acak, kemudian diantara seluruh kata-kata yang terdapat dalam topik disisipkan kata yang memiliki probabilitas rendah pada topik tersebut. Pada akhirnya, pengujian dilakukan dengan cara meminta beberapa responden untuk menebak manakah kata yang tidak termasuk dalam topik tersebut. Topic Intrusion Task menguji apakah topik yang dihasilkan dari dokumen melalui Topic Modelling sesuai dengan topik yang dihasilkan dari pandangan manusia terhadap dokumen yang sama. Pengujian kesesuaian topik dilakukan dengan menyajikan cuplikan dokumen yang disertai beberapa pilihan topik kepada beberapa responden, yang mana satu dari beberapa topik tersebut merupakan topik yang memiliki probabilitas rendah dalam dokumen (Chang, 2009).



Gambar 2.3 Word Intrusion Task dan Topic Intrusion Task (Chang, 2009)

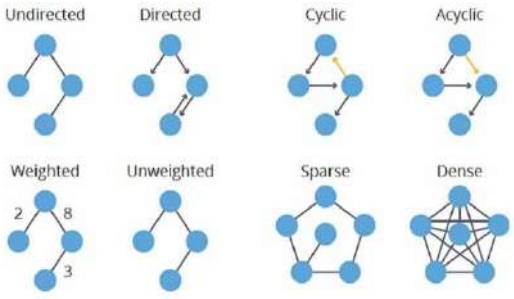
2.3 Pachinko Allocation Model

Teknik Pemodelan Topik Populer Seperti Alokasi Dirichlet Laten, Indeks Semantik Laten, Faktorisasi Matriks Non-Negatif, dll. Semua model ini sangat kuat dalam memperoleh hubungan semantik antara kata-kata yang ada dalam dokumen. Tapi mereka tidak menunjukkan hubungan yang koheren antara topik yang dihasilkan. Jadi dibutuhkan model yang lebih kuat untuk mencapai itu. Ini sangat diperlukan untuk teknik peringkasan teks. Terutama dalam kasus metode peringkasan teks berbasis Abstraksi, fitur ini akan sangat diperlukan. Saat ini sudahbanyak aplikasi berita yang mengimplementasikan teknik ini yang

menggunakan metode ringkasan teks berbasis abstraksi untuk menghasilkan ringkasan berita.

Hubungan semantik adalah asosiasi yang ada antara arti kata atau phrase atau kalimat. Hubungan yang koheren ditandai dengan hubungan kata-kata yang konsisten secara teratur, logis, gramatis dan estetis dalam kalimat.Ketidakmampuan untuk mengekstrak hubungan antar topik menimbulkan batasan yang sangat besar dari metode LDA. Karena dalam bagian yang terus menerus, garis berikutnya akan memiliki koherensi tertentu dengan garis sebelumnya. Jadi, sangat penting untuk mendapatkan koherensi yang erat di antara bagian-bagian untuk mendapatkan topik yang tepat. Kesulitan ini diatasi dengan menggunakan Model Alokasi Pachinko (Pachinko Allocation Model (PAM)) (Manthiramoorthi, 2020).

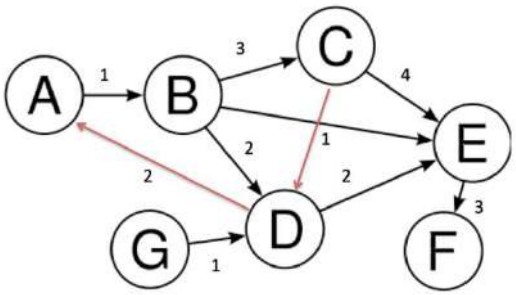
PAM menangkap korelasi berubah-ubah, bersarang, dan bahkan jarang antara topik menggunakan Grafik Asiklik Terarah (Directed Acyclic Graph). Daftar semua kata yang diperoleh dari korpus setelah menghapus stopwords dan pemrosesan teks mewakili distribusi dirichlet. Dalam PAM, setiap topik yang dihasilkan dikaitkan dengan distribusi dirichlet melalui Directed Acyclic Graph.



Sumber: https://iq.opengenus.org/pachinko-allocation-model

Gambar 2.4 Jenis Graf Asiklik Terarah

Graf Asiklik Terarah adalah graf berarah berhingga non siklik. Gambar2.4 menjelaskan jenis jaringan grafik DAG.Gambar 2.5 menjelaskan dua komponen graf asiklik terarah, yaitu node dan link.



Sumber: https://iq.opengenus.org/pachinko-allocation-model

Gambar 2.5 Node dan Link Graf Asiklik Terarah

1. Simpul (Node)

Topik-topik yang akan dihubungkan bersama disebut node. Oleh karena itu, setiap node mewakili sebuah topik. Atribut topik mewakili kata-kata yang berhubungan dengan topik tertentu.

2. Tautan (Link)

Tautan mewakili koneksi antar node. Dalam model, tautan bersifat asiklik dan diarahkan. Dengan demikian muncullah nama.

Fungsi Graf Asiklik Terarah

Dalam model ini, setiap simpul daun sesuai dengan kata-kata yang ada dalam kosakata dan setiap simpul non-daun mewakili topik. Dalam DAG berubah-ubah, model LDA tidak akan memiliki hubungan antara node interior non daun tersebut, sedangkan model PAM menggunakan hubungan di antara mereka untuk memiliki koherensi lengkap yang erat di seluruh korpus (Manthiramoorthi, 2020).

BAB III

METODE PENELITIAN

3.1 Teknik Pengumpulan Data

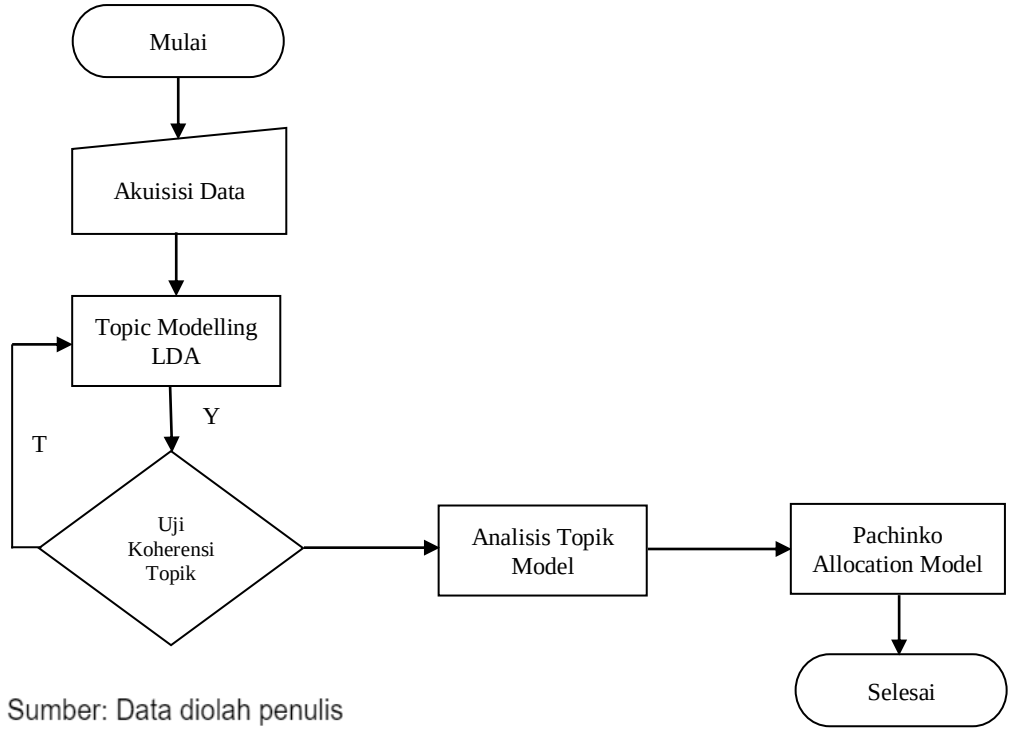
Sebagai tahap pertama, identifikasi masalah dilakukan dengan cara observasi pada media berita online tentang saham, yaitu kontan.co.id. Hasil observasi menunjukkan bahwa permasalahan terdapat pada topik yang sering berubah sewaktu-waktu dan beragam, baik dari segi topik maupun format, sehingga penggalian informasi saham melalui halaman web dirasa belum optimal. Berdasarkan kondisi tersebut, dirasa perlu untuk melakukan identifikasi topik apa yang sedang menjadi pembicaraan pada situs berita online tentang saham untuk mengetahui apa yang sedang terjadi di pasar saham Indonesia.

Tahap studi pustaka dilakukan dengan tujuan agar dapat memahami konsep, metode, dan teknologi sesuai bahasan dan permasalahan, sehingga dapat memberi solusi mengenai permasalahan yang akan digunakan dalam penyusunan penelitian ini. Tahap studi pustaka dilakukan dengan menggali informasi sesuai benang merah penelitian melalui pustaka-pustaka sebagai sumber terkait konsep-konsep atau penelitian sebelumnya yang pernah dilakukan yang terkait dengan permasalahan dalam bentuk jurnal, buku maupun referensi online. Adapun pustaka utama pada pembahasan *Topic Modelling* dan *LDA* mengacu pada penelitian David M. Blei, dkk dengan judul *Latent Dirichlet Allocation, Journal of Machine learning Research 3* (2003), dan pembahasan *Topic Coherence* mengacu pada penelitian Jonathan Chang, dkk dengan judul *Reading Tea Leaves: How Humans Interpret Topic Models* yang terdapat pada jurnal *Neural*

Information Processing Systems (2009), dan penelitian *Pachinko Allocation Model*oleh Manthiramoorthi (2020).

3.2 Alur Penelitian

Penelitian ini mempunyai bagan alir penelitian dapat dilihat pada Gambar 3.1.



Gambar 3.1 Pemodelan Topik menggunakan LDA dan PAM

3.2.1 Akuisisi Data

Tahap akuisisi data merupakan proses yang dilakukan untuk mengumpulkan data yang digunakan dalam penelitian. Data penelitian yang digunakan berupa data file berita yang diperoleh melalui proses pengambilan data secara scraping pada situs kontan.co.id. Data penelitian yang dikumpulkan sebanyak 181 berita selama bulan Februari hingga Juli 2021. Situs kontan.co.id dipilih karena link berita saham sudah jadi satu dengan ringkasan saham Google dari perusahaan yang ada di Bursa Efek Indonesia. Selain itu, situs kontan.co.id

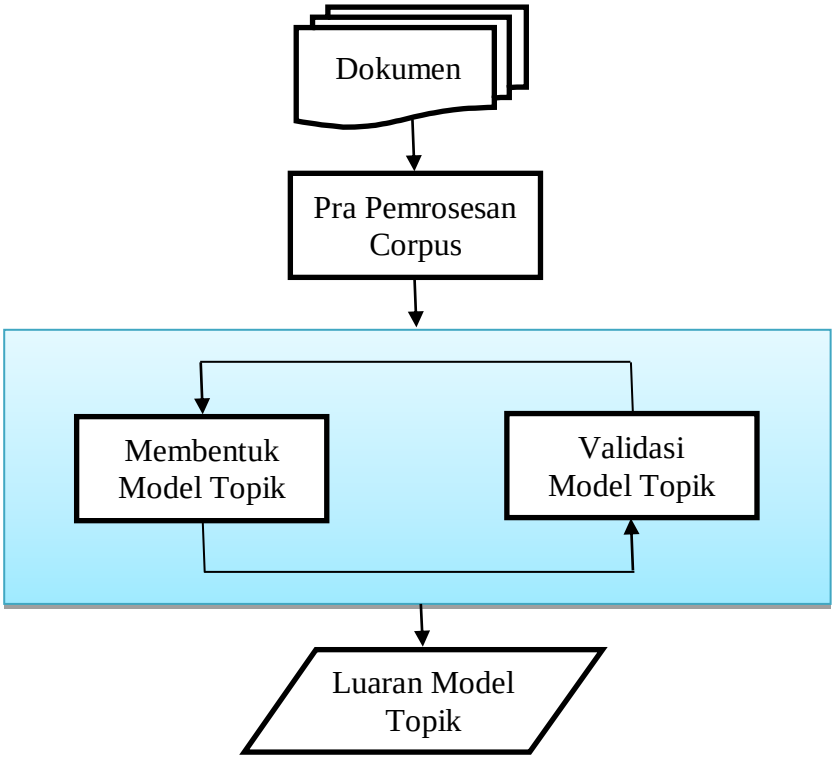
memuat lebih banyak berita tentang saham perusahaan di Indonesia dibandingkan dengan situs lain seperti bisnis.com, kompas.com, suara.com, sindonews.com, medcom.id, viva.co.id, dan m.cnnindonesia.com

Tahap menyiapkan data terdiri dari sub aktifitas pengumpulan, pemahaman, dan pemilihan data, tahap ini bertujuan untuk menyiapkan dokumen yang akan dianalisis menggunakan *Topic Modelling*. Adapun dokumen yang dianalisis adalah data artikel berita saham dari kontan.co.id yang dikumpulkan dengan cara *webscraping*.Scraping merupakanteknikmengumpulkan data padasebuah website melalui proses ektraksiinformasimenggunakanHypertext Transfer Protocol (HTTP). Dataset hasil scraping menggunakan format file *.csv

Berdasarkan data hasil *scraping*tersebut, tidak seluruh atribut data digunakan untuk *Topic Modelling*, sehingga diperlukan tahap pemilihan data. Adapun pemilihan data yang dimaksud adalah pemilihan kolom apa saja yang selanjutnya akan dianalisis. Apabila diperhatikan lebih lanjut, data yang tersimpan dalam artikel berita online perlu dilakukan pembersihan data terlebih dahulu, seperti penghapusan link,maupun data lain yang dianggap dapat mengganggu proses-proses berikutnya.

3.2.2 *Topic Modelling*dengan *Latent Dirichlet Allocation*

Tahap *Topic Modelling* dengan *Latent Dirichlet Allocation (LDA)* terhadap dokumen yang berasal dari artikel berita online kontan.co.id terdiri dari beberapa tahap, yang dapat dilihat pada Gambar 3.2.

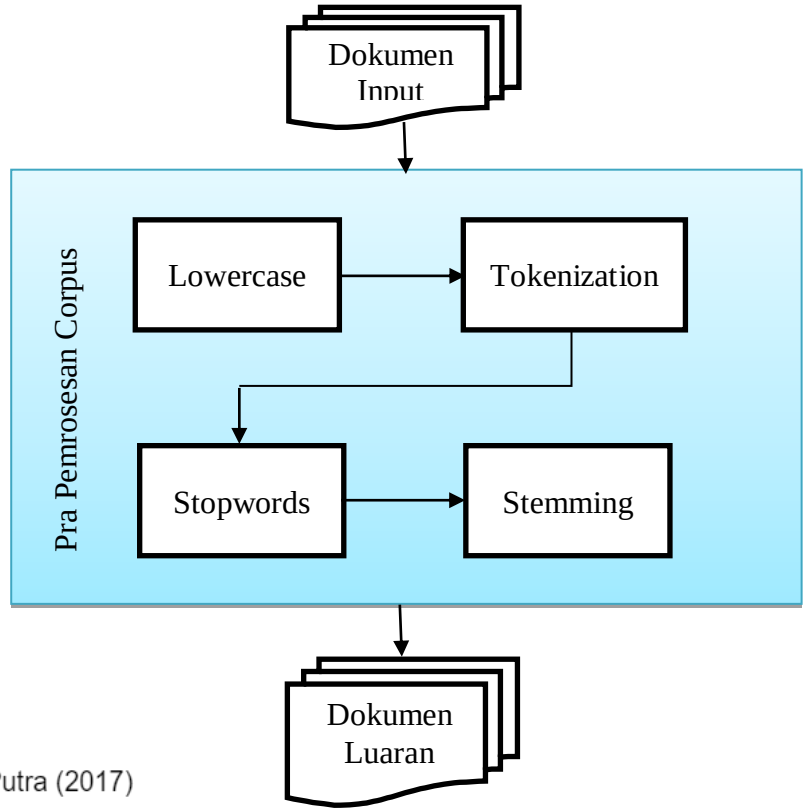


Sumber: Putra (2017)

Gambar 3.2 Tahap Topic Modelling dengan Latent Dirichlet Allocation

3.2.2.1 Pra-pemrosesan *Corpus*

Dalam melakukan *Topic Modelling*dengan *LDA*, diperlukan langkah-langkah untuk untuk mempersiapkan data sehingga dapat diolah pada tahap berikutnya, tahap ini disebut tahap pra-pemrosesan *corpus*. Adapun sub-aktivitas dari tahap pra-pemrosesan *corpus* dapat dilihatpada Gambar 3.3.



Sumber: Putra (2017)

Gambar 3.3 Tahap Pra Pemrosesan Corpus

Data teksperludibentukmenjadimenjadi*lowercase* dengantujuan agar kata yang samanamunberbedasecarapenulisanhurufkapitaldanbukan kapital, tidakdianggap kata yang berbeda.

Tokenization adalahaktivitasatau proses memisahkanderetan kata di dalamkalimatatauparagrafmenjadipotongan kata tunggalatau*termmed word*. Proses *tokenization* bertujuanuntukmempersiapkandokumenuntuk proses berikutnya, yaitu proses *stopwords*dan*Stemming* dapatdilakukan.

Stopwords merupakan kata umum (*common words*) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna, misal kata depan, kata sambung, dan kata ganti.

Daftar *stopwords* yang digunakan adalah *stopwords* Bahasa Indonesia yang disusun berdasarkan penelitian Fadillah Z Tala (2003).

Menghilangkan *stopwords* merupakan tahap yang penting, mengingat tingginya frekuensi kemunculan *stopwords* dalam dokumen, yang berujung pada tingginya probabilitas kata-kata *stopwords* dalam topik, sehingga topik tidak dapat diinterpretasi dengan baik.

Stemming digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut. *Stemming* bekerja dengan menghilangkan semua imbuhan *(affixes)*, baik yang terdiri dari awalan *(prefixes)*, sisipan *(infixes)*, akhiran *(suffixes)* dan kombinasi dari awalan dan akhiran *(confixes)* pada kata turunan. Data teks perlu dibentuk menjadi kata dasarnya dengan tujuan agar tidak terdapat kata yang sama, namun berbeda karena adanya imbuhan *(affixes)*. Proses *Stemming* dilakukan dengan menggunakan*library* Sastrawi, yaitu *library Stemmer* Bahasa Indonesia.

3.2.2.2 Membentuk Model Topik

Tahap membentuk topik model bertujuan untuk menghasilkan model topik yang paling tepat untuk dokumen. Model topik dikatakan tepat apabila mampu menghasilkan luaran yang baik pada tahap validasi model topik. Untuk menghasilkan model topik yang tepat, hal yang dilakukan adalah dengan melakukan eksperimen pada nilai *input parameter*. Adapun parameter yang

dimaksud adalah *number of topics* dan *words in topic*. Parameter *number of topics* menentukan jumlah topik dalam satu dokumen, sementara parameter *number of words in topic* menunjukkan jumlah kata penyusun topik.

3.2.2.3 Validasi Model Topik

Tahap validasi topik bertujuan untuk memastikan model topik yang dihasilkan dari hasil *Topic Modelling*yang dilakukan pada dokumen adalah benar, baik luaran berupa topik maupun kata-kata dalam topik. Dalam hal ini, tingkat kebenaran topik disesuaikan dengan dua metode, secara otomatis dengan *Perplexity* dan berdasarkan tingkat koherensi atau tingkat kemudahaannya dalam diinterpretasi manusia.

Ide dasar *Perplexity* adalah membagi secara acak dokumen menjadi data *training* dan data uji, kemudian menghitung rata-rata probabilitas *log* setiap kata dari data uji pada model yang dihasilkan dari data *training*, atau secara sederhana, *perplexity* dilakukan dengan mengambil n sampel dari N populasi data untuk diuji, apakah n sampel tersebut memiliki kesesuaian topik dengan kelompok topik dalam N populasi. Secara otomatis, penghitungan *Perplexity* sudah termasuk dalam *package gensim* untuk bahasa *python*, cara kerjanya adalah dengan menghitung rata-rata jarak geometris dari matriks data yang mewakili setiap kata menggunakan potongan dokumen evaluasi *corpus*. Apabila nilai *Perplexity* dirasa sudah mencapai titik optimal, maka akan dilakukan validasi model topik berdasarkan tingkat koherensinya sementara apabila model topik belum mencapai nilai *perplexity* yang optimal, maka akan kembali ke tahap membentuk topik model untuk kembali dilakukan eksperimen pada *input parameter*.

