

**EKSTRAKSI ASPEK EKSPLISIT PADA KOMENTAR PRODUK DI
TOKO ELEKTRONIK**

THESIS

**Oleh:
ALVIAN BURHANUDDIN
NIM. 19841009**



**PROGRAM MAGISTER TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2023**

**EKSTRAKSI ASPEK EKSPLISIT PADA KOMENTAR PRODUK DI
TOKO ELEKTRONIK**

THESIS

Diajukan kepada:

**Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Magister Komputer (M.Kom)**

Oleh:

**ALVIAN BURHANUDDIN
NIM. 19841009**

**PROGRAM MAGISTER TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2023**

TESIS

**EKSTRAKSI ASPEK EKSPLISIT PADA KOMENTAR PRODUK
DI TOKO ELEKTRONIK**

**OLEH
ALVIAN BURHANUDDIN
19841009**

**PEMBIMBING
Prof. Dr. Suhartono, M.Kom
196805192003121001**



**PROGRAM MAGISTER TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
MALANG
2023**


**EKSTRAKSI ASPEK EKSPLISIT PADA KOMENTAR PRODUK DI
TOKO ELEKTRONIK**

THESIS

Oleh :
ALVIAN BURHANUDDIN
NIM. 19841009

Telah diperiksa dan Disetujui untuk Diuji
Tanggal:

Pembimbing 1



Prof. Dr. Suhartono, M.Kom
NIP. 196805192003121001

Pembimbing 2


Dr. Sri Harini, M.Si
NIP. 197310142001122002

Mengetahui,
Ketua Program Studi Magister Informatika
Fakultas Sains dan Teknologi
Universitas Sekejur Maulana Malik Ibrahim Malang




W. Cryslian
197404242009011008

**EKSTRAKSI ASPEK EKSPLISIT PADA KOMENTAR PRODUK DI
TOKO ELEKTRONIK**

THESIS

Oleh :
ALVIAN BURHANUDDIN
NIM. 19841009

Telah Dipertahankan di Depan Dewan Penguji Tesis
dan Dinyatakan Diterima sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Magister
Tanggal: 18 September 2023

Susunan dewan Penguji		Tanda tangan
Penguji Utama	: <u>Dr. Usman Pagalay, M.Si</u> NIP.196504142003121001	(.....)
Ketua Penguji	: <u>Dr. Muhammad Faisal, M.T</u> NIP.197405102005011007	(.....)
Sekretaris Penguji	: <u>Prof. Dr. Suhartono, M.Kom</u> NIP.196805192003121001	(.....)
Anggota Penguji	: <u>Dr. Sri Harini, M.Si</u> NIP.197310142001122002	(.....)

Mengetahui dan Mengesahkan
Ketua Program Studi Magister Informatika
Fakultas Sains Dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Dr. Crysdiyan
NIP. 197404242009011008

HALAMAN PERSEMBAHAN

This page left blank

HALAMAN PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini :

Nama : Alvian Burhanuddin
NIM : 19841009
Tahun terdaftar : 2019
Program Studi : Magister Informatika
Fakultas : Sains dan teknologi

Menyatakan bahwa dalam dokumen ilmiah Tesis ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan sumbernya secara lengkap dalam daftar pustaka.

Dengan demikian saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi. Apabila dokumen ilmiah Tesis ini di kemudian hari terbukti merupakan plagiasi dari hasil karya penulis lain dan/atau dengan sengaja mengajukan karya atau pendapat yang merupakan hasil karya penulis lain, maka penulis bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Malang, 18 September 2023

Yang membuat pernyataan,

A 10,000 Rupiah postage stamp is placed over the signature. The stamp features the Garuda Pancasila emblem and the text '10000', 'METESAI TEMPEL', and '93AKX655796817'.

Alvian Burhanuddin

NIM. 19841009

PRAKATA

Segala Puji syukur terucap ke hadirat Allah SWT yang telah melimpahkan rahmat dan barokah-Nya. Tak lupa shalawat serta salam untuk junjungan kami nabi Muhammad SAW. Dengan barakah dan syafaat beliau, Penulis dapat menyelesaikan keseluruhan penulisan tesis dengan judul "EKSTRAKSI ASPEK EKSPRESIF PADA KOMENTAR PRODUK DI TOKO ELEKTRONIK". Laporan tesis ini disusun untuk memenuhi salah satu syarat dalam memperoleh gelar Magister komputer (M.Kom) pada Program Studi S2 Magister Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Dalam melakukan penelitian dan penyusunan laporan tesis ini penulis telah mendapatkan banyak dukungan dan bantuan dari berbagai pihak. Penulis mengucapkan terima kasih yang tak terhingga kepada:

1. Seluruh jajaran dosen Magister informatika, Khususnya Bapak Prof. Dr. Suhartono, M.Kom dan Ibu Prof. Dr. Sri Harini, M.Si yang telah memberikan pengarahan, pengetahuan, serta pandangan terkait topik penelitian dalam tesis ini.
2. Keluarga penulis, khususnya istri dan orang tua penulis yang telah memberikan dukungan baik materiil maupun moril.
3. Seluruh pihak yang turut memberikan kontribusi pada pandangan penulis dalam penyelesaian penulisan tesis, yang tidak mungkin disebutkan satu persatu.

Penulis menyadari sepenuhnya bahwa tesis ini masih jauh dari sempurna. Maka dari itu, penulis sangat terbuka perihal saran, kritik dan masukan lain yang bersifat membangun. Akhir kata, semoga tulisan ini dapat memberikan manfaat dan memberikan tambahan khazanah keilmuan.

Malang, 18 September 2023

Alvian Burhanuddin

DAFTAR ISI

HALAMAN PENGESAHAN	ii
HALAMAN PENGESAHAN	iv
HALAMAN PERSEMBAHAN	iv
PRAKATA	v
DAFTAR ISI	vii
DAFTAR TABEL	viii
DAFTAR GAMBAR	ix
ABSTRACT	x
ABSTRAK	xi
خلاصة	xii
BAB I Pendahuluan	1
1.1 Latar Belakang Masalah	1
1.2 Pertanyaan Masalah	4
1.3 Tujuan Penelitian	4
1.4 Manfaat Penelitian	4
1.5 Batasan Masalah	5
1.6 Sistematika Penulisan	5
BAB II Tinjauan Literatur	7
2.1 <i>Feature Extraction</i>	7
2.1.1 <i>Rule Based Feature Extraction</i>	7
2.1.2 TF-IDF	8
2.1.2.1 <i>Term Frequency</i>	9
2.1.2.2 <i>Inverse Document Frequency</i>	9
2.2 <i>Similarity Measurement</i>	9
2.2.1 NGD	10
2.2.2 Dictionary Based	11
2.3 Kerangka Teori	12
2.4 <i>Aspect Extraction</i> Perspektif Islam	18
BAB III Metode Penelitian	20
3.1 Prosedur Penelitian	20
3.2 Pengumpulan Data	20
3.3 Tahapan Penelitian	22
3.3.1 Preprocessing	23
3.3.2 Feature extraction	24
3.3.2.1 TF-IDF	24
3.3.2.2 Rule based Extraction	26
3.3.3 Similarity Measurement	27
3.3.3.1 NGD	27
3.3.3.2 Dictionary Based	28
3.3.4 Clustering	29
3.3.5 Quality Measurement	30
BAB IV Clustering Similarity Measurement Dictionary Based	32

4.1	Desain Sistem	32
4.2	<i>Feature extraction</i>	33
	4.2.1 <i>TF-IDF</i>	33
	4.2.2 <i>Rule based</i>	34
4.3	Similarity measurement	35
4.4	Quality Measurement	36
BAB V	Clustering Similarity Measurement NGD	39
5.1	Desain Sistem	39
5.2	<i>Feature extraction</i>	40
	5.2.1 <i>TF-IDF</i>	40
	5.2.2 <i>Rule based</i>	41
5.3	Similarity measurement	42
5.4	Quality Measurement	44
BAB VI	Pembahasan	47
6.1	Pembahasan	47
BAB VII	Kesimpulan	50
7.1	Kesimpulan	50
7.2	Saran	51
DAFTAR PUSTAKA	55

DAFTAR GAMBAR

Gambar 2.1	Kerangka teori.....	12
Gambar 2.2	Ilustrasi Alquran surat An Nisa Ayat 29	19
Gambar 3.1	Diagram Prosedur Penelitian.....	20
Gambar 3.2	Tahap cleansing document.....	23
Gambar 3.3	Tahap stopword removal document	23
Gambar 3.4	Tahap stemming document	24
Gambar 3.5	Tahap Normalisasi document.....	24
Gambar 3.6	Ilustrasi POS Tagger	26
Gambar 3.7	Ilustrasi POS Tagger	29
Gambar 4.1	Desain sistem cluster NGD	32
Gambar 4.2	Python similarity measurement using WordNet bahasa .	35
Gambar 4.3	Similarity Matrix	36
Gambar 4.4	Kalkulasi Silhouette score TF-IDF Dictionary Based ...	37
Gambar 4.5	Kalkulasi Silhouette score Rule Dictionary Based	38
Gambar 5.1	Desain sistem cluster Dictionary Based.....	39
Gambar 5.2	Python similarity measurement using WordNet bahasa .	43
Gambar 5.3	Kalkulasi Silhouette score TF-IDF Menggunakan NGD	45
Gambar 5.4	Kalkulasi Silhouette score Rule NGD Based	46

DAFTAR TABEL

Tabel 2.1	Penelitian Sebelumnya	13
Tabel 3.1	Contoh dataset	21
Tabel 3.2	Similarity measurement	28
Tabel 4.1	Perhitungan <i>Term Frequency</i>	33
Tabel 4.2	Perhitungan <i>Inverse Document Frequency</i>	34
Tabel 4.3	Perhitungan <i>Inverse Document Frequency</i>	34
Tabel 4.4	Perhitungan <i>Silhouette score</i>	37
Tabel 4.5	Perhitungan <i>Silhouette score</i>	38
Tabel 5.1	Perhitungan <i>Term Frequency</i>	40
Tabel 5.2	Perhitungan <i>Inverse Document Frequency</i>	41
Tabel 5.3	Perhitungan <i>Inverse Document Frequency</i>	41
Tabel 5.4	Perhitungan <i>Silhouette score</i>	44
Tabel 5.5	Perhitungan <i>Silhouette score</i>	45
Tabel 5.6	Perhitungan <i>Silhouette score</i>	46
Tabel 6.1	Hasil Sillhouette score <i>Dictionary based</i>	47
Tabel 6.2	Hasil Sillhouette score <i>NGD</i>	48
Tabel 6.3	Perbandingan antara Sillhouette score <i>Dictionary based</i> dan <i>NGD</i>	49

ABSTRACT

Burhanuddin, Alvian. 2023. **Extraction of Explicit Aspects of Product Comments at Electronic Stores**, Magister Program in Informatics, Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Advisors: (I) Prof. Dr. Suhartono, M.Kom (II) Prof. Dr. Sri Harini, M.Si .

Keywords :NGD, Dictionary Based, Similarity Measurement.

Comments on products in online stores are one of the assessment items used by customers and manufacturers. The inclusion of discernible components within every customer comment can aid in comprehending their specific preferences and expectations regarding products. In this thesis, we analyze the relationship between one aspect of words and other words in Indonesian language comments on products in online stores using a similarity measurement algorithm. In this research, the similarity measurement algorithm used is Dictionary Based and Neutralized Google Distance (NGD). The implementation of Silhouette analysis, along with feature extraction utilizing TF-IDF and similarity measurement with Dictionary Based, resulted in achieving the highest scores. Feature extraction with TF-IDF has better results compared to rule-based. The silhouette score reaches 0.71 compared to Rule which only reaches 0.5.

ABSTRAK

Burhanuddin, Alvian. 2023. **Ekstraksi Aspek Eksplisit Pada Komentar Produk Di Toko Elektronik**, program magister informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Prof. Dr. Sri Harini, M.Si .

Kata kunci – NGD, Dictionary Based, Similarity Measurement.

Komentar dari produk di toko online, merupakan salah satu hal penilaian yang digunakan oleh pelanggan maupun produsen. Adanya aspek secara eksplisit pada tiap komentar tersebut, dianggap mampu memberikan pemahaman secara spesifik terhadap kriteria barang berdasarkan penggunaan dari pelanggan. Pada tesis ini, kami melakukan analisa keterkaitan antara satu aspek kata dengan kata lainnya dalam komentar berbahasa indonesia pada produk di toko online menggunakan algoritma similarity measurement. Dalam penelitian ini algoritma *similarity measurement* yang digunakan adalah *Dictionary Based* dan *Neutralized Google Distance* (NGD). Saat analisa dilakukan menggunakan *Silhouette analysis*, Hasil dari Ekstraksi fitur dengan TF-IDF dan *similarity measurement* dengan *Dictionary Based* memiliki skor tertinggi. Ekstraksi fitur dengan TF-IDF memiliki hasil yang lebih baik dibandingkan dengan *rule based*. Dimana nilai silhouette score mencapai 0,71 dibandingkan dengan Rule based yang hanya mencapai 0,5.

2023. Burhanuddin, Alvian. استخراج الجوانب الصريحة لتعليقات المنتج في المتاجر الإلكترونية , ماجستير في برنامج المعلوماتية في جامعة مولانا الإسلامية التابعة للدولة الإسلامية مالك إبراهيم مالانج .
(I) Prof. Dr. Suhartono, M.Kom (II) Prof. Dr. Sri Harini, . مستشار .
M.Si

استخراج الجوانب الصريحة لتعليقات المنتج في المتاجر الإلكترونية

الكلمات الدالة – NGD , القاموس ، قياس التشابه.

التعليقات على المنتجات في المتاجر عبر الإنترنت هي أحد التقييمات المستخدمة من قبل العملاء والمصنعين. يعتبر وجود جوانب صريحة في كل من هذه التعليقات قادرًا على توفير فهم محدد لمعايير البضائع بناءً على استخدام العميل. في هذه الأطروحة ، نقوم بتحليل العلاقة بين جانب واحد من الكلمة والآخر في التعليقات باللغة الإندونيسية على المنتجات في المتاجر عبر الإنترنت باستخدام خوارزمية قياس التشابه. في هذا البحث ، خوارزمية قياس التشابه المستخدمة هي القاموس القائم على القاموس ومحايدة (NGD) Google Distance . عندما تم إجراء التحليل باستخدام تحليل Silhouette ، حصلت النتائج من Dictionary Based على أعلى درجة.

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Dengan semakin berkembangnya internet, masyarakat makin mudah mendapatkan kebutuhan. Misalkan saja dalam jual beli barang atau berupa jasa, ada banyak sekali pilihan toko daring yang bisa digunakan. Beragamnya produk membuat pengguna harus lebih pintar dalam memilih dan memilah barang. Untuk mempermudah pembeli, beberapa ecommerce memberikan penilaian rating pada sebuah barang. Makin sedikit rating pada barang tersebut, maka kualitasnya dianggap lebih rendah dibandingkan dengan barang dengan rating yang lebih tinggi. Pada kenyataannya, fitur review tersebut menimbulkan polemik baru. Salah satunya adalah munculnya ketidak konsekuensi antara sebuah nilai review dengan skala dibandingkan dengan komentar pada suatu barang (Saumya et al., 2021). Mengatasi permasalahan tersebut beberapa penelitian dilakukan diantaranya adalah membandingkan antara *word of mouth* dengan nilai rating dari user. Penelitian terhadap *word of mouth* dilakukan karena hal tersebut termasuk salah satu hal yang berpengaruh dalam proses keputusan seseorang dalam membeli barang (Hidayanto et al., 2017) (Asgari et al., 2022).

Sebuah opini atau review pada barang memiliki karakteristik yang berbeda. Misalkan saja dalam panjang review dan jumlah review. pada halaman pedagang besar atau produk populer tertentu, sudah menjadi yang lumrah adanya banyak sekali ulasan hingga mencapai ribuan. Dalam banyaknya review tersebut, terdapat bermacam ukuran review. Beberapa memiliki kalimat yang panjang dengan keterangan jelas, sebagian lain hanya mencakup beberapa kata saja. "barangnya ok", "jelek" dengan tanpa adanya penjelasan lebih lanjut pada apa yang dimaksud. Adanya kasus tersebut dapat mempersulit calon konsumen untuk menyimpulkan kualitas barang berdasarkan review, sehingga akan berpengaruh juga terhadap kegiatan pembelian barang tersebut.

Adapun jika melihat pada sisi manufaktur atau produsen barang, banyaknya ulasan juga meningkatkan kerumitan untuk memahami apa yang harus diperbaiki pada produk yang mereka miliki. Utamanya dengan kondisi

saat ini, dimana sebuah produk dapat dijual oleh banyak penjual yang berbeda beda tempat dan toko online. Review dari produk atau *electronic word of mouth* merupakan bagian dari sumber data untuk mengukur tingkat kepuasan konsumen terhadap barang. Sumber data tersebut bisa diolah sehingga dapat menjadi penunjang keputusan penting dalam perusahaan. Sentimen analisis merupakan salah satu metode pengolahan data yang menggunakan sumber data *electronic word of mouth*, misalkan saja pada dalam pengukuran kepuasan pelanggan pada hotel (Parolin & Boeing, 2019) dan pada kasus tempat rekreasi (Gerdt et al., 2019). Algoritma sentimen analisis diklasifikasi menjadi tiga bagian, yakni; (1) *document level*, (2) *sentence level*, (3) *aspect-based level*. Pada *aspect-based level* menawarkan hasil yang lebih baik dibandingkan dengan 2 metode lainnya (Rana & Cheah, 2016). Hal ini dikarenakan *aspect-based level* mampu menangkap beberapa kategori dalam satu kalimat komentar atau review, mengkategorikannya serta mampu menentukan polaritas sebuah komentar pada toko online.