3.2.3 Uji Koherensi Topik

Untuk validasi topik berdasarkan tingkat koherensi atau tingkat kemudahannya dalam diinterpretasi manusia, validasi topik dilakukan dengan metode *Topic Intrusion task* dan validasi kata-kata dalam topik dilakukan dengan *Word Intrusion Task*. *Word Intrusion task* dilakukan dengan menyisipkan sebuah kata dengan probabilitas rendah dalam kata-kata hasil proses *Topic Modelling*pada suatu topik tertentu, kemudian meminta beberapa responden untuk menebak manakah kata yang bukan hasil proses *Topic Modelling*.

Intrusion task dilakukandenganmenyajikancuplikandokumen yang disertai beberapa pilihan topik kepada beberapa responden, salah satu dari beberapa topik tersebut merupakan topik yang memiliki probabilitas rendah dalam dokumen.

Mengingat tahap validasi topik ini bertujuan untuk memastikan luaran dari hasil *Topic Modelling*yang dilakukan pada dokumen memiliki koherensi yang baik, maka apabila topik belum memiliki tingkat koherensi yang optimal, maka akan kembali ke tahap membentuk topik model untuk kembali dilakukan eksperimen pada *input parameter*.

3.2.4 Menganalisis Topik Model

Tahap menganalisis topik model bertujuan untuk mendapatkan kesimpulan terkait hasil uji koherensi topik secara lebih spesifik. Menganalisis topik model dilakukan dengan metode statistika sehingga didapatkan kesimpulan terkait hasil perbandingan antara word intrusion task dengan topic intrusion task dan perbandingan hasil pengaruh perlakuan stem.

3.2.5 Pachinko Allocation Model (PAM)

Setelah luaran topik LDA diperoleh, langkah berikutnya mengerjakan model alokasi pachinko (Pachinko Allocation Model) untuk menunjukkan hubungan yang koheren antara topik yang dihasilkan.

Penelitian ini menggunakan kerangka PAM empat tingkat. Kerangka PAM empat tingkat dijelaskandalam Li dan McCallum (2006). Ada satu node di bagian atas Directed Acyclic Graph (DAG) yang mendefinisikan distribusinode di tingkat kedua, yang disebut sebagai super topik. Setiap node di tingkat kedua mendefinisikan distribusi ke semua node di tingkat ketiga, atau sub-topik.Setiap sub-topik dipetakan ke satu distribusi di ataskosa kata. Oleh karena itu, hanya sub-topik sebenarnyamenghasilkan kata-kata. Super-topik mewakili kelompoktopik yang sering muncul.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

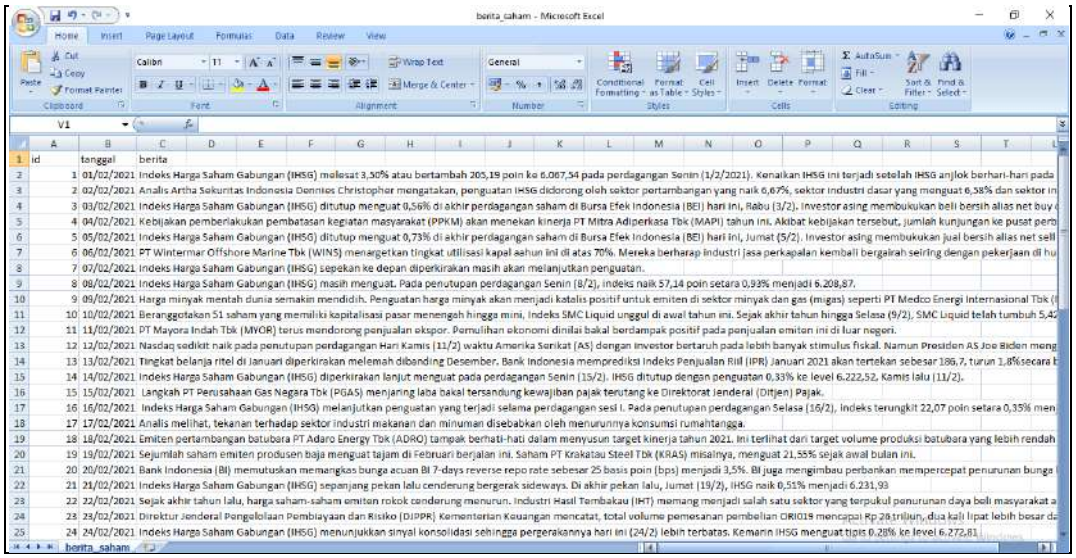
4.1.1 Membersihkan Data

Pada tahap membersihkan data, *library* yang digunakan adalah *library* "re" dan *library* "preprocessor". Tahap membersihkan data diawali dengan membuat *list* sebagai wadah untuk menyimpan data yang sudah dibersihkan dengan nama 'list_data_bersih'. Kemudian untuk melakukan pembersihan data, dilakukan secara bertahap dengan rincian sebagai berikut: dataBersih1 membersihkan *url*, *mention*, *reserved words* dan number, dataBersih2 membersihkan karakter *non-alphanumeric*, dataBersih3 membersihkan token sepanjang 1 digit. Setelah data dibersihkan, data disimpan dalam *list* 'berita_saham'.

Data penelitian yang digunakan berupa data file berita yang diperoleh melalui proses pengambilan data secara scraping pada situs kontan.co.id. Data penelitian yang dikumpulkan sebanyak 181 berita selama bulan Februari hingga Juli 2021. Situs kontan.co.id dipilih karena link berita saham sudah jadi satu dengan ringkasan saham Google dari perusahaan yang ada di Bursa Efek Indonesia.

a. Berita saham bulan Februari 2021

Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan Februari 2021 sebanyak 28 baris sebagai berikut.

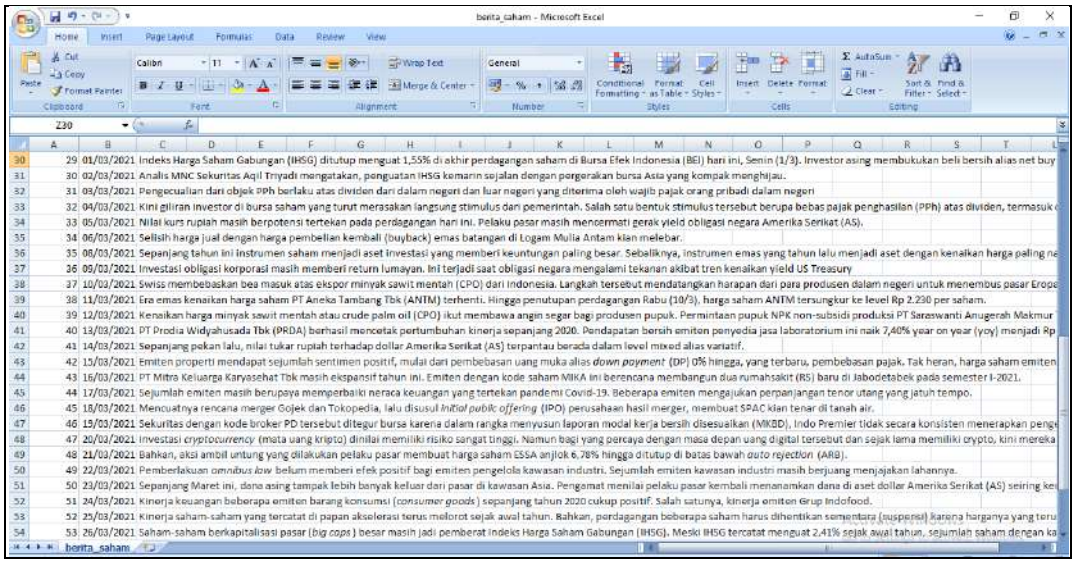


id	tanggal	berita
1	01/02/2021	Indeks Harga Saham Gabungan (IHSG) melesat 3,50% atau bertambah 203,19 poin ke 6.067,54 pada perdagangan Senin (1/2/2021). Kenaikan IHSG ini terjadi setelah IHSG anjlok berhari-hari pada
2	02/02/2021	Analis Artha Sekuritas Indonesia Dennies Christopher mengatakan, penguatan IHSG didorong oleh sektor pertambangan yang naik 0,67%, sektor industri dasar yang menguat 6,38% dan sektor in
3	03/02/2021	Indeks Harga Saham Gabungan (IHSG) ditutup menguat 0,56% di akhir perdagangan saham di Bursa Efek Indonesia (BEI) hari ini, Rabu (3/2). Investor asing membukukan beli bersih alias net buy
4	04/02/2021	Kebijakan pemberlakukan pembatasan kegiatan masyarakat (PPKM) akan menekan kinerja PT Mitra Adiperkasa Tbk (MAPI) tahun ini. Akibat kebijakan tersebut, jumlah kunjungan ke pusat pero
5	05/02/2021	Indeks Harga Saham Gabungan (IHSG) ditutup menguat 0,73% di akhir perdagangan saham di Bursa Efek Indonesia (BEI) hari ini, Jumat (5/2). Investor asing membukukan jual bersih alias net sell
6	06/02/2021	PT Wintermar Offshore Marine Tbk (WINS) menargetkan tingkat utilisasi kapal aahun ini di atas 70%. Mereka berharap industri jasa perkapalan kembali bergairah seiring dengan pekerjaan di hu
7	07/02/2021	Indeks Harga Saham Gabungan (IHSG) sepekan ke depan diperkirakan masih akan melanjutkan penguatan.
8	08/02/2021	Indeks Harga Saham Gabungan (IHSG) masih menguat. Pada penutupan perdagangan Senin (8/2), indeks naik 57,14 poin setara 0,93% menjadi 6.208,87.
9	09/02/2021	Harga minyak mentah dunia semakin mendidih. Penguatan harga minyak akan menjadi katalis positif untuk emiten di sektor minyak dan gas (migas) seperti PT Medco Energi Internasional Tbk (I
10	10/02/2021	Beranggotakan 51 saham yang memiliki kapitalisasi pasar menengah hingga mini, Indeks SMC Liquid unggul di awal tahun ini. Sejak akhir tahun hingga Selasa (9/2), SMC Liquid telah tumbuh 5,42
11	11/02/2021	PT Mayora Indah Tbk (MYOR) terus mendorong penjualan ekspor. Pemulihan ekonomi dinilai bakal berdampak positif pada penjualan emiten ini di luar negeri.
12	12/02/2021	Nasdaq sedikit naik pada penutupan perdagangan Hari Kamis (11/2) waktu Amerika Serikat (AS) dengan investor bertaruh pada lebih banyak stimulus fiskal. Namun Presiden AS Joe Biden meng
13	13/02/2021	Tingkat belanja ritel di Januari diperkirakan melemah dibanding Desember. Bank Indonesia memprediksi Indeks Penjualan Riil (IPR) Januari 2021 akan tertekan sebesar 186,7, turun 1,8%secara t
14	14/02/2021	Indeks Harga Saham Gabungan (IHSG) diperkirakan lanjut menguat pada perdagangan Senin (15/2). IHSG ditutup dengan penguatan 0,33% ke level 6.222,52, Kamis lalu (11/2).
15	15/02/2021	Langkah PT Perusahaan Gas Negara Tbk (PGAS) menjaring laba bakal tersandung kewajiban pajak terutang ke Direktorat Jenderal (Ditjen) Pajak.
16	16/02/2021	Indeks Harga Saham Gabungan (IHSG) melanjutkan penguatan yang terjadi selama perdagangan sesi I. Pada penutupan perdagangan Selasa (16/2), Indeks terungkit 22,07 poin setara 0,35% men
17	17/02/2021	Analis melihat, tekanan terhadap sektor industri makanan dan minuman disebabkan oleh menurunnya konsumsi rumahtangga.
18	18/02/2021	Emiten pertambangan batubara PT Adaro Energy Tbk (ADRO) tampak berhati-hati dalam menyusun target kinerja tahun 2021. Ini terlihat dari target volume produksi batubara yang lebih rendah
19	19/02/2021	Sejumlah saham emiten produsen baja menguat tajam di Februari berjalan ini. Saham PT Krakatau Steel Tbk (KRAS) misalnya, menguat 21,55% sejak awal bulan ini.
20	20/02/2021	Bank Indonesia (BI) memutuskan memangkas bunga acuan BI 7-days reverse repo rate sebesar 25 basis poin (bps) menjadi 3,5%. BI juga mengimbau perbankan mempercepat penurunan bunga
21	21/02/2021	Indeks Harga Saham Gabungan (IHSG) sepanjang pekan lalu cenderung bergerak sideways. Di akhir pekan lalu, Jumat (19/2), IHSG naik 0,51% menjadi 6.231,93
22	22/02/2021	Sejak akhir tahun lalu, harga saham-saham emiten rokok cenderung menurun. Industri Hasil Tembakau (IHT) memang menjadi salah satu sektor yang terpukul penurunan daya beli masyarakat a
23	23/02/2021	Direktur Jenderal Pengelolaan Pembiayaan dan Risiko (DJPPR) Kementerian Keuangan mencatat, total volume pemesanan pembelian DR019 mencapai Rp 28 triliun, dua kali lipat lebih besar da
24	24/02/2021	Indeks Harga Saham Gabungan (IHSG) menunjukkan sinyal konsolidasi sehingga pergerakannya hari ini (24/2) lebih terbatas. Kemarin IHSG menguat tipis 0,28% ke level 6.272,81

Gambar 4.1 Berita saham dari kontan.co.id bulan Februari 2021

b. Berita saham bulan Maret 2021

Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan Maret 2021 sebanyak 30 baris sebagai berikut.

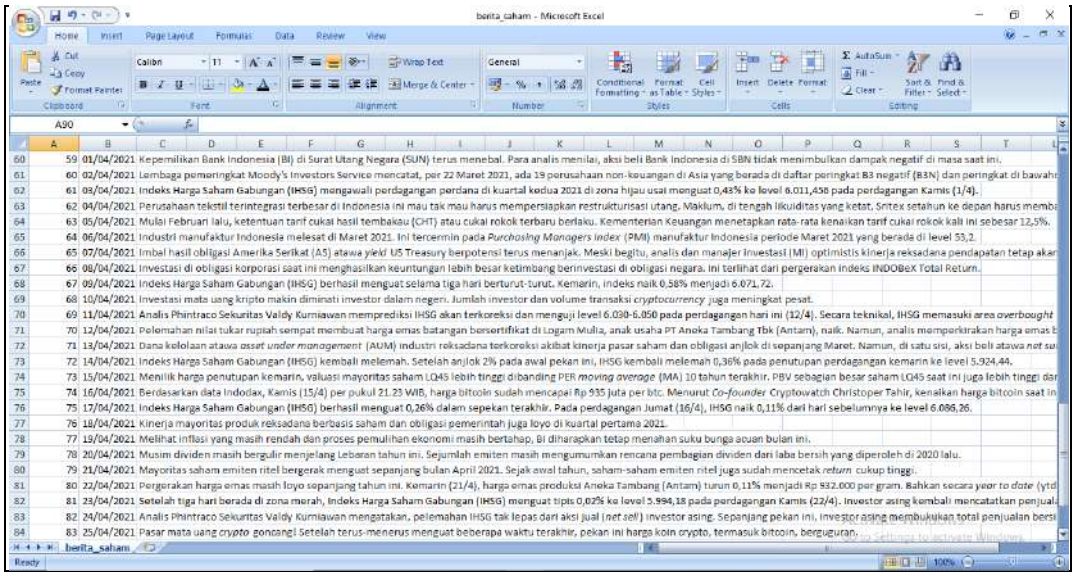


id	tanggal	berita
29	01/03/2021	Indeks Harga Saham Gabungan (IHSG) ditutup menguat 1,55% di akhir perdagangan saham di Bursa Efek Indonesia (BEI) hari ini, Senin (1/3). Investor asing membukukan beli bersih alias net buy
30	02/03/2021	Analis MNC Sekuritas Aqil Triyadi mengatakan, penguatan IHSG kemarin sejalan dengan pergerakan bursa Asia yang kompak menghijau.
31	03/03/2021	Pengecualian dari objek PPh berlaku atas dividen dari dalam negeri dan luar negeri yang diterima oleh wajib pajak orang pribadi dalam negeri
32	04/03/2021	Kini giliran investor di bursa saham yang turut merasakan langsung stimulus dari pemerintah. Salah satu bentuk stimulus tersebut berupa bebas pajak penghasilan (PPh) atas dividen, termasuk
33	05/03/2021	Nilai kurs rupiah masih berpotensi tertekan pada perdagangan hari ini. Pelaku pasar masih mencermati gerak yield obligasi negara Amerika Serikat (AS).
34	06/03/2021	Selisih harga jual dengan harga pembelian kembali (buyback) emas batangan di Logam Mulia Antam kian melebar.
35	08/03/2021	Sepanjang tahun ini instrumen saham menjadi aset investasi yang memberi keuntungan paling besar. Sebaliknya, instrumen emas yang tahun lalu menjadi aset dengan kenaikan harga paling ne
36	09/03/2021	Investasi obligasi korporasi masih memberi return lumayan. Ini terjadi saat obligasi negara mengalami tekanan akibat tren kenaikan yield US Treasury
37	10/03/2021	Swiss membebaskan bea masuk atas ekspor minyak sawit mentah (CPO) dari Indonesia. Langkah tersebut mendatangkan harapan dari para produsen dalam negeri untuk menembus pasar Eropa
38	11/03/2021	Era emas kenaikan harga saham PT Aneka Tambang Tbk (ANTM) terhenti. Hingga penutupan perdagangan Rabu (10/3), harga saham ANTM tersungkur ke level Rp 2.230 per saham.
39	12/03/2021	Kenaikan harga minyak sawit mentah atau crude palm oil (CPO) ikut membawa angin segar bagi produsen pupuk. Permintaan pupuk NPK non-subsidi produksi PT Saraswanti Anugerah Makmur
40	13/03/2021	PT Prodia Widyahusada Tbk (PRDA) berhasil mencetak pertumbuhan kinerja sepanjang 2020. Pendapatan bersih emiten penyedia jasa laboratorium ini naik 7,40% year on year (yoy) menjadi Rp
41	14/03/2021	Sepanjang pekan lalu, nilai tukar rupiah terhadap dollar Amerika Serikat (AS) terpantau berada dalam level mixed alias variatif.
42	15/03/2021	Emiten properti mendapat sejumlah sentimen positif, mulai dari pembebasan uang muka alias *down payment* (DP) 0% hingga, yang terbaru, pembebasan pajak. Tak heran, harga saham emiten
43	16/03/2021	PT Mitra Keluarga Karyasehat Tbk masih ekspansif tahun ini. Emiten dengan kode saham MIKA ini berencana membangun dua rumahsakit (RS) baru di Jabodetabek pada semester I-2021.
44	17/03/2021	Sejumlah emiten masih berupaya memperbaiki neraca keuangan yang tertekan pandemi Covid-19. Beberapa emiten mengajukan perpanjangan tenor utang yang jatuh tempo.
45	18/03/2021	Mencuatnya rencana merger Gojek dan Tokopedia, lalu disusul *Initial public offering* (IPO) perusahaan hasil merger, membuat SPAC kian tenar di tanah air.
46	19/03/2021	Sekuritas dengan kode broker PD tersebut ditegur bursa karena dalam rangka menyusun laporan modal kerja bersih disesuaikan (MKBD), Indo Premier tidak secara konsisten menerapkan penge
47	20/03/2021	Investasi *cryptocurrency* (mata uang kripto) dinilai memiliki risiko sangat tinggi. Namun bagi yang percaya dengan masa depan uang digital tersebut dan sejak lama memiliki crypto, kini mereka
48	21/03/2021	Bahkan, aksi ambil untung yang dilakukan pelaku pasar membuat harga saham ESSA anjlok 6,78% hingga ditutup di batas bawah *auto rejection* (ARB).
49	22/03/2021	Pemberlakuan *omnibus law* belum memberi efek positif bagi emiten pengelola kawasan industri. Sejumlah emiten kawasan industri masih berjuang menjajakan lahannya.
50	23/03/2021	Sepanjang Maret ini, dana asing tampak lebih banyak keluar dari pasar di kawasan Asia. Pengamat menilai pelaku pasar kembali menanamkan dana di aset dollar Amerika Serikat (AS) seiring ke
51	24/03/2021	Kinerja keuangan beberapa emiten barang konsumsi (*consumer goods*) sepanjang tahun 2020 cukup positif. Salah satunya, kinerja emiten Grup Indofood.
52	25/03/2021	Kinerja saham-saham yang tercatat di papan akselerasi terus melorot sejak awal tahun. Bahkan, perdagangan beberapa saham harus dihentikan sementara (suspensi) karena harganya yang teru
53	26/03/2021	Saham-saham berkapitalisasi pasar (*big caps*) besar masih jadi pemberat Indeks Harga Saham Gabungan (IHSG). Meski IHSG tercatat menguat 2,41% sejak awal tahun, sejumlah saham dengan ka

Gambar 4.2 Berita saham dari kontan.co.id bulan Maret 2021

c. Berita saham bulan April 2021

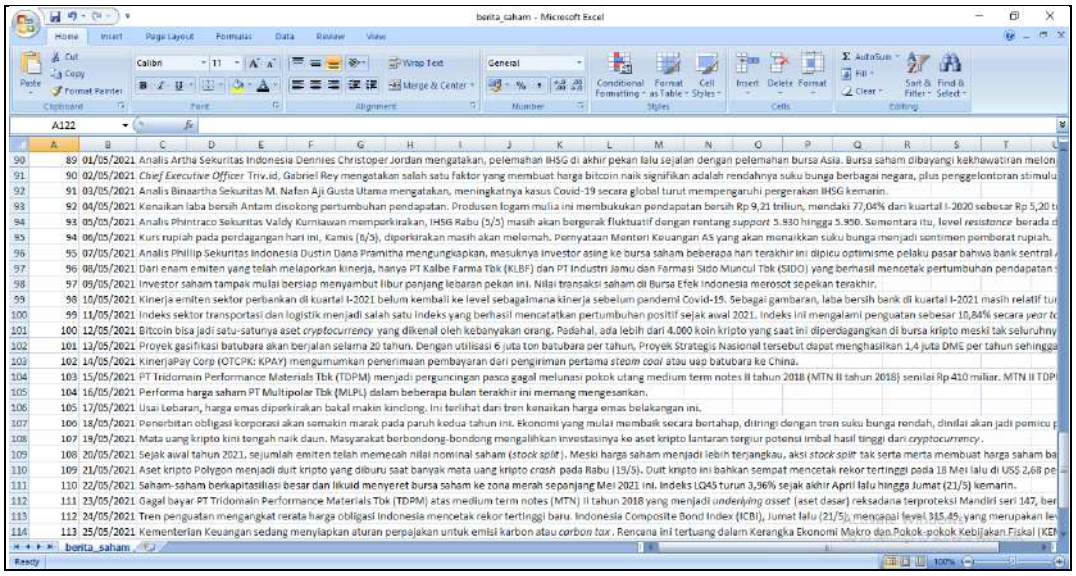
Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan April 2021 sebanyak 30 baris sebagai berikut.