Beberapa pendekatan untuk memudahkan ekstraksi informasi pada review telah dilakukan, salah satunya adalah adanya peringkat (*rating*). Peringkat merupakan sebuah cara pemberian nilai pada tiap review yang diberikan oleh pembeli barang. Nilai yang diberikan oleh pembeli merupakan nilai subjektif berdasarkan pengalaman dari tiap pembeli. Dari kebebasan tersebut, review terkadang tidak sesuai dengan apa yang dikeluhkan pada suatu barang. Hal ini dikuatkan pada penelitian sebelumnya (Hazarika et al., 2021), pada komparasi perbandingan rating aplikasi pada iTunes. yang mana terdapat ketidaksesuaian antara review yang diberikan oleh pengguna pada aplikasi. Sehingga, dapat ditemukan bahwa hal ini akan mengurangi nilai kepercayaan pelanggan untuk membeli pada suatu barang. Sehingga, ketika rating tersebut tidak dapat dipercayai, akan meningkatkan keinginan konsumen untuk memahami review atau komentar pengguna lain pada sebuah produk tersebut (Lee et al., 2021).

Didasari oleh masalah yang telah disebutkan tersebut, Ekstraksi aspek eksplisit pada opini produk elektronik di toko online merupakan salah satu tantangan penelitian yang bisa diselesaikan lewat metode komputerisasi. Dalam penelitian ini, kami menggunakan pendekatan *data mining* untuk melakukan ekstraksi aspek pada fitur kata dari review pelanggan. Secara garis besar, dalam penelitian ini kami akan melakukan proses mencakup:

(1) Identifikasi fitur kata yang berhubungan dengan barang pada review pembeli; Tahapan ini peneliti melakukan ekstraksi fitur kata yang sesuai dengan produk berdasarkan dengan aturan sintaksis dan TF-IDF.

(2) Melakukan pengukuran pada kesamaan antar kata menggunakan metode NGD dan *Dictionary Based*

(3) melakukan pengelompokan kata berdasarkan kedekatan kata yang ditentukan oleh metode NGD dan *Dictionary Based*;

Penelitian ekstraksi aspek eksplisit memiliki lebih banyak penelitian dibandingkan dengan aspek implisit (Rana & Cheah, 2016). Penelitian ekstraksi aspek eksplisit ini setidaknya memiliki 3 kelas utama, yakni *unsupervised*, *supervised*, dan *semi-supervised*. Dibanding dengan *supervised* dan *semi-supervised*, *unsupervised* ditemukan oleh rana (Rana & Cheah, 2016) lebih banyak digunakan. Dalam *unsupervised*, terdapat beberapa pendekatan yang bisa dilakukan untuk mendapatkan aspek dari kalimat. Pendekatan tersebut antara lain adalah *Frequency-based*(Marrese-Taylor et al., 2014)(Eirinaki et al., 2012), *Bootstrapping*(Liu et al., 2014)(Bagheri et al., 2013), dan *rule-based*(Bancken et al., 2014)(Poria et al., 2014).

Berdasarkan pada penelitian sebelumnya oleh bancken(Bancken et al., 2014), metode *rule-based* memiliki fleksibilitas dalam penggunaannya. metode *rule-based* ini menggunakan *syntactic dependency* dari tiap tiap bahasa. Sehingga, metode ini merupakan salah satu metode yang cukup relevan digunakan untuk bahasa indonesia. Sedikitnya corpus bahasa indonesia dibandingkan dengan bahasa inggris, membuat metode ini lebih kuat dibandingkan metode lain. Selaras dengan hal tersebut, metode TF-IDF juga merupakan salah satu metode pembandingan yang cukup relevan dengan *Rule based*, Misalkan saja pada penelitian milik (Fransiska et al., 2020), TF-IDF mampu memberikan hasil yang baik dalam melakukan ekstraksi fitur pada sebuah komentar di Google Play.

Berangkat dari latar belakang yang telah disebutkan, peneliti bermaksud melakukan penelitian dengan melakukan ekstraksi aspek eksplisit pada review produk di toko online dengan menggunakan rule based. Manfaat dari penelitian ini secara umum dapat membantu konsumen dalam memutuskan akan membeli suatu produk atau tidak. Dan bagi produsen, agar mampu meningkatkan produk sesuai dengan kebutuhan konsumen.

1.2 Pertanyaan Masalah

Berdasarkan paparan latar belakang yang telah disampaikan. Penulis merumuskan pertanyaan penelitian yang nantinya akan terjawab lewat penelitian ini. Yakni:

1. bagaimana perbandingan efektifitas aturan sintaksis dan TF-IDF dalam melakukan ekstraksi aspek pada review atau opini berbahasa indonesia.
2. bagaimana perbandingan NGD dan WordNet pada pengukuran kesamaan kata berbahasa indonesia.

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. mendapatkan nilai efektifitas dari aturan sintaksis dan TF-IDF dalam ekstraksi aspek pada opini review di bahasa indonesia
2. mendapatkan nilai perbandingan penerapan metode NGD dan WordNet dalam melakukan pengukuran kesamaan kata berbahasa indonesia

1.4 Manfaat Penelitian

Secara umum, pada penelitian ini terdapat 2 manfaat yang didapatkan yakni untuk akademisi dan praktisi. Manfaat tersebut antara lain:

1. Manfaat Untuk akademisi
 - a. Sebagai referensi akademisi dalam melakukan pengembangan model bahasa indonesia berbasis mesin.
 - b. Dapat dijadikan perbandingan pada penelitian selanjutnya, untuk dilakukan penyempurnaan dari sisi limitasi.
2. Manfaat untuk praktisi
 - a. Sebagai produsen, fitur yang telah diekstrak dapat dijadikan dasar pengembangan lanjutan dari fitur kebutuhan konsumen.
 - b. Sebagai konsumen, dapat meningkatkan kepedulian pada fitur dalam produk sehingga resiko untuk kecewa terhadap barang lebih berkurang.

1.5 Batasan Masalah

Penelitian ini memiliki batasan masalah antara lain:

1. penelitian ini akan berfokus untuk melakukan ekstraksi aspek pada komentar yang ada di toko online
2. Data yang digunakan dalam penelitian ini merupakan data review yang bisa diakses publik pada tokopedia dalam kategori elektronik

1.6 Sistematika Penulisan

Bab I Pendahuluan

Bagian bab yang menjabarkan latar belakang penelitian, dari penjabaran tersebut kemudian dirumuskan beberapa hal mencakup penelitian. Yakni; Masalah penelitian, tujuan Penelitian, manfaat penelitian, dan batasan masalah penelitian. Serta dilanjutkan dengan sistematika penulisan dalam penelitian ini.

Bab II Tinjauan Literatur

Merupakan bagian penelitian yang memuat beberapa literatur terdahulu dan memiliki relevansi pada penelitian. Pada bab ini pula terkandung bahasan kontekstual dan landasan teoritis bagi penelitian yang dilakukan

Bab III Metodologi Penelitian

Merupakan bagian penelitian yang memaparkan bagaimana keseluruhan penelitian dilakukan. Bagian ini memuat desain sistem, tahapan penelitian, dan analisis data

Bab IV Pengaplikasian klustering menggunakan kamus kata

Merupakan bagian penelitian yang menjelaskan intepretasi dan analisis secara mendalam atas hasil penelitian yang telah dilakukan menggunakan metode *dictionary based*.

Bab V Pengaplikasian klustering menggunakan NGD

Merupakan bagian penelitian yang menjelaskan intepretasi dan analisis secara mendalam atas hasil penelitian yang telah dilakukan menggunakan metode *Neutralized Google Distance*.

Bab VI Pembahasan

Bab ini menjelaskan hasil melakukan komparasi dari hasil penelitian yang telah dilakukan terhadap penelitian yang sudah ada sebelumnya.

Bab VII Kesimpulan

Bab ini menjelaskan kesimpulan dari hasil penelitian yang telah dilakukan. Dalam bagian ini, juga berisi saran yang dapat dijadikan acuan untuk penelitian selanjutnya.

BAB II

TINJAUAN LITERATUR

2.1 *Feature Extraction*

Secara umum, untuk melakukan *feature extraction* terdapat 2 cara, *supervised* dan *unsupervised*. Metode *supervised* merupakan salah satu metode *machine learning* dimana ekstraksi aspek didasari oleh data yang dilabeli secara manual untuk kemudian dijadikan sebagai data *training*. Dalam *feature extraction*, Hasil dari pengolahan data *training* ini kemudian, digunakan sebagai model untuk melakukan ekstraksi fitur pada sebuah dokumen. Misalnya, Kobayashi (Kobayashi et al., 2007), melakukan ekstraksi aspek pada beberapa blog dengan menyiapkan beberapa kamus kata. Selanjutnya dari kata tersebut dilakukan komparasi untuk menentukan antara hubungan satu kata dengan kata lain.

Disisi lain terdapat *unsupervised learning*, dimana bisa dilakukan ekstraksi fitur dari sebuah dokumen produk tanpa harus melalui proses training yang kompleks. Hu Pada penelitian sebelumnya (Hu & Liu, 2004) mengajukan metode dengan cara melakukan perhitungan pada banyaknya kata yang ada pada sebuah review untuk melakukan ekstraksi fitur. Namun, metode ini memiliki batasan pada ketidak sesuaian kata pada sebuah review. Misalkan saja, terdapat review "saya membeli handphone bagus ini di pasar". Kata "pasar", jika memiliki frekuensi yang tinggi secara otomatis akan dianggap sebagai aspek. Metode ini disempurnakan oleh Wei (Wei et al., 2010). Dengan mengajukan beberapa aturan *heuristic* berdasarkan pada *syntactic rules*. Hal ini mampu meningkatkan akurasi dari metode sebelumnya sebanyak 2 - 10%. Penyempurnaan juga dilakukan oleh bancken (Bancken et al., 2014), yakni dengan melakukan otomasi pada deteksi sentimen pada tiap aspek, sehingga bisa diketahui bagian mana dari sebuah produk yang memiliki sentimen negatif dan positif.