Gambar 4.3 Berita saham dari kontan.co.id bulan April 2021

d. Berita saham bulan Mei 2021

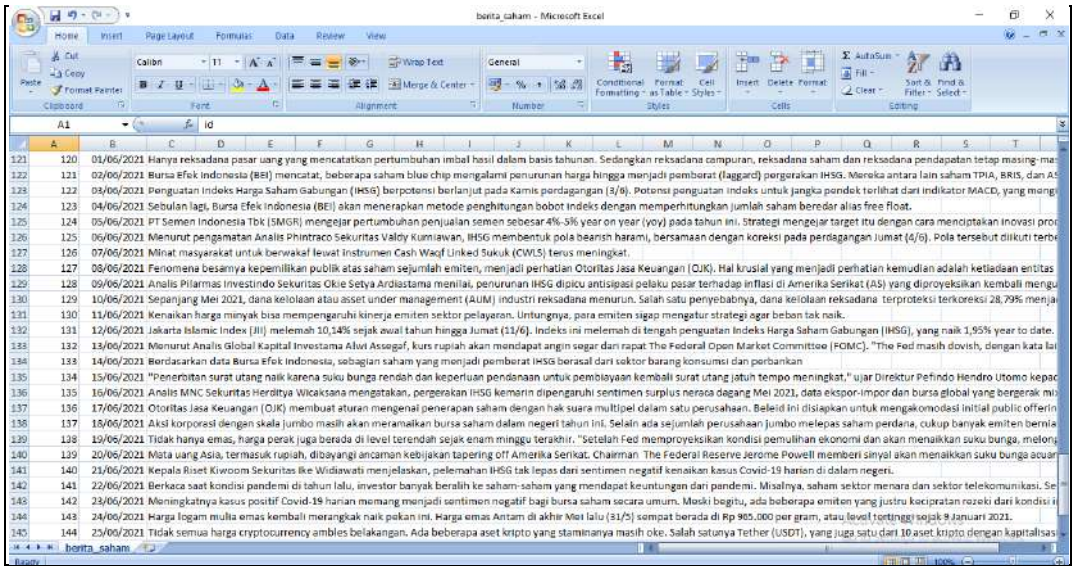
Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan Mei 2021 sebanyak 31 baris sebagai berikut.



Gambar 4.4 Berita saham dari kontan.co.id bulan Mei 2021

e. Berita saham bulan Juni 2021

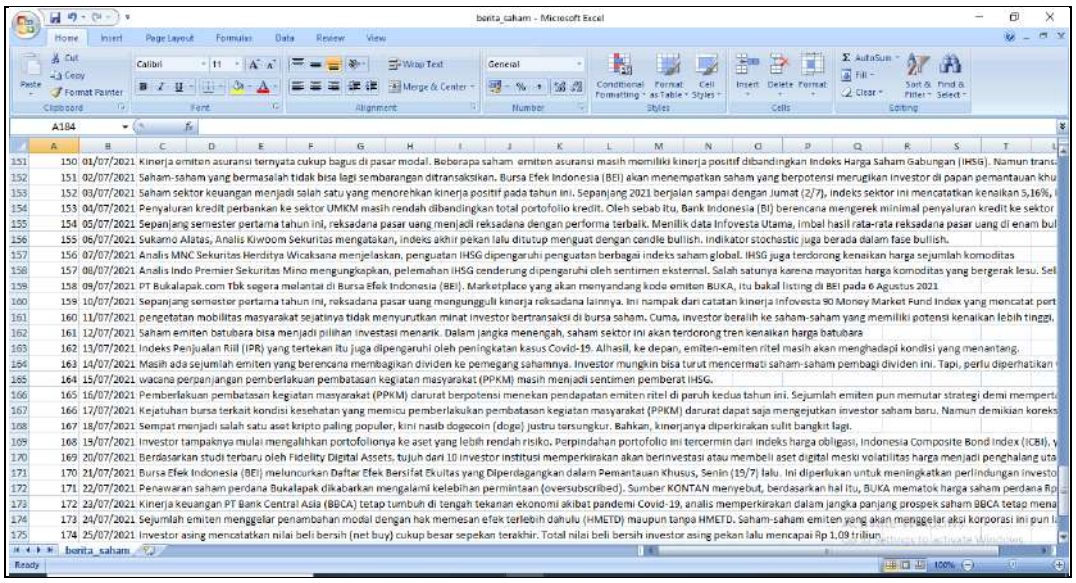
Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan Juni 2021 sebanyak 30 baris sebagai berikut.



Gambar 4.5 Berita saham dari kontan.co.id bulan Juni 2021

e. Berita saham bulan Juli 2021

Penelitian ini mengumpulkan berita saham kontan.co.id pada bulan Juli 2021 sebanyak 32 baris sebagai berikut.



Gambar 4.6 Berita saham dari kontan.co.id bulan Juli 2021

4.1.2Pra-Proses Data

Tahap pra-proses data merupakan tahapan yang dilakukan sebelum pemrosesan data agar tahap pemrosesan data dapat berjalan dengan optimal. Ada 4 tahapan utama

dalam melakukan implementasi penelitian, yaitu pendefinisian *stopwords*, menghapus *stopwords* dari dokumen, mengkonversidokumen ke dalam bentuk *dictionary* dan mengkonversi *dictionary* ke dalam bentuk matriks dokumen atau *corpus*.

4.1.2.1 Pendefinisian *Stopwords*

Dalam mendefinisikan *stopwords*, *stopwords* disimpan dalam *file* berformat *.csv* yang terdiri dari satu kata per baris, untuk kemudian dimuat dalam *list* dengan nama 'list_*stopword*' yang mana *list* tersebut telah dibuat sebelumnya.

4.1.2.2 Menghapus *Stopwords* dari Dokumen

Setelah *stopwords* didefinisikan dan disimpan dalam *list* dan dokumen telah tersimpan dalam *list*, langkah selanjutnya adalah dengan menghapus *stopwords* yang terdapat dalam dokumen. Dalam menghapus *stopwords* dari dokumen, perlu dilakukan tokenisasi. Adapun *library* yang digunakan adalah 'nltk.tokenize' dengan modul 'RegexpTokenizer'. Adapun tahap-tahap yang dilakukan adalah untuk setiap dokumen dalam *corpus* yang tersimpan dalam 'berita_saham' diubah menjadi *lowercase* dengan tipe data *string*, kemudian dokumen dilakukan tokenisasi, sehingga dokumen dipecah per-kata, kata-kata yang hanya terdiri dari angka dihapus, kemudian untuk setiap token yang tercatat sebagai *stopword* dalam 'list_*stopword*' dihapus dari dokumen, pada akhirnya data yang telah bersih dari *stopword* disimpan dalam *list* 'tokenlist_dataAgregat' yang telah dibuat sebelumnya.

4.1.2.3 Mengkonversi Dokumen ke Dalam bentuk *Dictionary*

Tahap berikutnya adalah tahap mengkonversi dokumen ke dalam bentuk *dictionary*, tujuan dari tahap ini adalah menyimpan data token ke dalam bentuk *dictionary*, kemudian memberikan indeks yang berfungsi untuk mengidentifikasi kata tersebut. Dalam melakukan konversi dokumen ke dalam bentuk *dictionary* digunakan *library gensim* dengan modul *corpora*. *File dictionary*kemudian disimpan dengan nama

‘dictionary_data_agregat.dict’ agar lebih mudah untuk digunakan dalam melakukan eksperimen pada tahap-tahap berikutnya.

4.1.2.4 Mengkonversi *Dictionary* dan *List* ke dalam Matriks Dokumen

Setelah terdapat *file dictionary* hasil konversi dokumen, tahap selanjutnya adalah mengkonversi *dictionary* dalam matriks dokumen yang selanjutnya disebut *corpus*. Adapun *library* yang digunakan adalah *gensim* dengan modul *corpora*. *Corpus* tersebut kemudian disimpan dalam *file* dengan nama ‘corpora.mm’ agar lebih mudah untuk digunakan dalam melakukan eksperimen pada tahap-tahap berikutnya.

4.1.3 Pemodelan Topik dengan *Latent Dirichlet Allocation*

Tahapan pemodelan topik dengan *Latent Dirichlet Allocation* merupakan tahapan yang dilakukan untuk membentuk model topik. Terdapat dua hal penting yang menjadi perhatian dalam tahap ini, yaitu alur pemodelan topik dengan *Latent Dirichlet Allocation* dan eksperimen pemodelan topik.

4.1.3.1 Alur Pemodelan Topik dengan *Latent Dirichlet Allocation*

a. *Loading Dictionary* dan *Corpus*

Dalam konteks ini, *loading* data yang dimaksud adalah memuat data *dictionary* dan *corpus* yang pada tahap sebelumnya telah disimpan *file*. Agar lebih mudah dalam pengolahan, data disimpan dalam variabel, variabel ‘dictionary_dataAgregat’ untuk *file* ‘dictionary_dataAgregat.dict’ dan variabel ‘*corpus*’ untuk *file* ‘file/corpora.mm’.

b. Pembentukan Model Topik

Pada tahap pembentukan model topik *library* yang digunakan adalah *gensim* dengan modul *MmCorpus* dan *Dictionary*. Dalam pembentukan model topik, diperlukan *input parameter*, yaitu jumlah topik, jumlah kata dalam topik, dan *passes*. Adapun yang dimaksud dengan *passes* adalah jumlah iterasi dalam pembentukan model topik. Ketiga *input parameter* ini nantinya akan dilakukan eksperimen untuk mencari nilai *perplexity*

yang optimal. Nilai *perplexity* yang semakin kecil menunjukkan model yang dibentuk semakin baik.

c. Pendokumentasian *Logging*

Dalam melakukan eksperimen, diperlukan *logging* untuk mengetahui catatan terkait kejadian yang terjadi dalam proses pembentukan model topik. Catatan penting yang dibutuhkan adalah nilai *perplexity* yang mana nilai *perplexity* sudah terkalkulasi secara otomatis sebagai fitur modul gensim. Untuk melakukan *logging*, diperlukan *library logging*, kemudian *file logging* disimpan dalam *file* berformat *.csv* dengan nama sesuai kebutuhan. Dalam contoh kasus di bawah, nama *file* disesuaikan dengan jumlah topik dan jumlah kata.

4.1.3.2 Eksperimen Pemodelan Topik dengan *LDA*

Tahapan eksperimen pemodelan topik dengan *Latent Dirichlet Allocation* merupakan tahapan yang dilakukan untuk membentuk model topik terbaik dengan melakukan eksperimen pada *input parameter*, yaitu *passes* atau jumlah iterasi dan jumlah topik. Sehingga eksperimen akan dibedakan kedalam dua tahap, yaitu penentuan jumlah iterasi dan penentuan jumlah topik.

a. Penentuan Jumlah Iterasi

Dalam *LDA*, istilah iterasi dikenal dengan *passes*. Penentuan iterasi merupakan tahap yang penting dalam menentukan model, hal ini untuk menghasilkan model yang terbaik, dalam konteks ini apabila jumlah iterasi terlalu sedikit, akan menghasilkan model yang belum stabil dan *under fitting*, sementara iterasi yang terlalu banyak akan menghasilkan model yang *overfitting*. Penentuan jumlah *passes* dilakukan dengan melakukan eksperimen pada jumlah topik. Penentuan jumlah *passes* diawali dengan memberikan nilai mula-mula sebesar 50, kemudian jumlah topik ditentukan yaitu sebanyak 5 kali, yaitu 10, 20, 30, 40, 50 topik. Berdasarkan eksperimen jumlah topik,

nilai *perplexity* yang muncul akan dicatat untuk dianalisis tren nilainya secara visual dan dilakukan penghitungan standar deviasinya. Sehingga pada akhirnya nilai *passes* yang akan digunakan adalah nilai *passes* yang paling awal setelah menunjukkan tren yang stabil.

b. Penentuan Jumlah Topik

Setelah menentukan jumlah iterasi, eksperimen dilakukan pada jumlah topik. Eksperimen jumlah topik merupakan tahap yang penting dalam menentukan model, hal ini untuk menghasilkan model yang terbaik, dalam konteks ini model yang baik adalah model dengan nilai *perplexity* yang rendah, semakin rendah nilai *perplexity*, menunjukkan akurasi model lebih baik. Penentuan jumlah topik dilakukan dengan melakukan eksperimen pada nilai jumlah topik. Penentuan jumlah topik diawali dengan memberikan nilai mula-mula yaitu sebanyak 5 kali, yaitu 10, 20, 30, 40, 50 topik. Berdasarkan eksperimen jumlah topik, nilai *perplexity* yang muncul akan dicatat untuk dianalisis tren nilainya secara visual dan dilakukan penghitungan standar deviasinya. Sehingga pada akhirnya jumlah topik yang dipilih adalah jumlah topik yang memiliki nilai rata-rata paling rendah dengan standar deviasi minimum.

c. Menyimpan Model

Model dengan jumlah iterasi dan jumlah topik yang telah ditentukan perlu disimpan untuk dapat digunakan kembali dengan cepat. Model disimpan dalam format *.model*. ‘namaFileModel’ yang digunakan adalah jumlah topik dan jumlah kata yang digunakan untuk eksperimen pembentukan model.

4.1.4Validasi Model Topik

4.1.4.1 Klasifikasi Dokumen ke dalam Topik dengan Metode *LDA*

Tahap klasifikasi dokumen ke dalam topik dengan metode *LDA* bertujuan untuk mengklasifikasikan setiap dokumen ke dalam topik yang telah dibentuk sebelumnya.

Secara umum, tahapan dalam melakukan klasifikasi adalah melakukan import *library*, *loading input*, operasi data, dan menampilkan distribusi dokumen dalam topik.

a. Import *Library*

Untuk melakukan klasifikasi data, diperlukan modul *MmCorpus* dan *Dictionary* yang terdapat pada *library gensim*, maka dari itu perlu dilakukan import *library*.

b. *Loading Input*

Terdapat empat *input* yang digunakan untuk melakukan klasifikasi data, yaitu model, *corpus*, *dictionary* dan *input* dokumen. *Input* model yang dipilih merupakan model (dengan format *.model*) yang sudah dibentuk sebelumnya dan terpilih sebagai model untuk melakukan klasifikasi. *Corpus* dan *dictionary* dijadikan *input* merupakan *corpus* dan *dictionary* yang telah dibentuk pada tahap prapemrosesan data. *Input* dokumen merupakan *file* yang berisi kumpulan dokumen yang akan diklasifikasi.

c. Operasi Data

Tahap operasi data dibagi menjadi tiga tahap, mengubah kumpulan dokumen ke dalam *list* dokumen, pembentukan *list* data hasil klasifikasi, dan klasifikasi data. Tahap mengubah kumpulan dokumen ke dalam *list* dokumen bertujuan untuk menyatukan berbagai sumber dokumen yang nantinya akan diklasifikasikan dalam bentuk *list* agar mudah dalam melakukan klasifikasi. Tahap pembentukan *list* data hasil klasifikasi bertujuan untuk membentuk *list* sebagai wadah yang nantinya akan diisi oleh dokumen-dokumen yang sudah diklasifikasi, *list* data dibuat sejumlah topik yang digunakan untuk mengklasifikasi. Tahap klasifikasi data bertujuan untuk melakukan klasifikasi data ke dalam masing-masing topik dengan mengabaikan dokumen yang tidak condong ke topik manapun, sehingga untuk setiap *list* akan tersimpan dokumen dari masing-masing dokumen yang terklasifikasi dalam topik tersebut.

d. Menampilkan Distribusi Dokumen dalam Topik

Setelah dokumen diklasifikasikan, untuk mempermudah dalam menganalisis distribusi dokumen dalam topik.

4.1.4.2 Analisis Distribusi Probabilitas Dokumen Per Topik dengan Metode *LDA*

a. Import *Library*

Untuk melakukan klasifikasi data, diperlukan modul *MmCorpus* dan *Dictionary* yang terdapat pada *library gensim*.

b. *Loading Input*

Terdapat empat *input* yang digunakan untuk melakukan klasifikasi data, yaitu model, *corpus*, *dictionary* dan *input* dokumen. *Input* model yang dipilih merupakan model (dengan format *.model*) yang sudah dibentuk sebelumnya dan terpilih sebagai model untuk melakukan klasifikasi. *Corpus* dan *dictionary* dijadikan *input* merupakan *corpus* dan *dictionary* yang telah dibentuk pada tahap prapemrosesan data. *Input* dokumen merupakan *file* yang berisi kumpulan dokumen yang akan diklasifikasi.

c. Operasi Data

Tahap operasi data dibagi menjadi dua tahap, inisiasi *list* data, dan klasifikasi data. Tahap inisiasi *list* data bertujuan untuk membentuk *list* sebagai wadah yang nantinya akan diisi oleh dokumen-dokumen yang sudah diklasifikasi, *list* data dibuat sejumlah topik yang digunakan untuk mengklasifikasi. Tahap klasifikasi data bertujuan untuk melakukan klasifikasi data ke dalam masing-masing topik dengan mengabaikan dokumen yang tidak condong ke topik manapun, sehingga untuk setiap *list* akan tersimpan probabilitas dari masing-masing dokumen yang terklasifikasi dalam topik tersebut.

d. Menampilkan Jumlah Distribusi Dokumen Per Topik

Setelah dokumen diklasifikasikan, untuk mempermudah dalam menganalisis

distribusi probabilitas dari dokumen dalam topik.

d.1. Visualisasi Distribusi Probabilitas Dominan dari Dokumen per Topik

Sebagai bagian dari tahap validasi topik model, distribusi probabilitas dominan dari dokumen per topik perlu divisualisasikan dengan tujuan untuk memudahkan dalam menyimpulkan apakah distribusi probabilitas cukup memberikan hasil yang meyakinkan melalui kecenderungan distribusinya.

d.1.a). Distribusi Probabilitas Dominan dari Dokumen per Topik dengan Metode *LDA*

d.1.a).1).Distribusi Probabilitas Dominan

Untuk menampilkan distribusi probabilitas dominan dari dokumen per topik dalam bentuk *list*.

d.1.a).2). Histogram

Distribusi probabilitas dominan dari dokumen per topik dalam bentuk *list*, sehingga masih sulit untuk dilakukan analisis, sehingga akan lebih mudah untuk melakukan analisis apabila divisualisassikan dalam bentuk histogram. Untuk memvisualisasikan ke dalam histogram, dibutuhkan *library* ‘matplotlib.pyplot’.

4.1.5Uji Koherensi Topik

Analisis topik merupakan tahap untuk menganalisis topik yang menjadi luaran dari tahap pembentukan model topik. Analisis topik dilakukan dengan melakukan pengamatan terhadap seluruh distribusi topik termasuk distribusi kata dalam topik.

4.1.5.1. Menampilkan Topik

Sebagai langkah awal untuk menganalisis topik, daftar topik perlu ditampilkan.Fungsi print_topics(-1) merupakan sebuah fungsi yang digunakan untuk mengurutkan probabilitas distribusi kata yang terdapat dalam topik sehingga lebih mudah untuk diamati.

4.1.5.2. Penyusunan Materi Kuesioner

Penyusunan materi kuesioner dibedakan menjadi empat, yaitu *Word Intrusion task* dengan *Stem*, *Word Intrusion task* tanpa *Stem*, *Topic Intrusion task* dengan *Stem* dan *TopicIntrusion task* tanpa *Stem*.

a. *Word Intrusion task*

Sesuai dengan langkah-langkah penyusunan materi kuesioner *Word Intrusion Task*, implementasi materi kuesioner dilakukan dengan menyusun tabel deret akumulatif untuk masing-masing topik, kemudian dilanjutkan dengan melakukan pemilihan kata dalam topik secara acak dengan metode *roulette wheel*,sehingga menghasilkan kelompok-kelompok kecil yang nantinya akan diuji keterkaitan setiap katanya, pada akhirnya, untuk setiap kelompok kecil, disisipkan satu kata lain yang memiliki probabilitas kecil terhadap topik tersebut. Berikut merupakan deret akumulatif dan hasil pengacakan dengan *roulette wheel* untuk setiap topik yang ditampilkan pada Tabel 4.1 sampai Tabel 4.2.

a.1). *Word Intrusion Task* dengan *Stem*

Tabel4.1DeretAkumulatifTopik#0

Distribusi Kata	Probabilitas	Probabilitas X 1000	Deret Akumulatif	Roulette Wheel Range
ihsg	0.031	31	31	0-31
saham	0.020	20	51	32-51
indeks	0.008	8	59	52-59
harga	0.007	7	66	60-66
gabungan	0.006	6	72	67-72
melesat	0.005	5	77	73-77
bertambah	0.005	5	82	78-82
poin	0.005	5	87	83-87
perdagangan	0.005	5	92	88-92
kenaikan	0.005	5	97	93-97
anjlok	0.005	5	102	98-102
berharihari	0.005	5	107	103-107
kemarin	0.005	5	112	108-112
senin	0.004	4	116	113-116

Tabel 4.2 PengacakandenganRoulettewheelpadaTopik#0

Pengacakan ke-	Angka yang muncul	Kata
1	51	saham
	74	melesat
	83	poin
	100	anjlok
	115	senin
2	83	poin
	90	perdagangan
	74	melesat
	113	senin
	58	indeks
3	106	berharihari
	102	anjlok
	78	bertambah
	112	kemarin
	4	ihsg
4	52	indeks
	85	poin
	19	ihsg
	79	bertambah
	84	poin

b. Topic Intrusion task

Sesuai dengan langkah-langkah penyusunan materi kuesioner *Topic Intrusion Task*, implementasi kuesioner dilakukan dengan mempersiapkan opsi kuesioner terlebih dahulu dengan menyusun opsi yang terdiri dari distribusi kata dalam topik, sehingga jumlah opsi akan sama dengan jumlah topik. kemudian dokumen diklasifikasikan ke dalam topik-topik sesuai probabilitas tertingginya. Dokumen-dokumen yang dipilih untuk dicantumkan ke dalam kuesioner merupakan dokumen dengan nilai probabilitas dominan lebih dari 90% untuk memastikan kecenderungannya. Kemudian setiap dokumen tersebut dipetakan dengan opsi yang ada

b.1). *Topic Intrusion task* dengan *Stem*

Daftar Opsi Kuesioner *Topic Intrusion task* dengan *Stem* Eksperimen 1 dapat dilihat pada tabel Tabel 4.3.

Tabel 4.3 Daftar Opsi Kuesioner Topic Intrusion task dengan Stem

Opsi	DistribusiKata
A	'indeks', 'harga', 'saham', 'gabungan', 'ihsg', 'melesat', 'bertambah', 'poin', 'perdagangan', 'senin', 'kenaikan', 'ihsg', 'ihsg', 'anjlok', 'berharihari', 'januari', 'kemarin'
B	'indeks', 'harga', 'saham', 'gabungan', 'ihsg', 'ditutup', 'menguat', 'perdagangan', 'saham', 'bursa', 'efek', 'indonesia', 'bei', 'rabu', 'investor', 'asing', 'membukukan', 'beli', 'bersih', 'alias', 'net', 'buy', 'memborong', 'saham', 'sido', 'bbni'
C	'wintermar', 'offshore', 'marine', 'wins', 'menargetkan', 'tingkat', 'utilisasi', 'kapal', 'tahun', 'berharap', 'industri', 'jasa', 'perkapalan', 'bergairah', 'seiring', 'pekerjaan', 'hulu', 'minyak', 'gas', 'migas', 'bergeliat'
D	'mayora', 'indah', 'myor', 'mendorong', 'penjualan', 'ekspor', 'pemulihan', 'ekonomi', 'dinilai', 'berdampak', 'positif', 'penjualan', 'emiten', 'negeri'

Tabel 4.4 Daftar Opsi Kuesioner Topic Intrusion task dengan Stem Eksperimen2 untukTopik#0

Pengacakanke-	Angkayang muncul	Kata
1	95	Kenaikan
	73	Melesat
	64	Harga
	80	Bertambah
	102	Anjlok
	19	Ihsg
	111	Kemarin
2	55	Indeks
	89	Perdagangan
	31	Ihsg
	60	Harga
	76	Melesat
	6	Ihsg
	79	Bertambah
3	45	Saham
	90	Perdagangan
	71	Gabungan
	77	Gabungan
	64	Harga
	93	Kenaikan
	97	Kenaikan
4	96	Kenaikan
	110	Kemarin
	81	Bertambah
	26	Ihsg
	103	Berharihari
	43	Saham
	38	Saham

5	84	Poin
	97	Kenaikan
	61	Harga
	73	Melesat
	63	Harga
	98	Anjlok
	18	Ihsg

Tabel 4.5 Daftar DokumenTopic Intrusion task dengan Stem

No	Topik	Dokumen
1	t0	'indeks', 'harga', 'saham', 'gabungan', 'ihsg', 'melesat', 'bertambah', 'poin', 'perdagangan', 'senin', 'kenaikan', 'ihsg', 'ihsg', 'anjlok', 'berharihari', 'januari', 'kemarin'
2	t1	'analis', 'artha', 'sekuritas', 'indonesia', 'dennies', 'christopher', 'penguatan', 'ihsg', 'didorong', 'sektor', 'pertambangan', 'sektor', 'industri', 'dasar', 'menguat', 'sektor', 'infrastruktur'
3	t2	'indeks', 'harga', 'saham', 'gabungan', 'ihsg', 'ditutup', 'menguat', 'perdagangan', 'saham', 'bursa', 'efek', 'indonesia', 'bei', 'rabu', 'investor', 'asing', 'membukukan', 'beli', 'bersih', 'alias', 'net', 'buy', 'memborong', 'saham', 'sido', 'bbni'
4	t3	'kebijakan', 'pemberlakukan', 'pembatasan', 'kegiatan', 'masyarakat', 'ppkm', 'menekan', 'kinerja', 'mitra', 'adiperkasa', 'mapi', 'akibat', 'kebijakan', 'kunjungan', 'pusat', 'perbelanjaan', 'turun', 'penjualan', 'mapi', 'melambat'

b.2). *Topic Intrusion task* tanpa *Stem*

Tabel 4.6 Daftar Dokumen Topic Intrusion task tanpa Stem

No	Topik	Dokumen
1	t0	indeks harga saham gabungan ihsg melesat bertambah poin perdagangan senin kenaikan ihsg ihsg anjlok berharihari januari kemarin
2	t1	analis artha sekuritas indonesia dennies christopher penguatan ihsg didorong sektor pertambangan sektor industri dasar menguat sektor infrastruktur

3	t2	indeks harga saham gabungan ihsg ditutup menguat perdagangan saham bursa efek indonesia bei rabu investor asing membukukan beli bersih alias net buy memborong saham sido bbni
4	t3	kebijakan pemberlakukan pembatasan kegiatan masyarakat ppkm menekan kinerja mitra adiperkasa mapi akibat kebijakan kunjungan pusat perbelanjaan turun penjualan mapi melambat

4.1.5.3. Sistematika Kuesioner Uji Koherensi Topik

Tujuan dalam melakukan perancangan sistematika kuesioner adalah untuk memudahkan responden dalam memahami konteks kuesioner, sehingga dapat memberikan hasil yang lebih objektif.

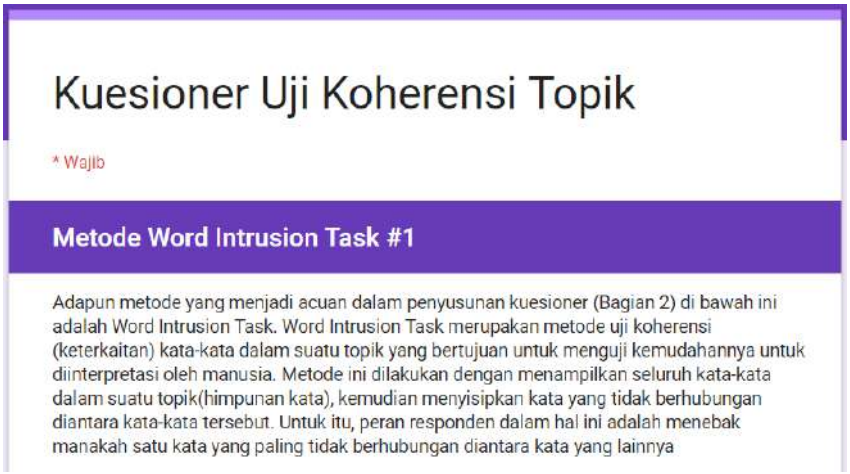
Sesuai dengan rancangan sistematika kuesioner uji koherensi topik, terdapat tiga bagian yaitu bagian judul, bagian deskripsi pembuka, dan bagian pertanyaan. Pada bagian judul berfungsi untuk mengidentifikasi aktivitas dan metode uji koherensi topik, pada bagian deskripsi pembuka berisi perkenalan penulis, tujuan pengumpulan data, deskripsi singkat metode pengumpulan data dan pengantar cara pengisian data. Berikut merupakan implementasi kuesioner dengan *tools Google Form* yang dapat dilihat pada Gambar 4.7.

Gambar 4.7 Bagian judul kuesioner uji koherensi topik

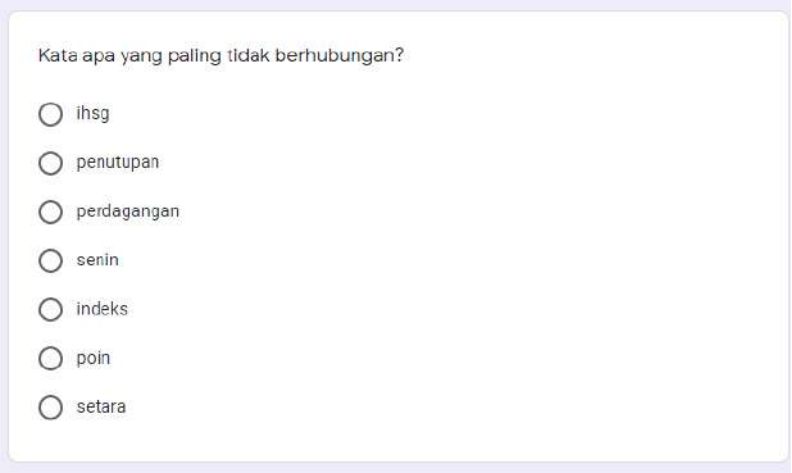
a. *Word Intrusion Task*

Pada kuesioner bagian *Word Intrusion Task*, pertanyaan ditujukan agar responden

memilih satu kata yang paling tidak berhubungan dari kumpulan opsi yang ditampilkan. Untuk setiap kali menampilkan kuesioner, opsi pilihan akan diacak urutannya. Berikut merupakan implementasi kuesioner



Gambar 4.8 Bagian deskripsi pembuka kuesioner uji koherensi topik bagian Word Intrusion Task

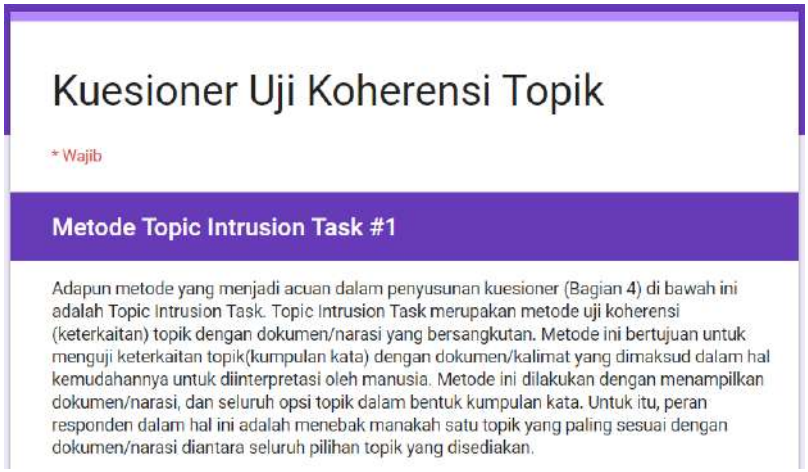


Gambar 4.9 Bagian pertanyaan kuesioner uji koherensi topik bagian Word Intrusion Task

b. *Topic Intrusion Task*

Pada kuesioner bagian *Topic Intrusion Task*, pertanyaan ditujukan agar responden memilih satu topik (himpunan kata) yang paling berhubungan dengan satu dokumen yang

ditampilkan. Untuk setiap kali menampilkan kuesioner, opsi pilihan tidak diacak urutannya. Berikut merupakan implementasi kuesioner pada bagian *Topic Intrusion task* yang dapat dilihat pada Gambar 4.10 sampai dengan Gambar 4.12



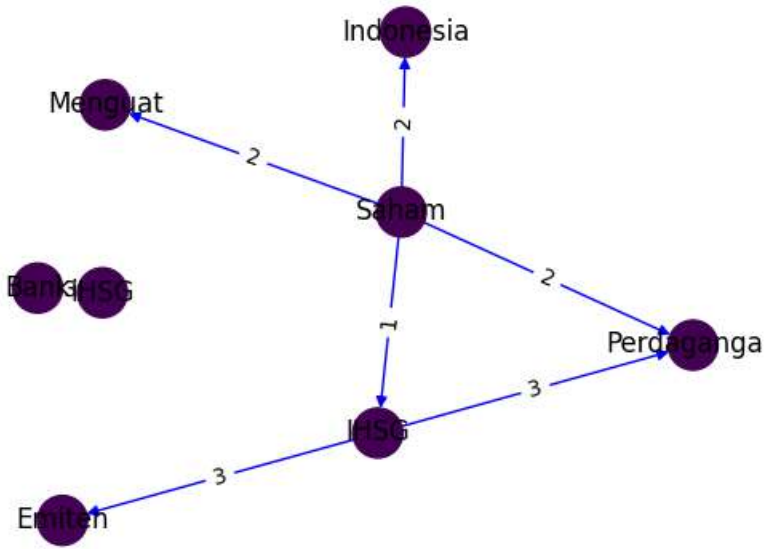
Gambar 4.10 Bagian deskripsi pembuka kuesioner uji koherensi topik bagian Topic Intrusion Task

Gambar 4.11 Bagian pertanyaan kuesioner uji koherensi topik bagian Topic Intrusion task Eksperimen1

Gambar 4.12 Bagian pertanyaan kuesioner uji koherensi topic bagian Topic Intrusion task Eksperimen2

4.1.6 Pachinko Allocation Model (PAM)

Ada satu node di bagian atas Directed Acyclic Graph (DAG) yang mendefinisikan distribusinode di tingkat kedua, yang disebut sebagai super topik. Setiap node di tingkat kedua mendefinisikan distribusi ke semua node di tingkat ketiga, atau sub-topik.Setiap sub-topik dipetakan ke satu distribusi di ataskosa kata. Oleh karena itu, hanya sub-topik sebenarnyamenghasilkan kata-kata. Super-topik mewakili kelompoktopik yang sering muncul. Gambar 4.13 menjelaskan hubungan antar topik berita saham.



Gambar 4.13 Hubungan antar topik berita saham menggunakan PAM

4.2 PEMBAHASAN

4.2.1 Mempersiapkan Data

4.2.1.1*Loading* Data

Jumlah data mentah yang dimuat sebagai sumber data masukan dari kontan.co.id sebanyak 181 baris berita saham.

4.2.1.2 Pembersihan Data

Jumlah data yang telah dibersihkan dari *url*, *mention*, *reserved words*, angka, karakter *non-alphanumeric* dan token sepanjang 1 digit dapat dilihat pada Gambar 4.1 hingga Gambar 4.6.

4.2.1.3 Pra-Proses Data

a. Pendefinisian *Stopwords*

Stopwords merupakan kata umum (*common words*) yang biasanya muncul dalam jumlah besar atau berfrekuensi tinggi dan dianggap tidak memiliki makna. Dalam kasus ini, pendefinisian *stopword* dilakukan dengan tiga justifikasi, yaitu

1. Berdasarkan pemaknaan kata sesuai sistem tata Bahasa Indonesia baku.

2. Berdasarkan analisis kemunculan kata dalam daftar intensitas kemunculan kata (1000 kata tertinggi).

3. Kode komunikasi internal pasar saham.

Dalam sistem tata Bahasa Indonesia baku, terdapat istilah kelas kata, yaitu istilah linguistik kelas atau golongan (kategori) kata berdasarkan bentuk, fungsi, atau maknanya (KBBI). Diantara kelas-kelas kata yang ada, terdapat beberapa kelas kata yang memiliki arti gramatikal, namun tidak memiliki arti leksikal. Dalam artian lain, bahwa suatu kata jika berdiri sendiri tanpa disertai dengan kata yang diterangkan, kata tersebut tidak memiliki makna, sehingga kelas kata yang dianggap memenuhi syarat untuk digolongkan ke dalam *stopword* dapat dilihat pada Tabel 4.7 Kelas kata yang memenuhi syarat sebagai *stopword.*

Tabel 4.7 Kelas kata yang memenuhi syarat sebagai stopword

Kelas Kata		Definisi
Adverbia		Kata Keterangan
Pronomina		Kata Ganti
Numeralia		Kata Bilangan
Kata Tugas	Preposisi	Kata Depan
	Konjungsi	Kata Sambung
	Artikula	Kata Sandang
	Interjeksi	Kata Seru
	Partikel	Partikel

Dalam melakukan pertimbangan berdasarkan analisis kemunculan kata dalam daftar intensitas kemunculan kata (1000 kata tertinggi), dapat memperhitungkan kelas kata di luar kelas kata yang dicantumkan. Adapun kata-kata yang dihilangkan merupakan kata yang memiliki intensitas kemunculan tinggi namun tidak memiliki makna.

b. Pembentukan Model *LDA*

Dalam melakukan pembentukan model *LDA*, terdapat model yang dibentuk, model yang melalui tahap *Stemming* dan model yang tidak melalui tahap *Stemming*

b.1). Hasil Pembentukan Model *LDA* dengan *Stemming*

Distribusi probabilitas kata dalam 4 Topik (20 kata per topik) dari model yang melalui proses *Stemming* dapat dilihat pada Tabel 4.8 Hasil Pembentukan Model *LDA* dengan *Stemming*.

Tabel 4.8 Hasil Pembentukan Model *LDA* dengan *Stemming*

topik #0:	topik #1:	topik #2:	topik #3:
0.021*indeks	0.017* analis	0.035* indeks	0.073* kebijakan
0.010*harga	0.016* artha	0.023* harga	0.069* pemberlakuan
0.008*saham	0.015* sekuritas	0.019* saham	0.044* pembatasan
0.007*gabungan	0.014* indonesia	0.015* gabungan	0.030* kegiatan
0.006*ihsg	0.014* penguatan	0.015* ihsg	0.027* masyarakat
0.005*melesat	0.011* ihsg	0.015* ditutup	0.024* menekan
0.005*bertambah	0.011* didorong	0.015* menguat	0.021* kinerja
0.005*poin	0.010* sector	0.014* perdagangan	0.020* mitra
0.005*perdagangan	0.010* pertambangan	0.014* saham	0.017* adi
0.005*senin	0.009* sector	0.013* bursa	0.016* perkasa
0.004*kenaikan	0.009* industri	0.013* efek	0.012* mapi
0.004*ihsg	0.008* dasar	0.012* indonesia	0.012* akibat
0.004*ihsg	0.007* menguat	0.012* bei	0.012* kebijakan
0.004*anjlok	0.006* sector	0.010* rabu	0.012* kunjungan
0.004*berharihari	0.006* infrastruktur	0.010* investor	0.012* pusat
0.004*januari	0.006* mengatakan	0.010* asing	0.010* perbelanjaan
0.004*kemarin	0.005* didorong	0.010* membukukan	0.010* turun
0.004*poin	0.005* naik	0.009* beli	0.010* penjualan
0.003*terjadi	0.005* denies	0.009* bersih	0.009* mapi
0.003*akhir	0.005* christopher	0.008* memborong	0.009* melambat

b.2). Hasil Pembentukan Model *LDA* tanpa *Stemming*

Distribusi probabilitas kata dalam 4 Topik (20 kata per topik) dari model yang tidak melalui proses *Stemming* dapat dilihat pada Tabel 4.9 Hasil Pembentukan Model *LDA* tanpa *Stemming*.