2.1.1 *Rule Based Feature Extraction*

Pada penelitian milik bancken (Bancken et al., 2014), ditemukan ASPECTATOR. ASPECTATOR merupakan salah satu metode yang menggunakan rule based untuk melakukan ekstraksi aspek. Pada

ASPECTATOR didefinisikan beberapa *syntactic rules*. Hasil dari *syntactic rules* tersebut, kemudian di kelompokkan berdasarkan nilai *similarity*. *Similarity* diukur berdasarkan frekuensi kemunculan kata pada WordNet. Sehingga, kata dianggap paling umum ketika muncul paling sering dengan nilai *similarity* terbesar.

Pada penelitian lain milik Rana(Rana & Cheah, 2017), *rule-based* juga digunakan untuk melakukan ekstraksi aspek. Seperti halnya pada ASPECTATOR, beberapa rules juga didefinisikan. Untuk selanjutnya dikelompokkan berdasarkan *similarity*. Metode pengukuran similarity yang digunakan adalah Google similarity distance(Cilibrasi & Vitanyi, 2007). Hasil dari penelitian ini, menunjukkan bahwa penggunaan *rule-based* cukup mampu melakukan ekstraksi. Selain itu, dapat meningkatkan akurasi dari penelitian sebelumnya.

Di penelitian terbaru oleh Luo(Luo et al., 2019), *rule-based* juga digunakan untuk melakukan ekstraksi fitur untuk mendapatkan aspek dari sebuah dokumen. Guna meningkatkan akurasi, dalam penelitian ini digunakan pula *directed graph* yakni dengan penerapan *aspect aggregation* dan penyusunan *graph* menggunakan Probase(Wu et al., 2012). Penggunaan probase sementara ini hanya mencakup pada bahasa inggris. Dan dihasilkan hasil yang lebih baik dari beberapa penelitian sebelumnya dalam ekstraksi aspek menggunakan rule.

2.1.2 TF-IDF

Term Frequency - Inverse Document Frequency (TF-IDF) merupakan salah satu metode yang umum digunakan. Metode ini digunakan untuk melakukan pembobotan pada data yang memiliki banyak variasi. TF-IDF menganggap dua dokumen sama, ketika dua dokumen tersebut memiliki kesamaan beberapa kata informatif. Dimana konsep ini merupakan salah satu cara untuk mengurangi *noise* dari sebuah dokumen. Rumus dasar dari pembobotan TF-IDF bisa di lihat pada rumus 2-1

$$TFIDF_{t,d} = TF_{d,t} \cdot \log\left(\frac{N}{|\{d \in \mathcal{D} : t \in d\}|}\right) \quad (2-1)$$

Dari rumus tersebut dapat diketahui bahwa nilai $|\mathcal{D}|$ merupakan total dokumen pada corpus. $TF_{d,t}$ adalah jumlah ditemukannya kata t pada

dokumen d . Adapun $DF_{t,D}$ ialah jumlah dokumen yang memuat t .

2.1.2.1 *Term Frequency*

Term Frequency (TF) mengukur banyaknya frekuensi kemunculan kata dalam sebuah dokumen. Dalam kenyataannya, perbedaan panjang dokumen menjadikan nilai TF berubah. Misalkan saja kata dalam dokumen dengan kalimat panjang, akan memiliki nilai lebih banyak dibandingkan dengan dokumen yang pendek. Sehingga, notasi untuk TF tersebut dapat dilihat pada persamaan 2-2.

$$TF_{d,t} = \frac{\Sigma_t}{\Sigma_d} \quad (2-2)$$

Σ_t merupakan total kemunculan sebuah kata dalam dokumen. Sedangkan Σ_d adalah banyaknya kata yang muncul dalam dokumen.

2.1.2.2 *Inverse Document Frequency*

Inverse Document Frequency digunakan untuk mengukur seberapa jauh informasi yang dikandung oleh sebuah kata. Perlakuan ini dilakukan untuk lebih menyaring nilai dari persamaan TF, dimana seluruh kata dianggap sama. Beberapa kata dengan kuantitas dan lebih sering muncul dianggap tidak terlalu penting, Misalkan saja pada kata "yang", "dan", maupun beberapa kata lain. Sehingga untuk mengurangi kata tidak informatif tersebut, digunakan persamaan 2-3.

$$idf_{d,t} = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right) \quad (2-3)$$

Dalam persamaan 2-3 N merupakan total keseluruhan dokumen, sedangkan pembaginya merupakan jumlah dokumen yang memuat kata terpilih. Dalam beberapa kasus perlu dilakukan normalisasi jika sebuah kata tidak terdapat didalam N yakni menambahkan nilai 1 pada pembaginya, menjadi $(1 + |\{d \in D : t \in d\}|)$

2.2 *Similarity Measurement*

Pengukuran *Similarity Measurement* yang peneliti gunakan adalah pengukuran dalam teks. Untuk mengukur *similarity* pada teks, beberapa variabel telah disebutkan pada penelitian(Lin et al., 2013). Nilai dari pengukuran *similarity* pada teks didapatkan dari ada tidaknya kesamaan

dari dua dokumen atau lebih. Nilai dari *similarity* berubah dipengaruhi oleh ada tidaknya kata dalam dokumen tersebut.

Untuk mengukur kesamaan kata dalam dokumen, beberapa model untuk menghitungnya telah diusulkan. Antara lain adalah TF-IDF dan NGD. Secara umum, TF-IDF merupakan model untuk menghitung kesamaan sebuah kata berdasarkan kemunculannya pada tiap dokumen. Sedangkan untuk NGD, memanfaatkan hasil dari mesin pencari yaitu Google untuk menghitung hubungan antar sebuah kata.

2.2.1 NGD

Normalized Google Distance (NGD) dikemukakan pertama kali oleh Cilibrasi (Cilibrasi & Vitanyi, 2007). NGD merupakan turunan dari *Normalized Compression Distance*, yang telah dilakukan pengujian pada beberapa penelitian semisal pada WordNet. Pada artikel tersebut, disebutkan pula untuk akurasi dari NGD dibandingkan dengan svm mampu mendapatkan nilai sebesar 75% untuk perhitungan similarity. Persamaan untuk pada model NGD dituliskan pada rumus 2-4

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x), \log f(y)\}} \quad (2-4)$$

Dalam rumus 2-4 diketahui $f(x)$ jumlah dari halaman yang diindeks Google pada kata x . Begitupula dengan $f(y)$, menunjukkan total halaman dari pencarian y . Sedangkan N adalah asumsi dari total halaman yang diindeks google. Ketika penelitian ini dilakukan, lebih dari 1×10^{12} gigabyte website telah di indeks pada Google.

Penelitian terkait NGD ini telah banyak diterbitkan. Pada Penelitian karve(Karve et al., 2019), improvisasi kesamaan semantik kata pada wikipedia dapat dilakukan. Dimana pada penelitian tersebut, *query* yang dipasangkan oleh intuisi manusia, dapat disamai dengan hasil prediksi keterkaitan dengan NGD.

Pada penelitian Mao(Mao et al., 2020) juga digunakan NGD sebagai metode untuk melakukan penilaian pada *semantic similarity*. Pada penelitian ini terdapat membandingkan dua metode yakni, kemunculan kata dan kesamaan kata. Untuk kesamaan kata digunakan tiga metode yakni penggunaan *dictionary based* menggunakan WordNet, Word Embedding, dan

menggunakan NGD. Hasil penelitian tersebut dihasilkan bahwa kombinasi dari kedua metode tersebut dengan penerapan NGD sebagai metode untuk pengukuran kesamaan kata memberikan hasil terbaik dibandingkan dengan 2 metode lainnya.

2.2.2 Dictionary Based

Untuk pengukuran kesamaan kata menggunakan *dictionary*, kami menggunakan WordNet. *Dictionary* ini merupakan kumpulan data leksikal yang digunakan untuk memodelkan sebuah keterkaitan antar kata. WordNet dikembangkan oleh Princeton University¹, dan hingga kini tetap aktif dikembangkan oleh komunitas peneliti dalam berbagai bahasa lainnya. Pada WordNet terdiri dari kata yang kemudian dikelompokkan kedalam susunan hirarki, dimana susunan ini disebut sebagai synsets. Pada tiap synset ini dihubungkan oleh konsep semantik kata yang terdiri dari hipernim, homonim, dan sinonim.

Dalam pengukuran kesamaan kata, salah satu algoritma yang digunakan oleh WordNet adalah *shortest path*. Dimana pada WordNet akan diberikan input 2 kata dan akan mengeluarkan hasil berupa kesamaan input tersebut dalam nilai 0 dan 1. Dimana 1 merupakan indikasi bahwa kata tersebut memiliki kesamaan. sebaliknya, jika nilai yang didapat adalah 0, maka kata tersebut tidak memiliki kesamaan. Secara matematis dapat dituliskan;

$$Sim(x,y) = \frac{1}{(lengthpath + 1)} \quad (2-5)$$

Dimana pada rumus 2-5, *lengthpath* merupakan jumlah dari banyaknya garis yang menghubungkan antara 1 kata dengan kata lainnya.