Tabel 4.9 Hasil Pembentukan Model *LDA* tanpa *Stemming*

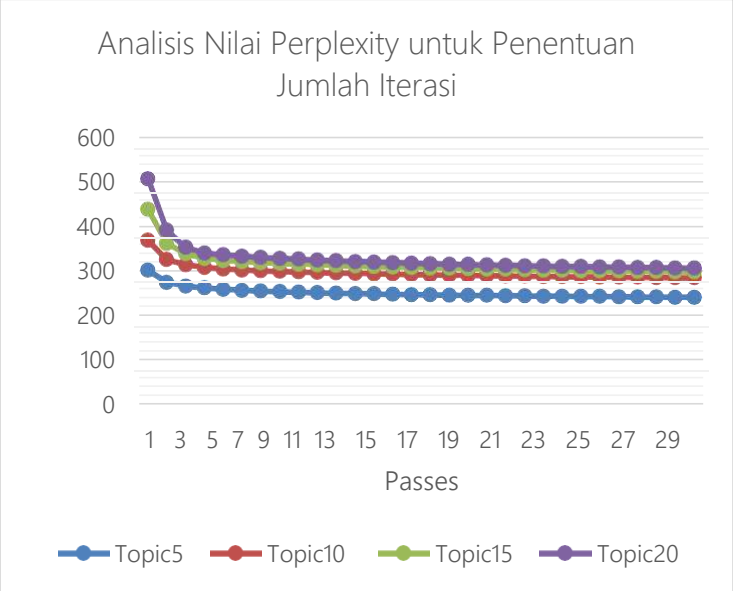
topik #0:	topik #1:	topik #2:	topik #3:
0.014* indeks	0.030* analis	0.070* indeks	0.021* kebijakan
0.008* harga	0.024* artha	0.068* harga	0.015* pemberlakuan
0.007* saham	0.021* sekuritas	0.043* saham	0.014* pembatasan
0.006* gabungan	0.019* indonesia	0.030* gabungan	0.014* kegiatan
0.006* ihsg	0.018* mengatakan	0.025* ihsg	0.012* masyarakat
0.006* melesat	0.017* penguatan	0.024* ditutup	0.011* ppkm
0.006* bartambah	0.013* ihsg	0.017* menguat	0.008* menekan

0.005* poin	0.013* didorong	0.017* akhir	0.008* kinerja
0.005* perdagangan	0.013* sektor	0.016* perdagangan	0.007* pt
0.004* senin	0.013* pertambangan	0.015* saham	0.007* mitra
0.003* kenaikan	0.011* naik	0.012* bursa	0.007* adi
0.003* ihsg	0.010* sektor	0.011* efek	0.007* perkasa
0.003* terjadi	0.009* industri	0.011* Indonesia	0.006* tbk
0.003* setelah	0.009* dasar	0.009* bei	0.006* tahun
0.003* ihsg	0.009* menguat	0.009* rabu	0.006* akibat
0.003* anjlok	0.009* sektor	0.009* investor	0.006* jumlah
0.003* berharihari	0.008* infrastruktur	0.009* asing	0.005* kunjungan
0.003* akhir	0.008* naik	0.009* membukukan	0.005* pusat
0.002* januari	0.008* dennies	0.008* beli	0.005* perbelanjaan
0.002* kemarin	0.008* christopher	0.008* bersih	0.005* turun

c. Validasi Model Topik

c.1). Penentuan Jumlah Iterasi

Dalam melakukan penentuan jumlah iterasi (*passes*), metode yang digunakan adalah dengan melakukan analisis nilai *perplexity*. Analisis nilai *perplexity* dilakukan dengan cara menjalankan pemodelan terhadap topik dengan setidaknya tiga parameter topik yang berbeda dengan nilai *passes* mula-mula 30. Dalam kasus ini, parameter jumlah topik yang dipilih adalah 5, 10, 15 dan 20. Hasil nilai *perplexity* yang muncul dari masing-masing parameter jumlah topik ini kemudian dicatat kemudian divisualisasikan pada Gambar4.14.



Gambar 4.14 Analisis Nilai Perplexity untuk Penentuan Jumlah Iterasi

Berdasarkan visualisasi Gambar 4.14, dapat dilihat bahwa nilai *perplexity* sudah mencapai kondisi cenderung stabil pada *passes* ke 10 untuk keseluruhan parameter jumlah topik, sehingga dapat disimpulkan bahwa iterasi yang digunakan adalah 10.

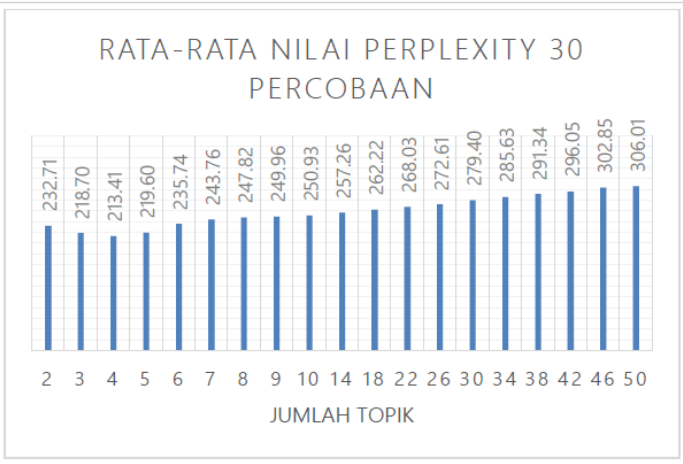
4.2.2. Penentuan Jumlah Topik

Penentuan jumlah topik dilakukan dengan menganalisis nilai *perplexity*, namun analisis nilai *perplexity* dalam konteks untuk penentuan jumlah topik dilakukan dengan melakukan eksperimen pada parameter jumlah topik dalam rentang nilai yang lebih luas, dalam hal ini rentang nilai yang digunakan dalam eksperimen ditampilkan dalam tabel di bawah. Analisis nilai *perplexity* dalam konteks untuk penentuan jumlah topik dilakukan dengan melakukan running sebanyak 30 kali untuk mendapatkan rata-rata nilai *perplexity* yang akurat untuk masing-masing parameter jumlah topik.

4.2.2.1. Penentuan Jumlah Topik dengan *Stemming*

Hasil eksperimen analisis nilai *perplexity* untuk penentuan jumlah topik dengan *Stemming* dapat dilihat pada

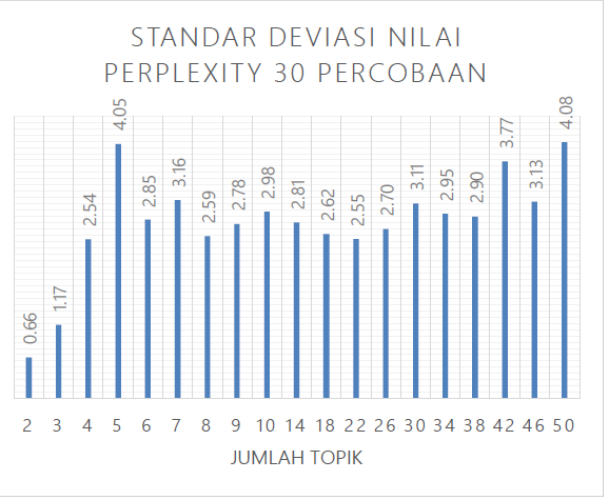
4.2.2.1.1. Analisis Hasil Eksperimen Nilai *Perplexity* untuk Penentuan Jumlah Topik



Gambar 4.15 Rata-rata nilai *Perplexity* 30 percobaan

Berdasarkan Gambar 4.15, nilai *perplexity* terendah terdapat pada jumlah topik 4 yaitu 213.41, dan tren nilai *perplexity* meningkat untuk jumlah topik yang semakin tinggi,

sehingga 4 topik merupakan jumlah topik terbaik berdasarkan analisis nilai perplexity.



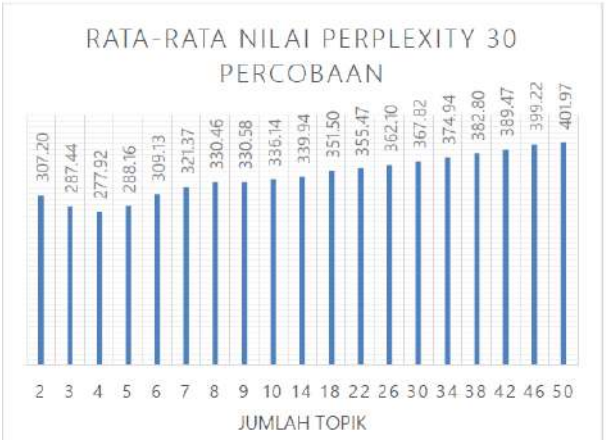
Gambar4.16 Standar Deviasi Nilai *Perplexity* 30 Percobaan

Berdasarkan Diagram 4.16, diketahui bahwa standar deviasi nilai *perplexity* 4 topik untuk 30 kali percobaan adalah 2.54. jika dibandingkan dengan nilai rata-rata standar deviasi keseluruhan yaitu 2.81, angka 2.54 dapat dikatakan cukup stabil.

4.2.2.2. Penentuan Jumlah Topik tanpa *Stemming*

Hasil eksperimen analisis nilai *perplexity* untuk penentuan jumlah topik tanpa *Stemming* dapat dilihat pada

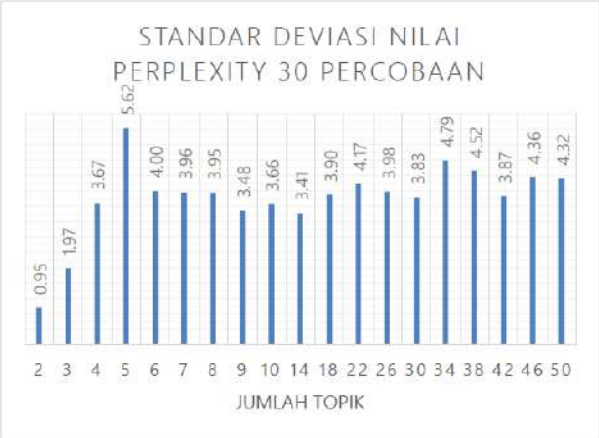
4.2.2.2.1. Analisis Hasil Eksperimen Nilai *Perplexity* untuk Penentuan Jumlah Topik



Gambar4.17 Rata-rata nilai *Perplexity* 30 percobaan

Berdasarkan Gambar 4.17, nilai *perplexity* terendah terdapat pada jumlah topik 4 yaitu 277.92, dan tren nilai *perplexity* meningkat untuk jumlah topik yang semakin tinggi,

sehingga 4 topik merupakan jumlah topik terbaik berdasarkan analisis nilai perplexity.



Gambar4.18 Standar Deviasi Nilai *Perplexity* 30 Percobaan

Berdasarkan Gambar 4.18, diketahui bahwa standar deviasi nilai *perplexity* 4 topik untuk 30 kali percoban adalah 3.67. Jika dibandingkan dengan nilai rata-rata standar deviasi keseluruhan yaitu 3.81, angka 3.67 dapat dikatakan cukup stabil.

4.2.3. Jumlah Distribusi Dokumen Per Topik

Distribusi dokumen per topik dilakukan sebagai bentuk analisis terhadap dominansi antara satu topik dengan topik yang lain dengan parameter jumlah dokumen per topik. Analisis distribusi dokumen dilakukan dengan melakukan klasifikasi dokumen mentah ke dalam model yang telah dibentuk. Peran analisis distribusi dokumen per topik dalam tahap validasi adalah untuk melihat apakah persebaran dokumen terpusat pada salah satu dari topik yang dibentuk atau tersebar secara merata. Analisis distribusi dokumen per topik dilakukan untuk topik yang dihasilkan melalui tahap *Stemming* dan yang tidak melalui tahap *Stemming.*

4.2.3.1. Distribusi Dokumen Per Topik dengan *Stemming*

Distribusi dokumen mayoritas berada pada topik #3, diikuti dengan topik #2

4.2.3.2. Distribusi Dokumen Per Topik tanpa *Stemming*

Distribusi dokumen mayoritas berada pada topik #1, diikuti dengan topik #0

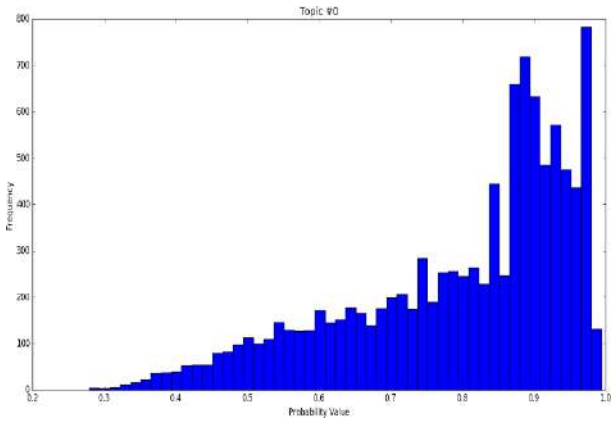
4.2.4. Visualisasi Distribusi Probabilitas Dokumen per Topik

Analisis distribusi dokumen per topik dilakukan sebagai bentuk analisis persebaran dokumen per topik dengan tujuan untuk mengetahui apakah terdapat dominansi pada satu topik terhadap topik lainnya. Namun analisis ini perlu dilakukan analisis lanjutan untuk mengetahui tingkat keyakinan suatu dokumen terhadap topik yang menjadi golongannya melalui probabilitas dokumen terhadap topiknya. Untuk itu perlu dilakukan analisis distribusi probabilitas dokumen per topik yang divisualisasikan dalam bentuk histogram agar lebih mudah dalam melakukan pengamatan. Analisis distribusi probabilitas dokumen per topik dilakukan untuk topik yang dihasilkan oleh model yang melalui tahap *Stemming* dan yang tidak melalui tahap *Stemming*

4.2.4.1. Distribusi Probabilitas Dokumen per Topik dengan *Stemming*

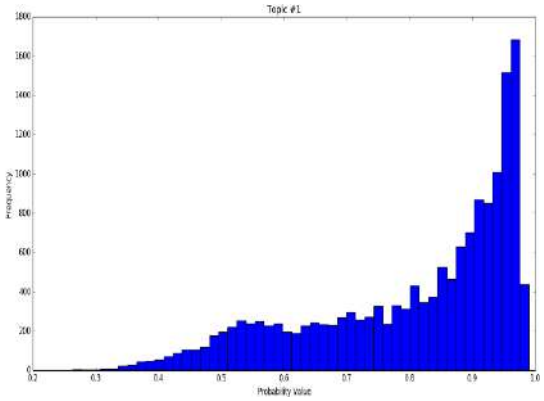
Berikut merupakan histogram distribusi probabilitas dokumen per topik dengan *Stemming*

4.2.4.1.1. Histogram



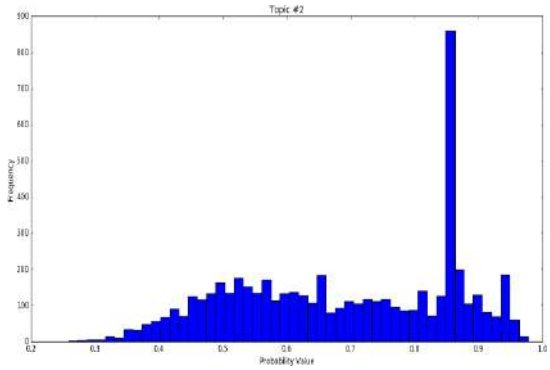
Gambar 4.19 Histogram distribusi probabilitas dokumen pada Topik #0

Berdasarkan Gambar 4.19, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.6 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.85 hingga 0.99



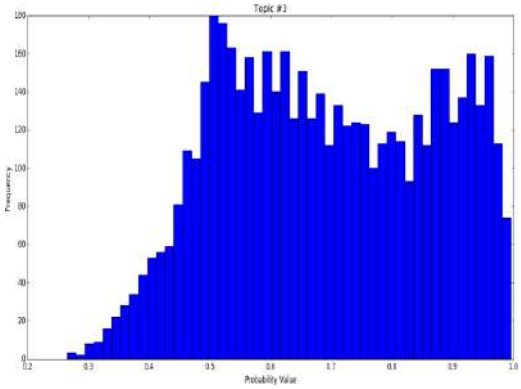
Gambar 4.20 Histogram distribusi probabilitas dokumen pada Topik #1

Berdasarkan Gambar 4.20, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.9 hingga 0.99.



Gambar4.21 Histogram distribusi probabilitas dokumen pada Topik #2

Berdasarkan Gambar4.21, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.85 hingga 0.87



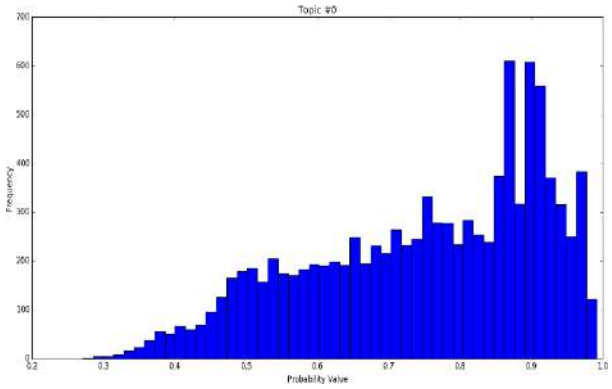
Gambar4.22 Histogram distribusi probabilitas dokumen pada Topik #3

Berdasarkan Gambar 4.22, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0. 5 hingga 0.99

4.2.4.2. Distribusi Probabilitas Topik Dokumen per Topik dengan *Stemming*

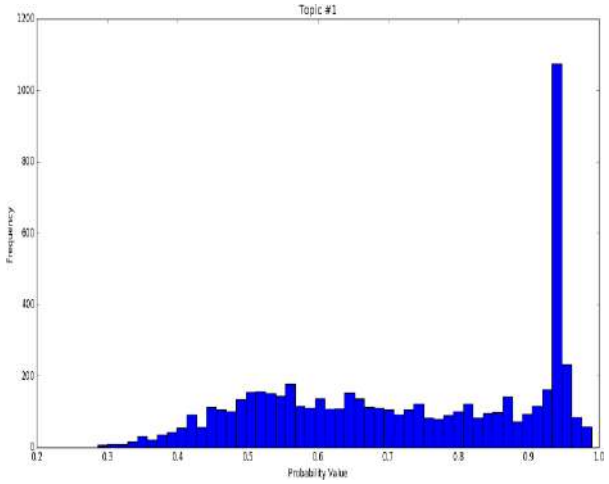
Berikut merupakan histogram distribusi probabilitas dokumen per topik dengan *Stemming*

4.2.4.2.1. Histogram



Gambar 4.23 Histogram distribusi probabilitas dokumen pada Topik #0

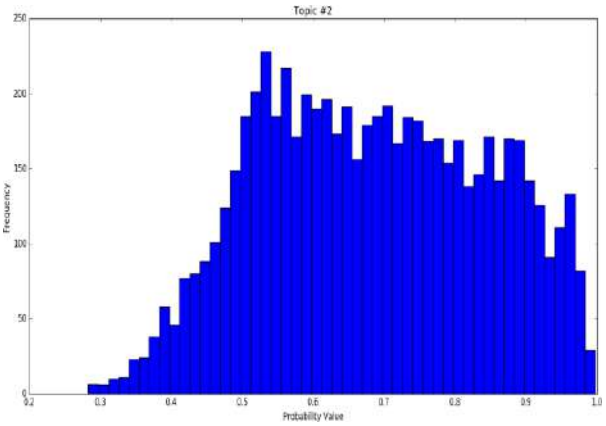
Berdasarkan Gambar 4.23, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.85 hingga 0.95



Gambar 4.24 Histogram distribusi probabilitas dokumen pada Topik #1

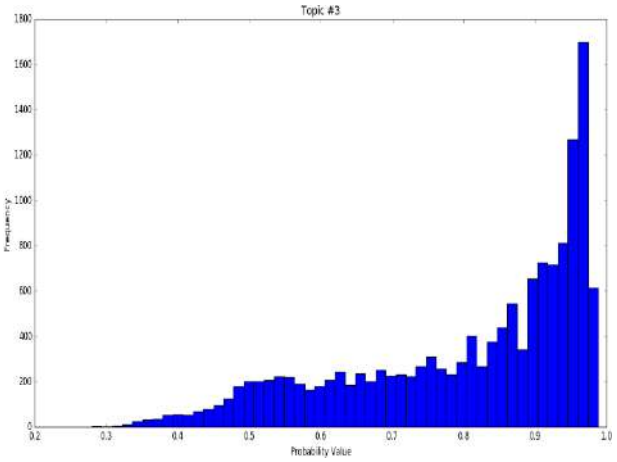
Berdasarkan Gambar 4.24, distribusi probabilitas dinilai meyakinkan dengan

mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.93 hingga 0.96



Gambar 4.25 Histogram distribusi probabilitas dokumen pada Topik #2

Berdasarkan Gambar 4.25, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.5 hingga 0.9



Gambar 4.26 Histogram distribusi probabilitas dokumen pada Topik #3

Berdasarkan Gambar4.26, distribusi probabilitas dinilai meyakinkan dengan mayoritas dokumen memiliki probabilitas topik lebih dari 0.5 dan tingginya jumlah dokumen yang memiliki probabilitas dalam rentang 0.9 hingga 0.99.