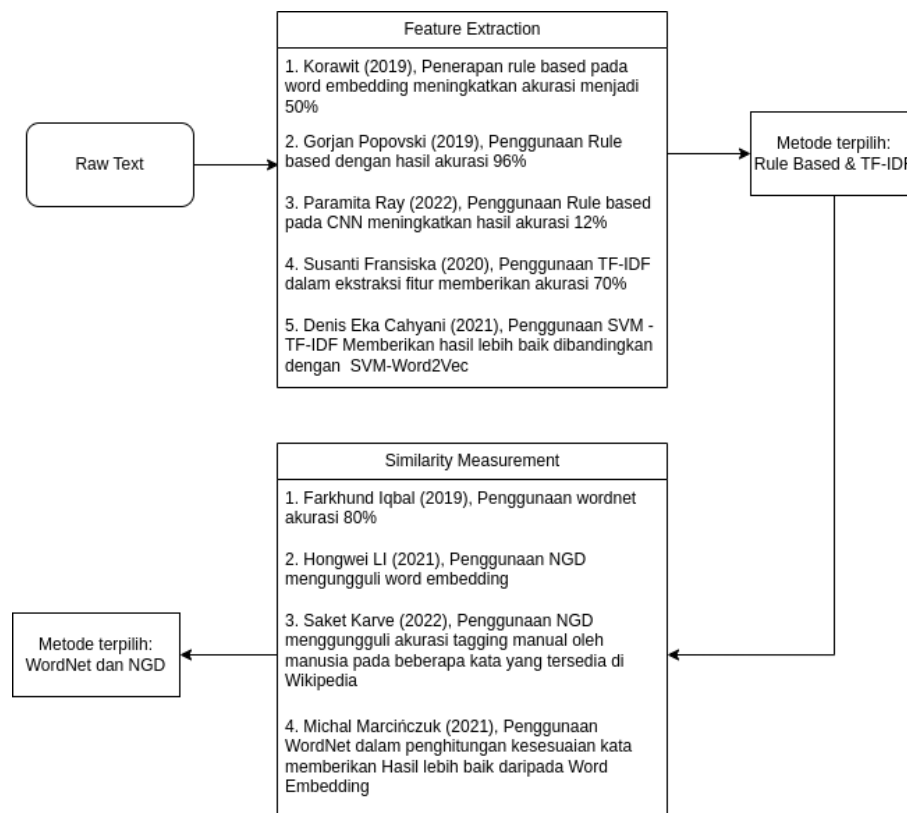
Penelitian oleh Michal Marcińczuk (Marcińczuk et al., 2021) melakukan komparasi antara metode lama yakni TF-IDF dan wordnet dengan metode baru yakni word-embedding. Pada 4 kategori dataset tentang kualifikasi kerja, Dalam akurasi, metode lama memiliki beberapa keunggulan pada 2 dataset dibandingkan dengan metode baru. Sementara untuk kecepatan training, metode lama memiliki waktu yang lebih lama dibanding *word-embedding*

¹<https://wordnet.princeton.edu/>

Adapun penelitian lain oleh Farkhund Iqbal (Iqbal et al., 2019) untuk melakukan deteksi pada teks pesan singkat pada kasus kriminal. Dengan menggunakan dataset asli, memberikan hasil yang cukup baik. Dimana lebih dari 80% teks pesan singkat tersebut dapat didapatkan aspek yang saling berkaitan dengan tindak pidana, dapat terekstrak dengan baik dan benar.

2.3 Kerangka Teori

Kerangka teori ini berlandaskan pada beberapa penelitian ilmiah yang telah ada sebelumnya, diantaranya bisa terlihat pada gambar 2.1. Dari hasil review tersebut, kami menemukan bahwa TF-IDF dan *Rule based* merupakan 2 diantara beberapa metode dengan hasil cukup baik. Sedangkan untuk pengukuran *Word Similarity*, kami menggunakan NGD dan Dictionary based, Dalam hal ini adalah WordNet. Hasil telaah penelitian ilmiah sebelumnya, kami jabarkan pada tabel 2.1



Gambar 2.1. Kerangka teori

Tabel 2.1. Penelitian Sebelumnya

1	Peneliti dan Judul penelitian	Metode yang digunakan
	Orkphol, K., and Yang, W. (2019) Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet.	Rule Based and word embedding
	Hasil Penelitian	Perbedaan
	Penelitian ini menyimpulkan bahwa penggunaan <i>rule based</i> dalam melakukan ekstraksi fitur untuk training akan mempertajam hasil maksud dari kata dibanding metode awal yang hanya berupa <i>word embedding</i> . Dengan improvisasi memiliki rentang dari 48% - 50%	Penggunaan sintaksis yang dinamis, pada penelitian ini menggunakan aturan sintaksis bahasa indonesia.
2	Peneliti dan Judul penelitian	Metode yang digunakan
	Marcińczuk, M., (2021) Text document clustering: Wordnet vs. TF-IDF vs. word embeddings.	WordNet, word embedding, TF-IDF
	Hasil Penelitian	Perbedaan

	<p>Penelitian ini melakukan komparasi antara metode lama dari ekstraksi fitur yakni TF-IDF dan wordnet dengan metode baru yakni word-embedding. Pada 4 kategori dataset tentang kualifikasi kerja, Dalam akurasi, metode lama memiliki beberapa keunggulan pada 2 dataset dibandingkan dengan metode baru. Sementara untuk kecepatan training, metode lama memiliki waktu yang lebih lama dibanding <i>word-embedding</i></p>	<p>Penggunaan dataset WordNet yang berbeda. Pada penelitian ini kami menggunakan WordNet berbasis bahasa indonesia. Sehingga akan terdapat scoring untuk nilai kata.</p>
3	<p>Peneliti dan Judul penelitian</p>	<p>Metode yang digunakan</p>
	<p>Karve, S., (2019) Semantic relatedness measurement from Wikipedia and WordNet using modified normalized google distance.</p>	<p>Neutralized Google Distance, Point Mutual Information (Menggunakan intuisi manusia)</p>
	<p>Hasil Penelitian</p>	<p>Perbedaan</p>

	<p>Penelitian ini melakukan komparasi kesamaan kata yang pada sebuah Dataset. Performa PMI-IR dalam menghitung kesamaan antara dua kata dilampaui oleh NGD, Dimana pada kasus ini NGD dapat memberikan nilai yang secara langsung telah dinormalisasi. Penelitian tersebut mengungkap keefektifan NGD meningkat ketika kedua kata tersebut tersedia di Wikipedia atau WordNet.</p>	<p>NGD pada penelitian tersebut, menggunakan versi lebih umum dari google. Dimana bahasa utama dari google menggunakan bahasa inggris, sedangkan penelitian ini menggunakan bahasa indonesia.</p>
4	Peneliti dan Judul penelitian	Metode yang digunakan
	Mao, X., (2020) Automatic keywords extraction based on co-occurrence and semantic relationships between words.	Neutralized Google Distance, Word Embedding
	Hasil Penelitian	Perbedaan
	Penelitian tersebut menyimpulkan bahwa penggunaan metode NGD sedikit mengungguli word embedding dalam melakukan kalkulasi hubungan semantik antar kata.	NGD pada penelitian tersebut, menggunakan versi lebih umum dari google. Dimana bahasa utama dari google menggunakan bahasa inggris, sedangkan penelitian ini menggunakan bahasa indonesia. Begitu pula dengan objek yang diteliti juga terdapat perbedaan.
	Peneliti dan Judul penelitian	Metode yang digunakan

	Popovski, G., (2019) FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction.	Rule based
	Hasil Penelitian	Perbedaan
	Penelitian menggunakan <i>syntactical rules</i> untuk melakukan pengenalan entitas kata makanan pada dataset resep. Penelitian ini menghasilkan akurasi hingga 96%.	Pada penelitian terdahulu, pengenalan entitas merujuk pada Kata benda saja dalam daftar langkah pembuatan makanan. Sedangkan pada penelitian ini, menggunakan aturan sintaksis untuk objek dan subjek.
6	Peneliti dan Judul penelitian	Metode yang digunakan
	Ray, P., and Chakrabarti, A. (2022) A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis.	CNN, CNN with Rule based
	Hasil Penelitian	Perbedaan
	Penelitian melakukan improvisasi pada ekstraksi implisit fitur dengan menggunakan rule based. Penggunaan <i>syntactical rules</i> dalam melakukan ekstraksi fitur, meningkatkan keakurasian hingga 12% dibandingkan dengan hanya menggunakan CNN.	Aturan sintaksis yang digunakan merupakan aturan sintaksis bahasa inggris, pada penelitian ini kami menggunakan bahasa indonesia
7	Peneliti dan Judul penelitian	Metode yang digunakan
	Li, H., (2021) Part-of-speech tagging with rule-based data preprocessing and transformer.	Bi-LSTM, Bi-LSTM with Rule based

	<table border="1"> <thead> <tr> <th>Hasil Penelitian</th> <th>Perbedaan</th> </tr> </thead> <tbody> <tr> <td> <p>Penelitian melakukan improvisasi pada ekstraksi implisit fitur dengan menggunakan rule based. Penggunaan <i>syntactical rules</i> dalam melakukan ekstraksi fitur, meningkatkan keakurasian hingga 12% dibandingkan dengan hanya menggunakan CNN.</p> </td> <td> <p>Aturan sintaksis yang digunakan merupakan aturan sintaksis bahasa inggris, pada penelitian ini kami menggunakan bahasa indonesia</p> </td> </tr> </tbody> </table>	Hasil Penelitian	Perbedaan	<p>Penelitian melakukan improvisasi pada ekstraksi implisit fitur dengan menggunakan rule based. Penggunaan <i>syntactical rules</i> dalam melakukan ekstraksi fitur, meningkatkan keakurasian hingga 12% dibandingkan dengan hanya menggunakan CNN.</p>	<p>Aturan sintaksis yang digunakan merupakan aturan sintaksis bahasa inggris, pada penelitian ini kami menggunakan bahasa indonesia</p>				
Hasil Penelitian	Perbedaan								
<p>Penelitian melakukan improvisasi pada ekstraksi implisit fitur dengan menggunakan rule based. Penggunaan <i>syntactical rules</i> dalam melakukan ekstraksi fitur, meningkatkan keakurasian hingga 12% dibandingkan dengan hanya menggunakan CNN.</p>	<p>Aturan sintaksis yang digunakan merupakan aturan sintaksis bahasa inggris, pada penelitian ini kami menggunakan bahasa indonesia</p>								
8	<table border="1"> <thead> <tr> <th>Peneliti dan Judul penelitian</th> <th>Metode yang digunakan</th> </tr> </thead> <tbody> <tr> <td> <p>Iqbal, F., (2019) Wordnet-based criminal networks mining for cybercrime investigation.</p> </td> <td> <p>WordNet, Agglomerative clustering</p> </td> </tr> <tr> <th>Hasil Penelitian</th> <th>Perbedaan</th> </tr> <tr> <td> <p>Penelitian melakukan deteksi pada teks pesan singkat pada kasus kriminal, Hasilnya lebih dari 80% teks pesan singkat tersebut dapat terekstrak dengan baik dan benar.</p> </td> <td> <p>Perbedaan pada model analisis, penelitian ini menggunakan silhouette analysis. Begitupula dengan objek, dimana pesan teks akan lebih terkontrol dibandingkan dengan komentar.</p> </td> </tr> </tbody> </table>	Peneliti dan Judul penelitian	Metode yang digunakan	<p>Iqbal, F., (2019) Wordnet-based criminal networks mining for cybercrime investigation.</p>	<p>WordNet, Agglomerative clustering</p>	Hasil Penelitian	Perbedaan	<p>Penelitian melakukan deteksi pada teks pesan singkat pada kasus kriminal, Hasilnya lebih dari 80% teks pesan singkat tersebut dapat terekstrak dengan baik dan benar.</p>	<p>Perbedaan pada model analisis, penelitian ini menggunakan silhouette analysis. Begitupula dengan objek, dimana pesan teks akan lebih terkontrol dibandingkan dengan komentar.</p>
	Peneliti dan Judul penelitian	Metode yang digunakan							
	<p>Iqbal, F., (2019) Wordnet-based criminal networks mining for cybercrime investigation.</p>	<p>WordNet, Agglomerative clustering</p>							
Hasil Penelitian	Perbedaan								
<p>Penelitian melakukan deteksi pada teks pesan singkat pada kasus kriminal, Hasilnya lebih dari 80% teks pesan singkat tersebut dapat terekstrak dengan baik dan benar.</p>	<p>Perbedaan pada model analisis, penelitian ini menggunakan silhouette analysis. Begitupula dengan objek, dimana pesan teks akan lebih terkontrol dibandingkan dengan komentar.</p>								
9	<table border="1"> <thead> <tr> <th>Peneliti dan Judul penelitian</th> <th>Metode yang digunakan</th> </tr> </thead> <tbody> <tr> <td> <p>Fransiska, S., (2020) Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method.</p> </td> <td> <p>SVM, TF-IDF</p> </td> </tr> <tr> <th>Hasil Penelitian</th> <th>Perbedaan</th> </tr> </tbody> </table>	Peneliti dan Judul penelitian	Metode yang digunakan	<p>Fransiska, S., (2020) Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method.</p>	<p>SVM, TF-IDF</p>	Hasil Penelitian	Perbedaan		
	Peneliti dan Judul penelitian	Metode yang digunakan							
	<p>Fransiska, S., (2020) Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method.</p>	<p>SVM, TF-IDF</p>							
Hasil Penelitian	Perbedaan								

	<p>Penelitian ini melakukan analisa sentimen pada komentar di Google play. Dalam penelitian ini TF-IDF digunakan untuk melakukan <i>feature extraction</i> dimana dihasilkan bahwa TF IDF bisa digunakan untuk melakukan ekstraksi fitur dengan hasil yang cukup baik.</p>	<p>Penggunaan metode dalam mengukur <i>similarity measurement</i>. Penelitian terdahulu menggunakan <i>support vector machine</i>, sedangkan penelitian ini menggunakan NGD dan <i>dictionary based</i></p>
10	Peneliti dan Judul penelitian	Metode yang digunakan
	Cahyani, D. E., and Patasik, I. (2021) Performance comparison of tf-idf and word2vec models for emotion text classification.	SVM dan TF-IDF, SVM - Word2Vec
	Hasil Penelitian	Perbedaan
	Penelitian ini melakukan komparasi antara TF-IDF dan Word2Vec dalam melakukan klasifikasi teks yang mengandung emosi. Kombinasi dari SVM dan TF-IDF memberikan hasil akurasi tertinggi dibandingkan metode lain.	Perhitungan <i>similarity measurement</i> menggunakan SVM, berbeda dengan penelitian ini yang menggunakan NGD dan <i>dictionary based</i>

2.4 Aspect Extraction Perspektif Islam

Penelitian ini berkaitan erat dengan *muamalah* jual beli. Dimana Pada Alquran surat An Nisa ayat 29, Disebutkan:

Penjelasan Shihab(Shihab, 2009) tentang surat An Nisa ini hendaknya transaksi jual beli tidak saling merugikan. Dimana salah satu parameter dari kesesuaian benda, dapat diketahui dari review orang lain pada produk tersebut. Penelitian ini memudahkan bagaimana menyimpulkan review dari

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَأْكُلُوا أَمْوَالَكُم بَيْنَكُم بِالْبَاطِلِ إِلَّا أَنْ تَكُونَ بِجَارَةً عَنْ تَرَاضٍ مِّنْكُمْ وَلَا تَقْتُلُوا
 أَنْفُسَكُمْ إِنَّ اللَّهَ كَانَ بِكُمْ رَحِيمًا

Hai orang-orang yang beriman, janganlah kamu saling memakan harta sesamamu dengan jalan yang batil, kecuali dengan jalan perniagaan yang berlaku dengan suka sama-suka di antara kamu. Dan janganlah kamu membunuh dirimu; sesungguhnya Allah adalah Maha Penyayang kepadamu.

Gambar 2.2. Ilustrasi Alquran surat An Nisa Ayat 29

orang lain pada benda tersebut. Sehingga, hal ini sangat bersesuaian dengan ajaran islam.

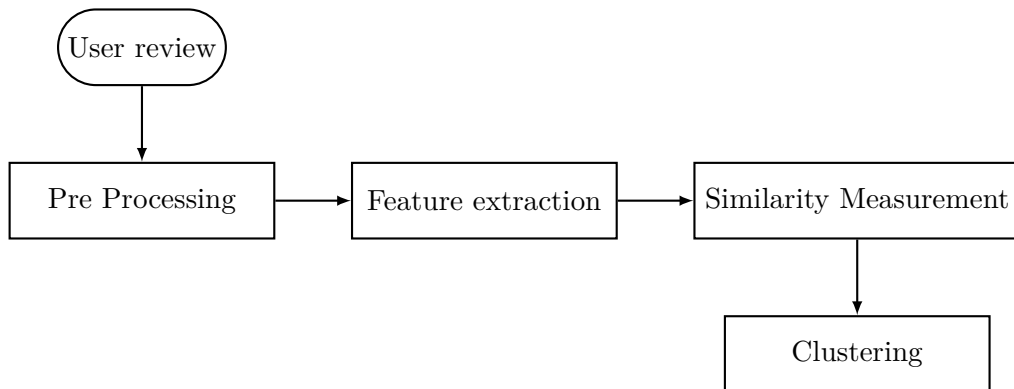
Dalam tinjauan hadits, bukhari telah menuliskan dalam shahih bukhari (Al-Bukhari et al., 1978) pada hadis nomor 1993. Bahwa perilaku ceroboh dalam berjual beli, yakni tidak menimbang barang dengan jelas. Tidak diperkenankan oleh rasulullah. Jika dikaitkan dengan hal ini, pembelian online perlu kejelian tersendiri dalam melihat barang yang dijual. Sehingga, lewat penelitian ini, mampu meningkatkan kesadaran pada pihak pembeli dalam melakukan pemilihan barang.