4.2.5 Hasil Uji Koherensi Topik

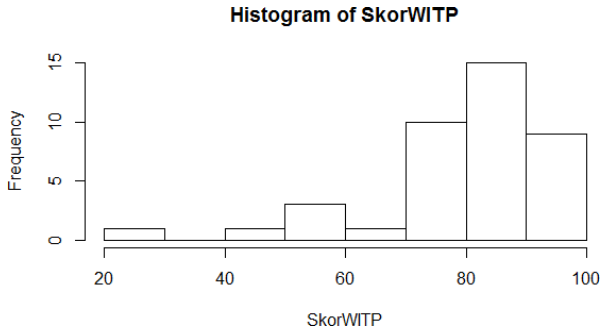
Penyebaran kuesioner uji koherensi topik dilakukan sebanyak dua kali.

Penyebaran pertama dilakukan pada tanggal tanggal 30 Mei - 4 Juni 2021dengan total 102 responden dan penyebaran kedua dilakukan pada tanggal 1 – 5Agustus 2021 dengan total 122 responden.

4.2.5.1. Histogram

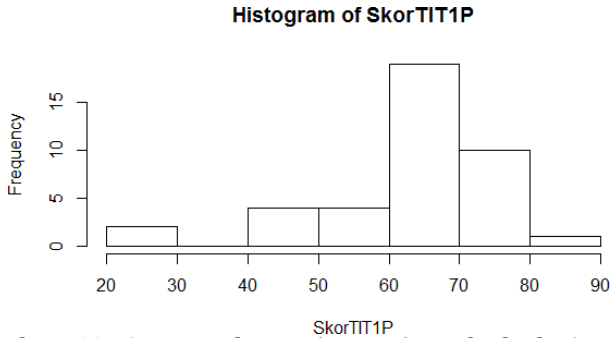
4.2.5.1.1. Berbasis Pertanyaan

Analisis ini bertujuan untuk mengetahui tingkat kesulitan dari pertanyaan yang diajukan secara keseluruhan. Tingkat kesulitan pertanyaan diukur dengan persentase ketepatan responden dalam menjawab pertanyaan sesuai *task* pada uji koherensi topik. Berikut merupakan histogram rekapitulasi hasil uji koherensi topik yang dirangkum berbasis skor dari setiap pertanyaan.



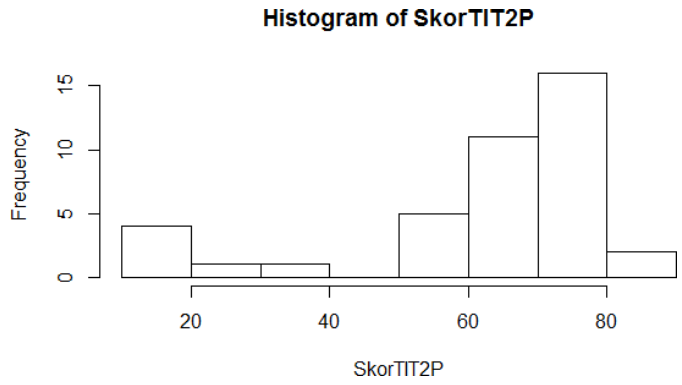
Gambar 4.27 Histogram skor *Word Intrusion task* berbasis pertanyaan

Gambar 4.27 menampilkan distribusi skor untuk pertanyaan kuesioner yang tergolong dalam *Word Intrusion Task*. Berdasarkan distribusi yang ditampilkan dalam histogram, mayoritas persentase ketepatan untuk setiap pertanyaan berada pada rentang 70% - 100%, sehingga pertanyaan-pertanyaan yang tergolong dalam kategori *Word Intrusion task* dapat dikatakan sangat mudah.



Gambar 4.28 Histogram skor *Topic Intrusion task* 1 berbasis pertanyaan

Gambar 4.28 menampilkan distribusi skor untuk pertanyaan kuesioner yang tergolong dalam *Topic Intrusion task* 1. Berdasarkan distribusi yang ditampilkan dalam histogram, mayoritas persentase ketepatan untuk setiap pertanyaan berada pada rentang 60% - 80%, sehingga pertanyaan-pertanyaan yang tergolong dalam kategori *Topic Intrusion task* 1 dapat dikatakan mudah.

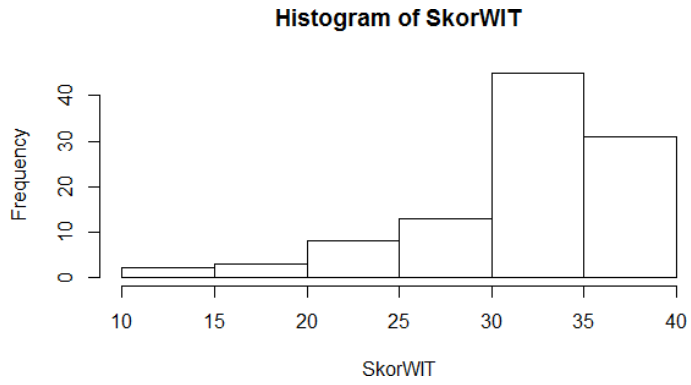


Gambar 4.29 Histogram skor *Topic Intrusion task* 2 berbasis pertanyaan

Gambar 4.29 menampilkan distribusi skor untuk pertanyaan kuesioner yang tergolong dalam *Topic Intrusion task* 2. Berdasarkan distribusi yang ditampilkan dalam histogram, mayoritas persentase ketepatan untuk setiap pertanyaan berada pada rentang 60% - 80%, sehingga pertanyaan-pertanyaan yang tergolong dalam kategori *Topic Intrusion task* 2 dapat dikatakan mudah.

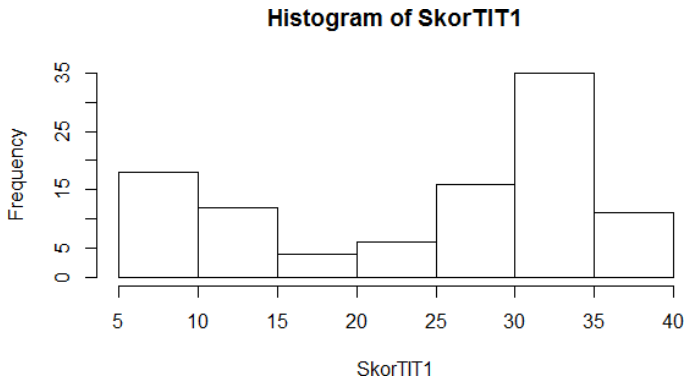
4.2.5.1.2. Berbasis Responden

Analisis ini bertujuan untuk mengetahui tingkat interpretasi responden terhadap topik melalui pertanyaan yang diajukan secara keseluruhan. tingkat interpretasi responden diukur dengan jumlah pertanyaan yang dijawab secara tepat untuk masing-masing *task* pada uji koherensi topik. Untuksetiap *task* dalam uji koherensi topik, terdapat 40 pertanyaan yang diajukan, sehingga skor maksimal yang dapat diperoleh oleh responden adalah 40. Berikut merupakan histogram rekapitulasi hasil uji koherensi topik yang dirangkum berbasis responden.



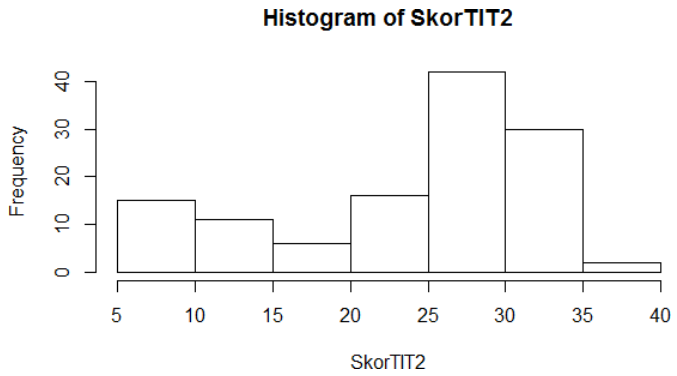
Gambar 4.30 Histogram skor *Word Intrusion task* berbasis responden

Gambar 4.30 menampilkan distribusi skor ketepatan responden dalam menjawab pertanyaan yang tergolong dalam *Word Intrusion Task*. Berdasarkan distribusi yang ditampilkan dalam histogram, sekitar 75 responden menjawab secara tepat 30-40 pertanyaan, sehingga tingkat interpretasi responden terhadap topik terhadap pertanyaan-pertanyaan yang tergolong dalam kategori *Word Intrusion task* dapat dikatakan sangat baik.



Gambar 4.31 Histogram skor *Topic Intrusion task* 1 berbasis responden

Gambar 4.31 menampilkan distribusi skor ketepatan responden dalam menjawab pertanyaan yang tergolong dalam *Topic Intrusion task* 1. Berdasarkan distribusi yang ditampilkan dalam histogram, sekitar 35 responden menjawab secara tepat 30-35 pertanyaan, diikuti sekitar 17 responden menjawab secara tepat 5-10 pertanyaan, sehingga tingkat interpretasi responden terhadap topik terhadap pertanyaan-pertanyaan yang tergolong dalam kategori *Topic Intrusion task* 1 dapat dikatakan baik.



Gambar 4.32 Histogram skor *Topic Intrusion task* 2 berbasis responden

Gambar 4.32 menampilkan distribusi skor ketepatan responden dalam menjawab pertanyaan yang tergolong dalam *Topic Intrusion task* 2. Berdasarkan distribusi yang ditampilkan dalam histogram, sekitar 75 responden menjawab secara tepat 25-35 pertanyaan, sehingga tingkat interpretasi responden terhadap topik terhadap pertanyaan-pertanyaan yang tergolong dalam kategori *Topic Intrusion task* 1 dapat dikatakan baik.

4.2.5.2. Uji Hipotesis

Uji Hipotesis dilakukan dengan tujuan untuk membuktikan hipotesis (dugaan sementara) berdasarkan metode statistika. Dalam melakukan uji hipotesis, tahap-tahap yang dilakukan adalah mendefinisikan pernyataan hipotesis, melakukan uji *variance*, melakukan uji *means* kemudian menyusun kesimpulan berdasarkan kedua uji tersebut. Kesimpulan uji *variance* dan uji *means* dilakukan dengan membandingkan *p-value* dengan tingkat signifikansi. Tingkat signifikansi (α) merupakan probabilitas penolakan hipotesis nol ketika hipotesis tersebut benar. Tingkat signifikansi akan menjadi pembanding terhadap nilai *p-value* untuk menentukanapakah hipotesis nol diterima atau ditolak, hipotesis nol ditolak jika nilai *p-value* kurang dari nilai α. Pada kasus pengujian hipotesis ini dipilih tingkat signifikansi $\alpha = 0.05$. Untuk mempertajam analisis, uji ANOVA dibagi berdasarkan metode uji koherensi topik dengan dibantu *tools* Minitab 17. Untuk melakukan uji hipotesis, diperlukan data yang terdapat pada Tabel 4.10.

Tabel 4.10 Tabel Pendahuluan Uji Hipotesis

	Rata-rata		StDev	
	#	%	#	%
WIT	31.96	79.90	5.57	14.00
WIT-*Stem*	15.66	78.28	2.67	13.33
WIT+*Stem*	16.30	81.52	3.99	19.96
TIT1	24.86	62.16	10.31	25.78
TIT 1-*Stem*	12.28	61.42	4.44	22.19
TIT 1+*Stem*	13.26	66.32	4.67	23.35
TIT2	24.32	60.80	8.33	20.83
TIT 1-*Stem*	12.13	60.66	5.07	25.35
TIT 1+*Stem*	12.88	64.39	5.93	29.66

4.2.5.2.1. Uji Hipotesis *Word Intrusion task* dan *Topic Intrusion task* 1

Pernyataan Hipotesis *Word Intrusion task* dan *Topic Intrusion task* 1

H_0: σ *Word Intrusion task* = σ *Topic Intrusion task* 1

H_1: σ *Word Intrusion task* ≠ σ *Topic Intrusion task* 1

Uji *Variance*

Test and CI for Two Variances: WIT, TIT 1

Statistics

Variable	N	StDev	Variance	95% CI for StDevs
WIT	102	5.599	31.345	(4.549, 7.025)
TIT 1	102	10.310	106.298	(9.531, 11.371)

Ratio of standard deviations = 0.543
Ratio of variances = 0.295

Tests

Method	DF1	DF2	Test Statistic	P-Value
Bonett	1	–	43.73	0.000
Levene	1	202	27.99	0.000

Gambar 4.33 Uji *variance* WIT dan TIT 1

Hasil pengujian yang ditampilkan pada Gambar 4.33 menunjukkan bahwa *p-value* dengan metode *Bonett* dan *Levene* senilai 0.000, yang mana nilai tersebut kurang dari nilai α, sehingga dapat dikatakan bahwa nilai *variance* dari *Word Intrusion task* dan *Topic Intrusion task* 1 dinyatakan berbeda.

Test and CI for Two Variances: TIT 1, TIT 2

```
Statistics

                                      95% CI for
Variable    N   StDev  Variance       StDevs
TIT 1     102  10.310   106.298  (9.531, 11.371)
TIT 2     122   8.333    69.442  (7.485,  9.429)

Ratio of standard deviations = 1.237
Ratio of variances = 1.531


Tests

                       Test
Method  DF1  DF2  Statistic  P-Value
Bonett    –    –          –    0.003
Levene    1  222       5.57    0.019
```

Uji *Means*

Two-Sample T-Test and CI: WIT, TIT 1

```
Two-sample T for WIT vs TIT 1

         N   Mean  StDev  SE Mean
WIT    102  31.96   5.60     0.55
TIT 1  102   24.9   10.3      1.0


Difference = μ (WIT) - μ (TIT 1)
Estimate for difference:  7.10
95% CI for difference:  (4.80, 9.39)
T-Test of difference = 0 (vs ≠): T-Value = 6.11  P-Value = 0.000  DF = 155
```

Gambar 4.34 Uji *means* WIT dan TIT 1

Hasil pengujian yang ditampilkan pada Gambar 4.34 menunjukkan bahwa *p-value* senilai 0.000, yang mana nilai tersebut lebih kecil dari nilai α, nilai tersebut menunjukkan bahwa terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol ditolak.

Berdasarkan uji hipotesis *Word Intrusion task* dan *Topic Intrusion task* 1, disimpulkan bahwa kedua sampel memiliki *variance* dan *Means* yang berbeda signifikan, sehingga diperlukan analisis lebih lanjut terhadap kesimpulan ini. Terdapat banyak kemungkinan rendahnya nilai TIT 1 jika dibandingkan dengan WIT, dalam kasus ini, hal yang paling disoroti adalah faktor kelelahan responden dan responden kesulitan dalam mengisi kuesioner akibat tampilan visual yang kurang baik.

Untuk menguji kedua asumsi yang muncul, dirasa perlu melakukan beberapa tindakan. Pertama, penyebaran ulang kuesioner khusus bagian *Topic Intrusion task* untuk mengantisipasi faktor kelelahan responden sehingga didapatkan tingkat fokus yang sama seperti WIT pada penyebaran kuesioner tahap pertama. Kedua, memperbaiki tampilan visual kuesioner untuk menjaga agar tidak terlalu banyak informasi yang ditampilkan dalam satu pertanyaan yang dapat membingungkan responden. Perbaikan tampilan visual dilakukan dengan membatasi pilihan jawaban yang sebelumnya terdapat 20 kata dalam satu topik menjadi 7 kata agar tetapdalam satu baris sesuai dengan perencanaan materi kuesioner.

4.2.5.2.2. Uji Hipotesis *Topic Intrusion task* 1 dan *Topic Intrusion task* 2

Pernyataan Hipotesis *Topic Intrusion task* 1 dan *Topic Intrusion task* 2

H_0: σ *Topic Intrusion task* 1 = σ *Topic Intrusion task* 2

H_1: σ *Topic Intrusion task* 1 ≠ σ *Topic Intrusion task* 2

Uji *Variance*

Test and CI for Two Variances: TIT 1, TIT 2

Statistics

Variable	N	StDev	Variance	95% CI for StDevs
TIT 1	102	10.310	106.298	(9.531, 11.371)
TIT 2	122	8.333	69.442	(7.485, 9.429)

Ratio of standard deviations = 1.237
Ratio of variances = 1.531

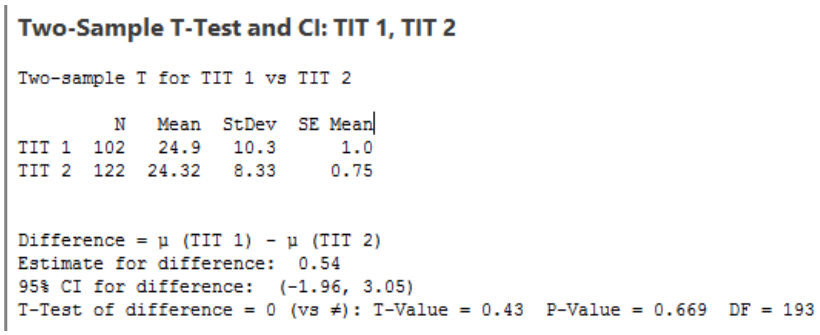
Tests

Method	DF1	DF2	Test Statistic	P-Value
Bonett	–	–	–	0.003
Levene	1	222	5.57	0.019

Gambar 4.35 Uji *variance* TIT 1 dan TIT 2

Hasil pengujian yang ditampilkan pada Gambar 4.35 menunjukkan bahwa *p-value* dengan metode *Bonett* senilai 0.003 dan metode *Levene* senilai 0.019, yang mana

kedua nilai tersebut kurang dari nilai α, sehingga dapat dikatakan bahwanilai *variance* dari *Topic Intrusion task* 1 dan *Topic Intrusion task* 2 dinyatakan berbeda.



Two-Sample T-Test and CI: TIT 1, TIT 2

```
Two-sample T for TIT 1 vs TIT 2

         N   Mean  StDev  SE Mean
TIT 1  102   24.9   10.3      1.0
TIT 2  122  24.32   8.33     0.75


Difference = μ (TIT 1) - μ (TIT 2)
Estimate for difference:  0.54
95% CI for difference:  (-1.96, 3.05)
T-Test of difference = 0 (vs ≠): T-Value = 0.43  P-Value = 0.669  DF = 193
```

Gambar 4.36 Uji *Means* TIT 1 dan TIT 2

Hasil pengujian yang ditampilkan pada Gambar 4.36 menunjukkan bahwa *p-value* senilai 0.669, yang mana nilai tersebut lebih besar dari nilai α, nilai tersebut menunjukkan bahwa tidak terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol gagal tolak. Berdasarkan uji hipotesis *Topic Intrusion task* 1 dan Topik Intusion *task* 2, disimpulkan bahwa kedua sampel memiliki *variance* yang berbeda namun *Means* yang sama, sehingga hasil ini dapat memberikan gambaran terhadap kedua asumsi yang muncul pada analisis lanjutan uji hipotesis *Word Intrusion task* dan Topik Intusion *task* 1.

Asumsi adanya kelelahan responden dan adanya kesulitan responden terhadap tampilan kuesioner bukan merupakan faktor yang berpengaruh terhadap kedua sampel, hal ini dibuktikan dengan uji *Means* yang menunjukkan bahwa H_0 gagal tolak, atau dalam artian lain *Means* untuk kedua sampel dinyatakan sama.