Rusyd (Rusyd, n.d.), dalam perspektif fiqh menjelaskan bahwa hukum dasar dalam jual beli, ketika terdapat sesuatu yang dianggap bisa mengurangi harga dari sebuah barang. maka barang tersebut wajib dikembalikan. Seiring dengan itu, terdapat penafsiran lebih luas perihal barang cacat dalam konteks jual beli tersebut.

BAB III

METODE PENELITIAN

3.1 Prosedur Penelitian



Gambar 3.1. Diagram Prosedur Penelitian

Prosedur penelitian untuk melakukan uji empiris dapat di temukan pada gambar 3.1. Pada tahap *system development* tersebut diawali dengan proses *preprocessing*. Pada tahapan *preprocessing*, tahapan pembersihan data dilakukan. Selanjutnya dilakukan proses *feature extraction*, dimana pada tahap ini akan diambil fitur dari tiap document. Kemudian dilakukan perhitungan pada kesamaan kata pada proses *similarity measurement*. Ditutupi dengan proses *clustering*

3.2 Pengumpulan Data

Penelitian ini menggunakan objek data berupa data komentar yang didapat dari toko elektronik, yakni Tokopedia. Data komentar ini bersifat publik. Adapun untuk menggunakannya kami menggunakan beberapa tools *parser* yakni Selenium ¹. Untuk mesin yang kami gunakan adalah sebuah perangkat komputer dengan processor Intel I5 dengan RAM 16 GB serta GPU Nvidia GTX 1050ti dengan kapasitas memory 4 GB.

Data yang awalnya berupa markup html, kami ekstrak menjadi csv dengan susunan.

- *text*, kalimat komentar untuk sebuah produk

¹<https://www.selenium.dev/>

- *rating*, jumlah bintang yang disematkan pada produk yang dilakukan oleh user.
- *category*, nama kategori dari produk
- *Product name*, Nama dari produk
- *product url*, link dari product

Sebagai gambaran, dataset yang kami gunakan dapat dilihat pada Tabel 3.1. Tahapan ini, data yang dihasilkan dari *crawler* merupakan data mentah. Selanjutnya, kami akan masuk ke tahapan *Preprocessing*.

Tabel 3.1: Contoh dataset

text	rating	Category	ProductName	ProductUrl
Berfungsi dengan baik....	5	elektronik	Alfalink EI 212 - Kamus Elektronik	https://www.tokopedia.com/omegaelectronic/alfalink-ei-212-kamus-elektronik
Seller fast response, pengiriman cepat, product sesuai deskripsi, semoga awet, buat kado ponakan semoga suka 😊	5	elektronik	Alfalink EI 212 - Kamus Elektronik	https://www.tokopedia.com/omegaelectronic/alfalink-ei-212-kamus-elektronik

Tabel 3.1: Contoh dataset (lanjutan)

text	rating	Category	ProductName	ProductUrl
Berfungsi dengan baik kualitas okey	5	elektronik	Alfalink EI 212 - Kamus Elektronik	https://www.tokopedia.com/omegaelectronic/alfalink-ei-212-kamus-elektronik
barang diterima dgn aman, penjual juga ramah	5	elektronik	Alfalink EI 212 - Kamus Elektronik	https://www.tokopedia.com/omegaelectronic/alfalink-ei-212-kamus-elektronik
sesuai gambar, respon cepat	5	elektronik	Alfalink EI 212 - Kamus Elektronik	https://www.tokopedia.com/omegaelectronic/alfalink-ei-212-kamus-elektronik

3.3 Tahapan Penelitian

Dalam tujuan untuk memperjelas prosedur yang dilakukan pada tahapan prosedur penelitian, pada bagian ini kami akan menjelaskan secara umum tahapan demi tahapan yang dilakukan pada proses penelitian. Mencakup:

1. Pengumpulan data, dimana tahap ini dapat dilihat pada bagian 3.2.
2. Preprocessing. Secara mendetail, proses ini dapat dilihat pada sub bab 3.3.1
3. *Feature Extraction*. Output dari proses *Preprocessing*, kemudian di olah pada tahapan ini. Secara mendetail dapat dilihat pada 3.3.2
4. *Similarity Measurement*. Output dari proses *Feature Extraction*,


kemudian di olah pada tahapan ini, pada tahapan ini akan didapatkan nilai numerik dari tiap kata. Secara mendetail dapat dilihat pada 3.3.3

5. *Clustering*. Hasil kata yang telah diketahui kesamaannya, kemudian dilakukan clustering. Secara detail di jelaskan pada 3.3.4
6. *Quality Measurement*. untuk mengetahui hasil *clustering* terbaik, maka dilakukan proses analisa. Secara detail di jelaskan pada 3.3.5

3.3.1 Preprocessing

Tahap *preprocessing* melakukan pembersihan data dari bagian yang tidak diperlukan dalam sistem. Dalam penelitian kali ini, berikut adalah penjabaran dari tahapan tersebut:

- a. *Cleansing*. Pada tahap ini dilakukan pembersihan karakter tidak diperlukan dari dokumen. Karakter tersebut mencakup simbol tanda baca, angka, emoji atau emoticon. Pada ilustrasi 3.2, kami menghilangkan tanda baca dan emoji.

<p>pengiriman cepat, semoga awet, buat kado ponakan semoga suka </p> <p style="text-align: center;"> </p> <p>pengiriman cepat semoga awet buat kado ponakan semoga suka</p>

Gambar 3.2. Tahap cleansing document

- b. Stopword removal Selanjutnya, kata yang tidak mengandung makna signifikan akan dihilangkan. Dalam penelitian ini kami menggunakan kamus stopwords yang telah ada pada penelitian sebelumnya (Tala, 2003). Pada ilustrasi di gambar 3.3, kami melakukan penghilangan pada kata ‘buat’ karena kata tersebut masuk pada stopwords.

<p>pengiriman cepat semoga awet buat kado ponakan semoga suka</p> <p style="text-align: center;"> </p> <p>pengiriman cepat semoga awet kado ponakan semoga suka</p>

Gambar 3.3. Tahap stopwords removal document

- c. Stemming Ditahap ini, kami melakukan ekstraksi kata dasar pada tiap kata dalam kalimat. Pada gambar 3.4, kami melakukan pembentukan kata dasar dari ‘pengiriman’ menjadi ‘kirin’

pengiriman cepat semoga awet buat kado ponakan semoga suka kirim cepat semoga awet kado ponakan semoga suka

Gambar 3.4. Tahap stemming document

- d. Normalisasi Tahapan selanjutnya dilanjutkan dengan melakukan *normalization*. Pada teks review banyak sekali kata yang kurang baku dan tidak ada dalam kamus. Misalkan saja ‘ponakan’. Kata ‘ponakan’, tidak mungkin ditemukan di kamus bahasa indonesia (Indonesia, 2008). Sehingga, untuk mengatasi masalah ini peneliti menggunakan pendekatan PBSMT (Wibowo et al., 2020). Adapun jika hasilnya tetap, maka kami akan melakukan masking pada kata tersebut dan menghilangkannya dari dokumen. Hal ini bisa di lihat pada gambar 3.5.

kirim cepat semoga awet kado ponakan semoga suka kirim cepat semoga awet kado semoga suka

Gambar 3.5. Tahap Normalisasi document

3.3.2 Feature extraction

Pada bagian *feature extraction*, kalimat yang sudah melalui tahap *preprocess* akan diekstrak bagian kata. Bagian dari tiap kata ini mengacu sesuai dengan masing masing metode yang digunakan. Dalam Hal ini, adalah TF-IDF dan *Syntactical rules*.

3.3.2.1 TF-IDF

TF-IDF adalah singkatan dari "*Term Frequency-Inverse Document Frequency*". Ini adalah fungsi statistik yang digunakan untuk mengevaluasi seberapa penting sebuah kata bagi dokumen dalam corpus. Fungsi ini melakukan perhitungan frekuensi kata dalam dokumen, Selanjutnya menghitung frekuensi kata di seluruh kumpulan dokumen. Sebagai bagian kecil dari keseluruhan data, berikut adalah sampel perhitungan:

Dokumen 1: "fungsi baik"

Dokumen 2: "fungsi baik kualitas"

Dokumen 3: "kirim cepat semoga awet kado semoga suka"

Dari 3 Dokumen tersebut, selanjutnya akan dilakukan perhitungan *term frequency*. *Term Frequency* merupakan banyaknya kemunculan kata pada dokumen. Perhitungan ini menggunakan rumus yang ada pada persamaan 2-2 per dokumen. Dalam sampel perhitungan, kami mengambil contoh perhitungan kata 'baik' pada seluruh dokumen.