4.2.5.2.3. Uji Hipotesis *Word Intrusion task* dan *Topic Intrusion task* 2

Pernyataan Hipotesis *Word Intrusion task* dan *Topic Intrusion task* 2

H_0: σ *Word Intrusion task* = σ *Topic Intrusion task* 2

H_1: σ *Word Intrusion task* ≠ σ *Topic Intrusion task* 2

Test and CI for Two Variances: WIT, TIT 2

Statistics

Variable	N	StDev	Variance	95% CI for StDevs
WIT	102	5.599	31.345	(4.549, 7.025)
TIT 2	122	8.333	69.442	(7.485, 9.429)

Ratio of standard deviations = 0.672
Ratio of variances = 0.451

Tests

Method	DF1	DF2	Test Statistic	P-Value
Bonett	—	—	—	0.000
Levene	1	222	11.91	0.001

Gambar 4.37 Uji *variance* WIT dan TIT 2

Hasil pengujian yang ditampilkan pada Gambar 4.37 menunjukkan bahwa *p-value* dengan metode *Bonett* senilai 0.000 dan metode *Levene* senilai 0.001, yang mana kedua nilai tersebut kurang dari nilai α, sehingga dapat dikatakan bahwa nilai *variance* dari *Word Intrusion task* dan *Topic Intrusion task* 1 dinyatakan berbeda.

Uji Means

Two-Sample T-Test and CI: WIT, TIT 2

Two-sample T for WIT vs TIT 2

	N	Mean	StDev	SE Mean
WIT	102	31.96	5.60	0.55
TIT 2	122	24.32	8.33	0.75

Difference = μ (WIT) - μ (TIT 2)
Estimate for difference: 7.641
95% CI for difference: (5.796, 9.487)
T-Test of difference = 0 (vs ≠): T-Value = 8.16 P-Value = 0.000 DF = 212

Gambar 4.38 Uji *Means* WIT 1 dan TIT 2

Hasil pengujian yang ditampilkan pada Gambar 4.38 menunjukkan bahwa *p-value* senilai 0.000, yang mana nilai tersebut lebih kecil dari nilai α, nilai tersebut menunjukkan bahwa terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol ditolak. Berdasarkan seluruh uji hipotesis yang telah dilakukan, dapat disimpulkan bahwa metode *Word Intrusion task* secara inheren lebih mudah diinterpretasi

oleh manusia jika dibandingkan dengan metode *Topic Intrusion task* baik TIT 1 maupun TIT 2.

4.2.5.2.4. Uji Hipotesis Pengaruh *Stemming* pada *Word Intrusion Task*

Pernyataan Hipotesis Pengaruh *Stemming* pada *Word Intrusion task*

H_0: σ WIT tanpa *Stem* = σ WIT dengan *Stem*

H_1: σ WIT tanpa *Stem* ≠ σ WIT dengan *Stem*

Test and CI for Two Variances: WIT -Stem, WIT +Stem

Statistics

Variable	N	StDev	Variance	95% CI for StDevs
WIT -Stem	102	5.096	25.968	(4.681, 5.656)
WIT +Stem	102	5.962	35.543	(5.482, 6.610)

Ratio of standard deviations = 0.855
Ratio of variances = 0.731

Tests

Method	DF1	DF2	Test Statistic	P-Value
Bonett	1	—	5.31	0.021
Levene	1	202	1.47	0.227

Gambar 4.39 Uji *variance* WIT -*Stem* dan WIT +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.39 menunjukkan bahwa *p-value* dengan metode *Bonett* senilai 0.021 dan metode *Levene* senilai 0.221, yang mana nilai dari metode *Bonett* tersebut kurang dari nilai α, sementara nilai dari metode *Levene* lebih besar dari α, namun rataan kedua nilai tersebut lebih besar dari α, sehingga dapat dikatakan bahwa nilai *variance* dari *Word Intrusion task* tanpa *Stem* dan *Word Intrusion task* dengan *Stem* dinyatakan sama.

Uji *Means*

Two-Sample T-Test and CI: WIT -Stem, WIT +Stem

Two-sample T for WIT -Stem vs WIT +Stem

	N	Mean	StDev	SE Mean
WIT -Stem	102	12.28	5.10	0.50
WIT +Stem	102	13.26	5.96	0.59

Difference = μ (WIT -Stem) - μ (WIT +Stem)
Estimate for difference: -0.980
95% CI for difference: (-2.512, 0.551)
T-Test of difference = 0 (vs ≠): T-Value = -1.26 P-Value = 0.208 DF = 197

Gambar 4.40 Uji *Means* WIT -*Stem* dan WIT +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.40 menunjukkan bahwa *p-value* senilai 0.208, yang mana nilai tersebut lebih besar dari nilai α, nilai tersebut menunjukkan bahwa tidak terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol gagal tolak.

4.2.5.2.5. Uji Hipotesis Pengaruh *Stemming* pada *Topic Intrusion task* 1

Pernyataan Hipotesis Pengaruh *Stemming* pada *Topic Intrusion task* 1

H_0: σ TIT 1 tanpa *Stem* = σ TIT 1 dengan *Stem*

H_1: σ TIT 1 tanpa *Stem* ≠ σ TIT 1 dengan *Stem*

Uji *Variance*

Test and CI for Two Variances: TIT 1 -Stem, TIT 1 +Stem

Statistics

Variable	N	StDev	Variance	95% CI for StDevs
TIT 1 -Stem	102	2.679	7.178	(2.285, 3.203)
TIT 1 +Stem	102	4.012	16.095	(3.132, 5.240)

Ratio of standard deviations = 0.668
Ratio of variances = 0.446

Tests

Method	DF1	DF2	Test Statistic	P-Value
Bonett	1	—	4.64	0.031
Levene	1	202	2.36	0.126

Gambar 4.41 Uji *variance* TIT 1 -*Stem* dan TIT 1 +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.41 menunjukkan bahwa *p-value* dengan metode *Bonett* senilai 0.031 dan metode *Levene* senilai 0.126, yang mana nilai dari metode *Bonett* tersebut kurang dari nilai α, sementara nilai dari metode *Levene* lebih besar dari α, namun rataan kedua nilai tersebut lebih besar dari α, sehingga dapat dikatakan bahwa nilai *variance* dari *Topic Intrusion task* 1 tanpa *Stem* dan *Topic Intrusion task* 1 dengan *Stem* dinyatakan sama.

Uji *Means*

Two-Sample T-Test and CI: TIT 1 -Stem, TIT 1 +Stem

```
Two-sample T for TIT 1 -Stem vs TIT 1 +Stem

              N   Mean  StDev  SE Mean
TIT 1 -Stem  102  15.66   2.68     0.27
TIT 1 +Stem  102  16.30   4.01     0.40


Difference = μ (TIT 1 -Stem) - μ (TIT 1 +Stem)
Estimate for difference:  -0.647
95% CI for difference:  (-1.590, 0.296)
T-Test of difference = 0 (vs ≠): T-Value = -1.35  P-Value = 0.177  DF = 176
```

Gambar 4.42 Uji *Means* TIT 1 -*Stem* dan TIT 1 +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.42 menunjukkan bahwa *p-value* senilai 0.177, yang mana nilai tersebut lebih besar dari nilai α, nilai tersebut menunjukkan bahwa tidak terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol gagal tolak.

4.2.5.2.6. Uji Hipotesis Pengaruh *Stemming* pada *Topic Intrusion task* 2

Pernyataan Hipotesis Pengaruh *Stemming* pada *Topic Intrusion task* 1

H_0: σ TIT 1 tanpa *Stem* = σ TIT 1 dengan *Stem*

H_1: σ TIT 1 tanpa *Stem* ≠ σ TIT 1 dengan *Stem*

Uji *Variance*

Test and CI for Two Variances: TIT 2 - Stem, TIT 2 +Stem

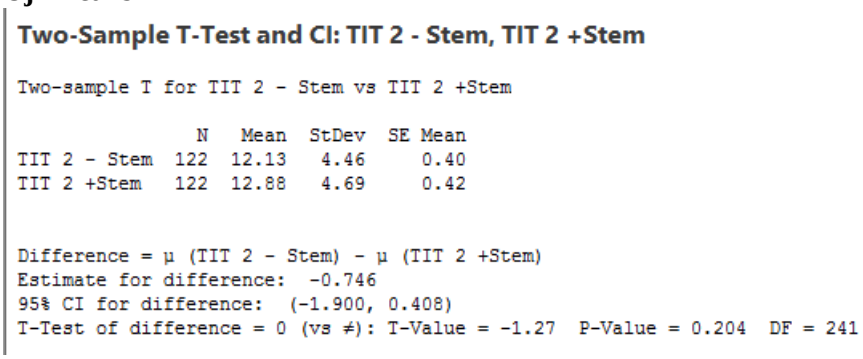
```
Statistics

                                      95% CI for
Variable        N  StDev  Variance      StDevs
TIT 2 - Stem  122  4.457    19.867  (3.981, 5.071)
TIT 2 +Stem   122  4.690    21.993  (4.202, 5.319)

Ratio of standard deviations = 0.950
Ratio of variances = 0.903


Tests

                     Test
Method  DF1  DF2  Statistic  P-Value
Bonett    1    —       0.36    0.549
Levene    1  242       0.64    0.424
```

Gambar 4.43 Uji *variance* TIT 2 -*Stem* dan TIT 2 +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.43 menunjukkan bahwa *p-*

value dengan metode *Bonett* senilai 0.549 dan metode *Levene* senilai 0.429, yang mana kedua nilai tersebut lebih besar dari nilai α, sehingga dapat dikatakan bahwa nilai *variance* dari *Topic Intrusion task* 2 tanpa *Stem* dan *Topic Intrusion task* 2 dengan *Stem* dinyatakan sama.

Uji Means



Two-Sample T-Test and CI: TIT 2 - Stem, TIT 2 +Stem

```
Two-sample T for TIT 2 - Stem vs TIT 2 +Stem

               N   Mean  StDev  SE Mean
TIT 2 - Stem  122  12.13   4.46     0.40
TIT 2 +Stem   122  12.88   4.69     0.42


Difference = μ (TIT 2 - Stem) - μ (TIT 2 +Stem)
Estimate for difference:  -0.746
95% CI for difference:  (-1.900, 0.408)
T-Test of difference = 0 (vs ≠): T-Value = -1.27  P-Value = 0.204  DF = 241
```

Gambar 4.44 Uji *Means* TIT 2 -*Stem* dan TIT 2 +*Stem*

Hasil pengujian yang ditampilkan pada Gambar 4.44 menunjukkan bahwa *p-value* senilai 0.204, yang mana nilai tersebut lebih besar dari nilai α, nilai tersebut menunjukkan bahwa tidak terdapat nilai *Means* yang signifikan antara kedua sampel, sehingga hipotesis nol

4.2.5.2.7. Kesimpulan Uji Hipotesis

Berdasarkan hasil uji hipotesis yang telah dilakukan, berikut merupakan rekapitulasi uji *variance* yang ditunjukkan pada Tabel 4.11 dan rekapitulasi uji *Means* yang ditunjukkan pada Tabel 4.12.

Tabel 4.11 Tabel Rekapitulasi Uji *Variance*

Uji *Variance* (a=0.05)			
Sampel	*p-value*		Kesimpulan
	Bonett	Lavene	
WIT-*Stem*&WIT+*Stem*	0.021	0.227	*Variance* sama
TIT 1-*Stem*&TIT 1+*Stem*	0.031	0.126	*Variance* sama
TIT 2-*Stem*&TIT 2+*Stem*	0.549	0.424	*Variance* sama
WIT &TIT1	0.000	0.000	*Variance* beda
WIT &TIT2	0.000	0.001	*Variance* beda
TIT1&TIT2	0.003	0.019	*Variance* beda

Tabel 4.12 Tabel Rekapitulasi Uji *Means*

Uji *Means*		
a=0.05	*p-value*	Kesimpulan
WIT-*Stem*vsWIT+*Stem*	0.208	*Means* sama, H_0 gagal tolak
TIT 1-*Stem* vsTIT1 +*Stem*	0.177	*Means* sama, H_0 gagal tolak
TIT 2-*Stem* vsTIT2 +*Stem*	0.204	*Means* sama, H_0 gagal tolak
WIT vsTIT 1	0.000	*Means* beda, H_0 ditolak
WIT vsTIT 2	0.000	*Means* beda, H_0 ditolak
TIT 1vsTIT2	0.669	*Means* sama, H_0 gagal tolak

Berdasarkan uji hipotesis yang dilakukan melalui uji *variance* dan uji *Means*, terdapat dua kesimpulan yang dapat diambil. Pertama, nilai rata-rata metode WIT lebih tinggi dan berbeda signifikan jika dibandingkan dengan nilai rata-rata metode TIT 1 dan TIT 2, hal ini menunjukkan bahwa metode *Word Intrusion task* secara inheren mampu diinterpretasi lebih baik oleh responden. Hal ini dibuktikan pula dengan nilai rata-rata metode TIT 2 yang tidak signifikan terhadap nilai rata-rata TIT 1 meskipun dalam pelaksanaan TIT 2 telah dilakukan perbaikan seperlunya dan dilakukan secara terpisah sebagai usaha untuk mengeliminasi asumsi penyebab dari skor pada TIT 1 yang lebih rendah dari WIT. Kedua, nilai rata-rata metode WIT, TIT 1, dan TIT 2 dengan *Stem* memberikan hasil yang lebih baik. Hal ini menunjukkan bahwa proses *Stemming* berpengaruh positif terhadap pemodelan yang dilakukan dalam kasus ini, namun berdasarkan uji *variance* dan uji *Means*, perbedaan ini tidak signifikan.

4.2.5.3. Uji ANOVA (Analysis of *Variance*)

Uji ANOVA bertujuan untuk untuk membuktikan hipotesis (dugaan sementara) berdasarkan metode statistika, biasanya digunakan untuk melihat signifikansi pada rata-ratadari dua atau lebih grup yang berbeda. Uji ANOVA dalam kasus ini bertujuan untuk menguji signifikansi rata-rata pada setiap topik, sehingga dapat diketahui topik mana

yang berbeda secara signifikan dari topik yang lain. Untuk mempertajam analisis, uji ANOVA dibagi berdasarkan metode uji koherensi topik dengan dibantu *tools* Minitab 17. Dalam melakukan uji ANOVA, tahap-tahap yang dilakukan adalah mendefinisikan pernyataan hipotesis, melakukan uji ANOVA kemudian menyusun kesimpulan berdasarkan kedua uji tersebut. Kesimpulan ANOVA dilakukan dengan membandingkan *p-value* dengan tingkat signifikansi. Tingkat signifikansi (α) merupakan probabilitas penolakan hipotesis nol ketika hipotesis tersebut benar. Tingkat signifikansi akan menjadi pembanding terhadap nilai *p-value* untuk menentukan apakah hipotesis nol diterima atau ditolak, hipotesis nol ditolak jika nilai *p-value* kurang dari nilai α. Pada kasus pengujian hipotesis ini dipilih tingkat signifikansi α = 0.05.

Sebelum melakukan uji anova, terlebih dahulu data kelompok topik dilakukan uji *variance*. Uji *variance* dilakukan untuk setiap kelompok topik pada setiap metode uji koherensi topik dengan hasil yang ditampilkan pada Tabel 4.13.

Tabel 4.13 Tabel Hasil Uji *Variance*

	Variancep-value		Kesimpulan
	Multiple Comparisons	Lavene	
WIT	0.644	0.518	*Variance* sama
TIT1	0.059	0.235	*Variance* sama
TIT2	0.000	0.004	*Variance* beda

Perbedaan varian ditunjukkan dengan melakukan perbandingan pada nilai *p-value* baik dengan metode multiple comparisons dan Lavene dengan nilai alpha (α=0.05), jika nilai p-value lebih tinggi dari nilai alpha, maka variance pada kedua kelompok uji koherensi topik dinyatakan berbeda signifikan.

Hasil uji *variance* menunjukkan *variance* pada topik topik dalam *Word Intrusion task* dan *Topic Intrusion task* 1 memiliki *variance* yang sama, sementara topik topik dalam *Topic Intrusion task* 2 memiliki *variance* berbeda. Topik-topik dalam uji koherensi

yang memiliki *variance* yang sama akan dilakukan uji ANOVA dengan metode Tukey, sementara topik-topik dalam uji koherensi yang memiliki *variance* yang berbeda akan dilakukan uji anova dengan metode Games-Howell. Perbedaan metode ini akan tetap menghasilkan luaran yang serupa. Metode Tukey maupun Games Howell akan mengelompokkan nilai rata-rata ke dalam kelompok huruf. Suatu topik akan dinyatakan memiliki rata-rata yang sama jika digolongkan pada kelompok huruf yang sama, demikian pula sebaliknya, rata-rata topik akan dinyatakan berbeda signifikan jika digolongkan pada kelompok huruf yang berbeda.

4.2.5.3.1. Uji ANOVA *Word Intrusion Task*

Pernyataan hipotesis uji ANOVA *Word Intrusion task*

H0: semua *Means* untuk setiap sampel sama

H1: terdapat setidaknya satu *Means* yang berbeda

Uji Anova

Berikut merupakan informasi pengelompokan dengan Metode Tukey dengan Confidence Level 95% (α = 0.05) yang ditunjukkan pada Gambar 4.45

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
WIT T1	102	8.559	A
WIT T0	102	8.039	A B
WIT T2	102	7.716	B
WIT T3	102	7.647	B

Means that do not share a letter are significantly different.

Gambar 4.45 Informasi pengelompokan sampel pada *Word Intrusion Task*

Untuk kasus uji ANOVA *Word Intrusion Task*, yang tergolong dalam kelompok huruf A adalah topik 1 dan topik 0, sementara yang tergolong dalam kelompok huruf B adalah topik 0, 2 dan topik 3. Topik 0 tergolong ke dalam kelompok huruf A dan B karena memiliki nilai *Means* yang sama dengan topik-topik yang terdapat dalam kedua

kelompok huruf.

Games-Howell Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
TIT 2 T1 - TIT 2 T0	0.574	0.324	(-0.265, 1.413)	1.77	0.290
TIT 2 T2 - TIT 2 T0	-1.467	0.272	(-2.171, -0.763)	-5.40	0.000
TIT 2 T3 - TIT 2 T0	0.164	0.315	(-0.646, 0.974)	0.52	0.954
TIT 2 T2 - TIT 2 T1	-2.041	0.293	(-2.799, -1.283)	-6.96	0.000
TIT 2 T3 - TIT 2 T1	-0.410	0.334	(-1.267, 0.447)	-1.23	0.609
TIT 2 T3 - TIT 2 T2	1.631	0.283	(0.898, 2.365)	5.75	0.000

Gambar 4.46 Perbandingan berpasangan setiap sampel untuk *Word Intrusion Task*

Gambar 4.46 menunjukkan perbandingan berpasangan dari masing masing topik dalam uji ANOVA *Word Intrusion Task.* Berdasarkan Gambar 4.46, perbedaan varian ditunjukkan dengan melakukan perbandingan pada nilai *p-value* dan nilai alpha (α=0.05), jika nilai *p-value* lebih tinggi dari nilai alpha, maka rata-rata kedua topik dinyatakan berbeda signifikan. Dengan demikian, perbandingan berpasangan untuk setiap topik telah menunjukkan kesesuaian dengan yang ditunjukkan dalam tabel informasi pengelompokan.

4.2.5.3.2. Uji ANOVA *Topic Intrusion task* 1

Pernyataan hipotesis uji ANOVA *Topic Intrusion task* 1

H0: semua *Means* untuk setiap sampel sama

H1: terdapat setidaknya satu *Means* yang berbeda

Uji Anova

Berikut merupakan informasi pengelompokan dengan Metode Tukey dengan Confidence Level 95% yang ditunjukkan pada Gambar 4.47.

Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

Factor	N	Mean	Grouping
TIT 1 T1	102	6.892	A
TIT 1 T0	102	6.382	A
TIT 1 T3	102	6.284	A
TIT 1 T2	102	5.990	A

Means that do not share a letter are significantly different.

Gambar 4.47 Informasi pengelompokan sampel pada *Topic Intrusion task* 1

Untuk kasus uji ANOVA *Topic Intrusion task* 1, seluruh topik digolongkan pada huruf A, yang artinya nilai rata-rata untuk setiap topik dinyatakan sama.

Gambar 4.48 menunjukkan perbandingan berpasangan dari masing masing topik dalam uji ANOVA *Topic Intrusion task* 1.

Tukey Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
TIT 1 T1 - TIT 1 T0	0.510	0.408	(-0.537, 1.557)	1.25	0.595
TIT 1 T2 - TIT 1 T0	-0.392	0.408	(-1.439, 0.655)	-0.96	0.771
TIT 1 T3 - TIT 1 T0	-0.098	0.408	(-1.145, 0.949)	-0.24	0.995
TIT 1 T2 - TIT 1 T1	-0.902	0.408	(-1.949, 0.145)	-2.21	0.120
TIT 1 T3 - TIT 1 T1	-0.608	0.408	(-1.655, 0.439)	-1.49	0.443
TIT 1 T3 - TIT 1 T2	0.294	0.408	(-0.753, 1.341)	0.72	0.889

Individual confidence level = 98.94%

Gambar 4.48 Perbandingan berpasangan setiap sampel untuk *Topic Intrusion task* 1

Berdasarkan Gambar 4.48, perbedaan varian ditunjukkan dengan melakukan perbandingan pada nilai *p-value* dan nilai alpha (α=0.05), jika nilai *p-value* lebih tinggi dari nilai alpha, maka rata-rata kedua topik dinyatakan berbeda signifikan. Dengan demikian, perbandingan berpasangan untuk setiap topik telah menunjukkan kesesuaian dengan yang ditunjukkan dalam tabel informasi pengelompokan.

4.2.5.3.3. Uji ANOVA *Topic Intrusion task* 2

Pernyataan hipotesis uji ANOVA *Topic Intrusion task* 2

H_0: semua *Means* untuk setiap sampel sama

H_1: terdapat setidaknya satu *Means* yang berbeda

Uji Anova

Berikut merupakan informasi pengelompokan dengan Metode Games-Howell dengan Confidence Level 95% yang ditunjukkan pada Gambar 4.49.

Games-Howell Pairwise Comparisons

Grouping Information Using the Games-Howell Method and 95% Confidence

Factor	N	Mean	Grouping
TIT 2 T1	122	6.828	A
TIT 2 T3	122	6.418	A
TIT 2 T0	122	6.254	A
TIT 2 T2	122	4.787	B

Means that do not share a letter are significantly different.

Gambar 4.49 Informasi pengelompokan sampel pada *Topic Intrusion task* 2

Untuk kasus uji ANOVA *Topic Intrusion task* 2, yang tergolong dalam kelompok huruf A adalah topik 1 dan topik 3, sementara yang tergolong dalam kelompok huruf B adalah topik 0, 2

Gambar 4.50 menunjukkan perbandingan berpasangan dari masing masing topik dalam uji ANOVA *Word Intrusion Task.*

Games-Howell Simultaneous Tests for Differences of Means

Difference of Levels	Difference of Means	SE of Difference	95% CI	T-Value	Adjusted P-Value
TIT 2 T1 - TIT 2 T0	0.574	0.324	(-0.265, 1.413)	1.77	0.290
TIT 2 T2 - TIT 2 T0	-1.467	0.272	(-2.171, -0.763)	-5.40	0.000
TIT 2 T3 - TIT 2 T0	0.164	0.315	(-0.646, 0.974)	0.52	0.954
TIT 2 T2 - TIT 2 T1	-2.041	0.293	(-2.799, -1.283)	-6.96	0.000
TIT 2 T3 - TIT 2 T1	-0.410	0.334	(-1.267, 0.447)	-1.23	0.609
TIT 2 T3 - TIT 2 T2	1.631	0.283	(0.898, 2.365)	5.75	0.000

Gambar 4.50 Perbandingan berpasangan setiap sampel untuk *Topic Intrusion task* 2

Berdasarkan Gambar 4.50, perbedaan varian ditunjukkan dengan melakukan perbandingan pada nilai *p-value* dan nilai alpha (α=0.05), jika nilai *p-value* lebih tinggi dari nilai alpha, maka rata-rata kedua topik dinyatakan berbeda signifikan. Dengan demikian, perbandingan berpasangan untuk setiap topik telah menunjukkan kesesuaian dengan yang ditunjukkan dalam tabel informasi pengelompokan.

4.2.5.3.4. Kesimpulan Uji ANOVA

Berdasarkan uji ANOVA yang telah dilakukan, analisis terhadap kualitas pemodelan topik hingga tingkat topik menghasilkan dua kesimpulan. Pertama, topik 1 merupakan topik yang paling mudah diinterpretasi oleh manusia, hal ini didukung dengan nilai rata-rata topik 1 yang paling tinggi baik untuk metode WIT, TIT 1 maupun TIT 2, sementara topik 2 merupakan topik yang paling sulit diinterpretasi oleh manusia, hal ini didukung dengan nilai rata-rata topik 2 yang paling rendah pada metode TIT 1 dan TIT 2, dan kedua terendah pada metode WIT. Kedua, dari keempat topik yang dihasilkan, dapat diketahui bahwa derajat pemahaman responden terhadap topik untuk metode WIT dan TIT 2 dapat dibagi ke dalam dua tingkat, sementara untuk metode TIT 1, hanya dalam satu tingkat. Dengan kata lain, topik-topik yang berada dalam kelompok A dapat dikatakan sebagai kelompok topik yang memiliki performa baik sebagai classifier,

sementara topik-topik yang berada dalam kelompok B dapat dikatakan sebagai kelompok topik yang memiliki performa kurang baik.

4.2.6 Pachinko Allocation Model (PAM)

Ketika menggunakan metode LDA biasa, metode ini dapat dengan jelas memilih topik yang baik dari bagian tersebut. Tetapi ketika melihat hubungan antar topik, itu tidak bisa membantu. Tapi PAM dapat menjaga hubungan antara beberapa topik.

Oleh karena itu dalam bagian di atas, kata-kata individu membentuk kosa kata, topik yang telah disebutkan mewakili sub topik. Kata-kata koheren seperti indeks harga saham gabungan dan saham mewakili topik super.

Grafik dibuat berdasarkan frekuensi, hubungan koheren antara kata-kata. Topik, subtopik diekstraksi oleh model ini sendiri.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, maka kesimpulan yang dapatdiambil dari penelitian ini adalah sebagai berikut.

1. Telah dilakukan berbagai eksperimen dalam proses *Topic modeling* dengan metode *LDA*. Eksperimen yang terkait jumlah *passes* menghasilkan 10 *passes* sebagai jumlah *passes* terbaik berdasarkan nilai *perplexity* yang telah stabil pada iterasi ke 10 untuk setiap distribusi topik yang terpilih secara acak.
2. Eksperimen yang terkait dengan jumlah topik dalam model dibagi menjadi dua eksperimen, yaitu *Topic modeling* dengan *Stem* dan *Topic modeling* tanpa *Stem*. Berdasarkan hasil eksperimen ini, dihasilkan beberapa kesimpulan. Kesimpulan pertama, bahwa 4 topik merupakan jumlah topik terbaik dalam membentuk topik model berdasarkan analisis nilai *perplexity* dan dapat dikatakan cukup stabil berdasarkan nilai standar deviasinya. Kesimpulan kedua, proses *Stemming* dalam kasus ini memberikan pengaruh positif terhadap pembentukan model yang ditunjukkan dengan penurunan nilai *perplexity* sebesar 64.51.
3. Eksperimen terkait klasifikasi dokumen ke dalam topik telah dilakukan baik untuk model dengan *Stem*, maupun model tanpa *Stem*. Hasil eksperimen klasifikasi dokumen menunjukkan hasil yang meyakinkan yang ditunjukkan dengan distribusi probabilitas dokumen terhadap topik untuk model dengan atau tanpa *Stem* secara mayoritas berada pada rentang 0.5 hingga 0.99.

4. Dalam mengukur tingkat koherensi topik yang menjadi luaran *Topic modeling*, metode yang digunakan adalah *Word Intrusion task* dan *Topic Intrusion Task*. Tahap-tahap yang dilakukan adalah merancang sistematika dan materi kuesioner, kemudian melakukan analisis hasil tanggapan responden dengan uji hipotesis dan uji ANOVA.

5. Berdasarkan uji hipotesis yang dilakukan melalui uji *variance* dan uji *Means*, terdapat dua kesimpulan yang dapat diambil. Pertama, metode *Word Intrusion task* secara inheren mampu diinterpretasi lebih baik oleh responden. Kedua, nilai rata-rata metode WIT, TIT 1, dan TIT 2 dengan *Stem* memberikan hasil yang lebih baik. Hal ini menunjukkan bahwa proses *Stemming* berpengaruh positif terhadap pemodelan yang dilakukan dalam kasus ini, namun berdasarkan uji *variance* dan uji *Means*, perbedaan ini tidak signifikan.

6. Berdasarkan uji ANOVA yang telah dilakukan, analisis terhadap kualitas pemodelan topik hingga tingkat topik menghasilkan dua kesimpulan. Pertama, topik 1 merupakan topik yang paling mudah diinterpretasi oleh manusia, sementara topik 2 merupakan topik yang paling sulit diinterpretasi oleh manusia. Kedua, dari keempat topik yang dihasilkan, dapat diketahui bahwa derajat pemahaman responden terhadap topik untuk metode WIT dan TIT 2 dapat dibagi ke dalam dua tingkat, sementara untuk metode TIT 1, hanya dalam satu tingkat.

7. PAM membuat grafik berdasarkan frekuensi, hubungan koheren antara kata-kata. Model PAM mengektraksi topik dan sub topik dari berita saham online.

5.2 Saran

Berdasarkan kesimpulan di atas, ada beberapa saran dapat digunakan untuk ekstraksi berita saham online, yaitu

1 Data yang digunakan dalam penelitian ini bersumber dari halaman web berita saham kontan.co.id yang menggunakan kebahasaan berita media online. Pada pengembangannya, dapat diujicobakan dataset dari media online lain yang terpercaya.

2. Dari hasil pemodelan topik, terdapat kata-kata yang tergolong sebagai nama perusahaan. Terdapat pula kata-kata yang berupa singkatan yang memiliki arti sama namun dalam pemodelan topik menjadi dua entitas yang berbeda, seperti ihsg dengan indeks harga saham gabungan dan bursa efek Indonesia dengan bei. Untuk itu diperlukan suatu normalisasi sebelum dilakukan pemodelan topik.

DAFTAR PUSTAKA

A. Gaur. (2015). "Topic Models as A Novel Approach To Identify Themes In Content Analysis: The Example of Organizational Research Methods".

Avellaneda, M., Stoikov, S. (2008). High-frequency trading in a limit order book. Quant. Financ. **8**(3), 217–224.

Balu, Raghavendran. (2013). What are the advantages and disadvantages of Latent Semantic Analysis? Diakses dari https://www.quora.com/What-are-the-advantages-and-disadvantages-of-Latent-Semantic-Analysis tanggal 1 Desember 2020.

Bansal, Harsh. (2020). Latent Dirichlet Allocation. Diakses dari https://iq.opengenus.org/latent-dirichlet-allocation tanggal 12 November 2020.

Banu, S. Halima, & S. Chitrakala. (2016). Trending Topic Analysis Using Novel Sub Topic Detection Model. *2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*. IEEE. pp. 157-161.

Blei, David M. (2003). "Latent Dirichlet Allocation," *Machine Learning Research 3*, pp. 933-1022.

Blei, David M. (2012). Probabilistic Topic Models. *Communications of the ACM*, Volume 55, Issue 4. pp. 77-84.

Coussement, K. and D Van den Poel. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two

parameter-selection techniques. Expert Systems with Applications. Volume 34, Issue 1, January 2008, Pages 313-327.

Dadgar, Seyyed Mohammad Hossein, *et al*. (2016). A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification. *IEEE International Conference on Engineering and Technology (ICETECH)*. IEEE. pp.112-116.

David Newman. (2010). "Automatic Evaluation of Topic Coherence," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, p. 100–108.

F. Z. Tala. (2003). "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia." *Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands*.

H. Z. Zhou Tong. (2016). "A Text Mining Research Based on LDA Topic Modelling," *Jodrey School of Computer Science, Acadia University, Wolfville, NS, Canada*, vol. 10.5121/csit.2016.60616, p. 201–210.

J. C. Campbell, A. Hindle and A. E. Stroulia. (2014). "Latent Dirichlet Allocation: Extracting Topics".

Jey Han Lau, Nigel Collier, Timothy Baldwin. (2012). "On-line Trend Analysis with Topic Models: #twitter trends detection topic model online," *Proceedings of COLING 2012: Technical Papers*, p. 1519–1534.

J. F. Yeh, Y. S. Tan and C. H. Lee. (2016). "Topic detection and tracking for conversational content by using conceptual dynamic Latent Dirichlet Allocation," *Neurocomputing*.

Jonathan Chang. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models," in *Neural Information Processing Systems*, Vancouver, BC.

Keith Stevens. (2012). "Exploring Topic Coherence over many models and many topics," *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 952–961.

Li, X., Xie, H., Chen, L., Wang, J., Deng, X. (2014). News impact on stock price return via sentiment analysis. Knowl.-Based Syst. 69, 14–23.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. ICML.

Manthiramoorthi, Murugesh. (2020). Pachinko Allocation Model. Diakses dari https://iq.opengenus.org/pachinko-allocation-model tanggal 17 Februari 2020.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 10-pp. IEEE.

M. Muslich. (2010). Garis-Garis Besar Tata Bahasa Baku Bahasa Indonesia. Bandung: Refika Aditama.

M. R. Brett, "Journal of Digital Humanities," Desember 2012. [Online]. Available: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Putra, I Made Kusnanta Bramantya. (2017). Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan *Latent Dirichlet Allocation (LDA)*. Tugas Akhir. Jurusan Sistem Informasi. Fakultas Teknologi Informasi. Institut Teknologi Sepuluh Nopember, Surabaya.

Radim Rehurek dan Petr Sojka. (2010). "Software Framework for Topic Modelling with Large Corpora". *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks*, pp. 46-50, at Malta.

Setijohatmo, Urip T., d.k.k. (2020). Analisis Metoda *Latent Dirichlet Allocation* untuk Klasifikasi Dokumen Laporan Tugas Akhir Berdasarkan Pemodelan Topik. *Prosiding The 11th Industrial Research Workshop and National Seminar Bandung, 26-27 Agustus 2020*.

Vapnik, V. (2005). Universal Learning Technology: Support Vector Machines. *NEC Journal of Advanced Technology*, Vol. 2, No. 2, pp. 137-144.

JUSTIFIKASI ANGGARAN PENELITIAN

Biaya penelitian yang dianggarkan dapat dilihat dalam tabel berikut.

NOMER	TANGGAL	URAIAN	JUMLAH
1	14/01/2021	Persiapan Proposal	272.000
2	02/02/2021	Foto copy & jilid proposal	28.000
3	01/04/2021	HP Cartridge 680 Black	145.000
4	01/04/2021	ATK	137.000
5	08/04/2021	Data kontan.co.id 1	480.000
6	09/04/2021	Data kontan.co.id 2	480.000
7	12/04/2021	Data kontan.co.id 3	480.000
8	13/04/2021	Data kontan.co.id 4	480.000
9	17/06/2021	Kabel USB OTG Micro	10.000
10	17/06/2021	CD Casing Bulat	19.000
11	21/06/2021	Analisis penelitian	386.500
12	23/06/2021	Pelaksana Penelitian	300.000
13	25/06/2021	Desain penelitian	300.000
14	28/06/2021	Pelaksana Penelitian	300.000
15	05/07/2021	Data kontan.co.id 5	480.000
16	20/07/2021	Implementasi penelitian	300.000
17	21/07/2021	Scan gambar	28.500
18	26/07/2021	Pelaksana Penelitian	300.000
19	30/07/2021	Uji coba dan integrasi penelitian	237.500
20	30/07/2021	Pelaksana Penelitian	300.000
21	30/07/2021	Data kontan.co.id 6	480.000
22	19/08/2021	HP Cartridge 680 Colour	145.000
23	19/08/2021	ATK	43.200
24	21/08/2021	Laporan Kemajuan Penelitian	443.800
25	21/08/2021	Foto copy & jilid laporan kemajuan	74.500
JUMLAH TAHAP I (70%)			**6.650.000**
26	23/08/2021	Pembuatan Jurnal Penelitian	600.000
27	24/08/2021	Publikasi Jurnal Penelitian (Cetak)	700.000
28	24/08/2021	Publikasi Jurnal Penelitian (Online)	700.000
29	24/08/2021	Seminar nasional penelitian	780.000
30	25/08/2021	Foto copy dan Jilid Laporan Hasil Penelitian	70.000
JUMLAH TAHAP II (30%)			**2.850.000**
JUMLAH KESELURUHAN (TAHAP I + TAHAP II)			**9.500.000**

Surabaya, 25 Agustus 2021

Ketua Peneliti

Ekka Pujo A. A., S.E., M. Kom.

JADWAL PELAKSANAAN

Jadwal kegiatan penelitian meliputi kegiatan persiapan, pelaksanaan, dan penyusunan laporan penelitian dalam bentuk bar-chart.

Kegiatan	Th. 2020/2021											
	1	2	3	4	5	6	7	8	9	10	11	12
Persiapan	■											
Analisis kebutuhan	■											
Pembuatan proposal		■										
Seminar proposal			■									
Pelaksanaan penelitian				■								
- Analisis				■								
- Perencanaan					■							
- Pengembangan						■						
- Uji Coba							■					
- Penilaian							■					
Pembuatan laporan					■	■	■					
Seminar hasil penelitian								■				
Penyerahan hasil penelitian								■				

PERSONALIA PENELITIAN

Personalia yang terlibat dalam penelitian ini umumnya terdiri dari:

1. Ketua Peneliti
   - a. Nama lengkap : Ekka Pujo Ariesanto Akhmad, S.E., M.Kom.
   - b. Jenis kelamin : Laki-laki
   - c. NIDN : 0724037402
   - d. Disiplin ilmu : Manajemen/Teknologi Informasi
   - e. Pangkat/Golongan : III/d
   - f. Jabatan fungsional : Lektor
   - g. Fakultas/Jurusan : Program Diploma Pelayaran/KPN
   - h. Waktu penelitian : 6 bulan

2. Anggota Peneliti
   - a. Nama lengkap : Carlos L. Prawirosastro, S.Pd.I., M.Pd.I.
   - b. Jenis kelamin : Laki-laki
   - c. NIDN : 0710078302
   - d. Disiplin ilmu : Pendidikan Agama Islam
   - e. Pangkat/Golongan : III/b
   - f. Jabatan fungsional : -
   - g. Fakultas/Jurusan : Program Diploma Pelayaran/KPN
   - h. Waktu penelitian : 6 bulan

3. Pekerja Lapangan/Pencacah : Samuel
   Arif
   M. Rayhan
   Ghusma
   M. Habibi
   Kukuh

SURAT KETERANGAN *REVIEW*
HASIL PENELITIAN

Bersama ini kami beritahukan bahwa hasil penelitian dosen berikut:

Nama : Ekka Pujo Ariesanto Akhmad, S.E., M.Kom.

NIK/NIDN : 01197/0724037402

Jabatan Fungsional : Lektor

Fakultas/Jurusan : Program Diploma Pelayaran/KPN

Anggota : Carlos L. Prawirosastro, S.Pd.I., M.Pd.I.

Judul Penelitian : Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online

(1 *) SUDAH diperbaiki sesuai masukan reviewer.

2 *) PERBAIKAN BELUM sesuai dengan reviewer. Oleh karena itu kami sarankan:

.....

.....

.....

Surabaya, 25 Agustus 2021

Reviewer,

Nurul Rosana, S.Pi., M.T.
NIP. 01137

Catatan: *) Lingkari salah satu

SURAT KETERANGAN *REVIEW*
HASIL PENELITIAN

Bersama ini kami beritahukan bahwa hasil penelitian dosen berikut:

Nama	: Ekka Pujo Ariesanto Akhmad, S.E., M.Kom.
NIK/NIDN	: 01197/0724037402
Jabatan Fungsional	: Lektor
Fakultas/Jurusan	: Program Diploma Pelayaran/KPN
Anggota	: Carlos L. Prawirosastro, S.Pd.I., M.Pd.I.
Judul Penelitian	: Pemodelan Topik Menggunakan Latent Dirichlet Allocation dan Pachinko Allocation Model Untuk Ekstraksi Berita Saham Online

(1 *) SUDAH diperbaiki sesuai masukan reviewer.

2 *) PERBAIKAN BELUM sesuai dengan reviewer. Oleh karena itu kami sarankan:

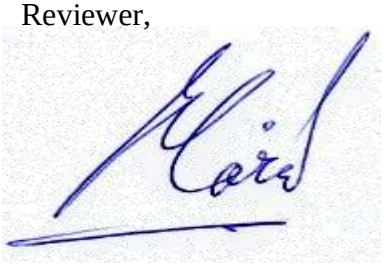
.....

.....

.....

Surabaya, 25 Agustus 2021

Reviewer,



Muhammad Taufiqurrohman, S.T., M.T.
NIK. 01235

Catatan: *) Lingkari salah satu