Dokumen 1: Kata baik keluar 1 kali dari 2 kata, sehingga nilai kata baik pada document 1 adalah $1/2$

Dokumen 2: Kata baik keluar 1 kali dari 3 kata, sehingga nilai kata baik pada document 2 adalah $1/3$

Dokumen 3: Kata baik keluar 0 kali dari 7 kata, sehingga nilai kata baik pada document 3 adalah $0/7$

Selanjutnya dilakukan perhitungan *Inverse Document Frequency* (IDF). *Inverse Document Frequency* merupakan nilai logaritma dari total dokumen didalam korpus dibagi dengan dokumen memuat kata dalam pencarian. Dalam teks sebelumnya, kami mencontohkan penggunaan kata 'baik'. Sehingga untuk idf dari kata baik adalah $\log(3/2)$

Untuk menghitung nilai TF-IDF, dari perhitungan diatas disubstitusikan kedalam persamaan 2-1. Sehingga menjadi:

Dokumen 1: Nilai TF-IDF untuk Kata baik pada dokumen 1 adalah $(1/2) \times \log(3/2) = 0.08$

Dokumen 2: Nilai TF-IDF untuk Kata baik pada dokumen 2 adalah $(1/3) \times \log(3/2) = 0.05$

Dokumen 3: Nilai TF-IDF untuk Kata baik pada dokumen 3 adalah $0 \times \log(3/2) = 0$

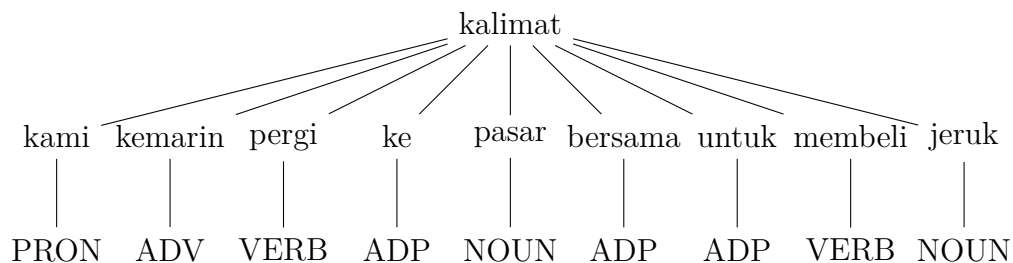
Perhitungan TF-IDF ini dilakukan per dokumen. Sehingga untuk melakukan kalkulasi dari nilai TF-IDF dari sebuah kata pada sebuah dokumen kami menggunakan persamaan Salient TF-IDF Features (Horn et al., 2017):

$$\sum_{d=1}^D \text{TF-IDF}(w,d) \quad (3-1)$$

Dalam persamaan tersebut, \sum adalah simbol penjumlahan skor TF-IDF dalam semua dokumen dalam koleksi, dan subskrip $d = 1$ di bawah simbol penjumlahan menunjukkan perhitungan dimulai pada dokumen pertama dalam koleksi. Superskrip D merupakan simbol bahwa penghitungan dilakukan hingga dokumen terakhir dalam koleksi. Terakhir, $\text{TF-IDF}(w,d)$ adalah skor TF-IDF dari kata w dalam dokumen d . Sehingga nilai TF-IDF dari kata ‘baik’ adalah 0.13.

3.3.2.2 Rule based Extraction

Penggunaan *rule based extraction* dalam penelitian ini adalah menggunakan aturan sintaksis. Pada penelitian ini kami mendefinisikan rule sintaksis yakni pengambilan subjek dan objek dari kalimat. Pengambilan sintaksis tersebut melibatkan penggunaan *part of speech tagging* (POS tagging). Untuk memudahkan pemahaman, berikut adalah ilustrasi pos tagging yang dimaksud:



Gambar 3.6. Ilustrasi POS Tagger

Pada proses *feature extraction* ini, kami menggunakan aturan dengan mengambil kata yang memiliki tag NOUN, ADJ. Dimana maksud dari 3 tag kata tersebut ialah:

- a. NOUN, merupakan kata benda. Dimana pada aturan sintaksis bahasa indonesia, kata ini akan menjadi subjek atau objek dari sebuah kalimat
- b. ADJ, merupakan kata sifat.

Untuk melakukan pemberian label dari tiap dokumen dalam penelitian ini, kami menggunakan metode *Conditional Random Field*. Sedangkan untuk

dataset yang digunakan adalah dataset yang telah ada pada penelitian sebelumnya (Rashel et al., 2014).

Dalam prosesnya, peneliti akan melakukan training menggunakan dataset yang telah ditag secara manual tersebut. Kemudian, model yang telah kami dapatkan dari hasil training akan kami gunakan untuk menentukan label pada dataset yang kami miliki. Selanjutnya, sesuai dengan rule yang telah kami sebutkan kami akan mendapatkan fitur yang akan digunakan.

3.3.3 Similarity Measurement

Pada proses ini, kami telah mendapat input berupa kata yang didapatkan dari proses *feature extraction*. Selanjutnya untuk melakukan pengukuran kesamaan dari tiap kata, kami akan menggunakan 2 metode yakni NGD dan *dictionary based*

Pada fitur tersebut dilakukan komparasi tiap kata, dimana pada kata W_1, W_2, \dots, W_n akan dibandingkan pada kata tersebut untuk mengukur persamaannya. Untuk lebih jelasnya bisa dilihat pada Tabel 3.2

3.3.3.1 NGD

Keluaran yang dihasilkan dari *feature extraction*, selanjutnya akan dijadikan input pada proses *similarity measurement*. Untuk pengukurannya, salah satunya kami menggunakan NGD. Untuk melakukan kalkulasi NGD, Proses pertama kali diawali dengan melakukan query pada Google dengan kata yang paling umum. Pada penelitian dasar, Cilibrasi (Cilibrasi & Vitanyi, 2007) melakukan query pada kata 'the'. Kemudian hasilnya akan disubstitusikan kedalam persamaan 2-4.

Perlakuan ini dapat dilihat pada, ilustrasi perhitungan kesamaan kata antara 'bikin' dan 'buat'. Diketahui penggunaan kata ketika melakukan query pada kata 'the' adalah 2527×10^7 . Selanjutnya dilakukan query untuk pencarian kata 'bikin' dan 'buat'. kata bikin memiliki hasil sejumlah 888×10^6 , sedangkan kata 'buat' memiliki hasil query sejumlah 171×10^7 . Selanjutnya perlu diketahui pula kemunculan dari dua kata secara bersamaan adalah 824×10^6 . Sehingga ketika dilakukan substitusi maka akan

Tabel 3.2. Similarity measurement

Kata pada dokumen W_1, W_2, \dots, W_n

		W_1	W_2	W_3	W_4	W_n
Kata pada dokumen W_1, W_2, \dots, W_n	W_1	0.55	0.16	0.04	0.30	...
	W_2	0.13	0.75	0.02	0.11	...
	W_3	0.03	0.01	0.91	0.00	...
	W_4	0.25	0.07	0.01	0.58	...
	W_n

$$\begin{aligned}
 &NGD(bikin, buat) \\
 &= \frac{\max\{\log(bikin), \log(buat)\} - \log(bikin, buat)}{\log N - \min\{\log(bikin), \log(buat)\}} \\
 &= \frac{\max\{\log(888 \times 10^6), \log(171 \times 10^7)\} - \log(824 \times 10^6)}{\log(2527 \times 10^7) - \min\{\log(888 \times 10^6), \log(171 \times 10^7)\}} \\
 &= 0,46
 \end{aligned}$$

Nilai yang didapatkan tersebut, akan menjadi parameter sesuai dengan Tabel 3.2. Sehingga, pada proses *clustering*, kata bisa dikelompokkan sesuai dengan kesamaan nilai pada tiap kata.

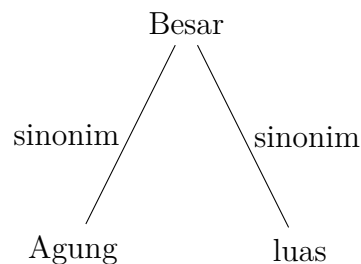
3.3.3.2 Dictionary Based

Adapun untuk pengukuran dengan metode *dictionary based*, kami menggunakan WordNet. WordNet merupakan database leksikal yang mengelompokkan kumpulan kata dan mengaturnya menjadi kumpulan

sinonim, yang disebut synsets, serta menjelaskan hubungan di antara kata tersebut.

Pada Setiap synset di WordNet akan memiliki konsep sendiri, kemudian terhubung ke synset lain melalui berbagai hubungan semantik bahasa. Hubungan ini termasuk hipernim (hubungan antara kata yang lebih umum dan kata yang lebih spesifik, misalnya "binatang" adalah hipernim dari "anjing"), hiponim (hubungan antara kata yang lebih spesifik dan kata yang lebih umum, misalnya "anjing" adalah hiponim dari "hewan"), antonim (hubungan antara kata-kata dengan makna yang berlawanan, misalnya "panas" dan "dingin"), meronimia (hubungan antara keseluruhan dan bagian-bagiannya, misalnya "roda" adalah meronim dari "mobil"), dan holonimi (hubungan antara bagian dan keseluruhannya, misalnya "mobil" adalah holonim dari "roda").

Sebagai contoh untuk pengambilan data sinonim, pada kata 'besar'. Pada wordnet kata ini memiliki sinonim kata yakni 'agung' dan 'luas'. Untuk ilustrasi graph dapat dilihat pada gambar 3.7



Gambar 3.7. Ilustrasi POS Tagger

3.3.4 Clustering

Setelah nilai dari kesamaan tiap kata didapatkan, selanjutnya akan dilakukan *clustering* untuk mengelompokkan aspek dengan berdasarkan kemiripan kata. Pada tahap ini, metode *clustering* yang digunakan adalah metode *Agglomerative clustering*. Algoritma *Agglomerative clustering* merupakan algoritma pengelompokan berbentuk hierarki. Dalam penggunaannya algoritma ini berasumsi bahwa setiap kata memiliki klusternya sendiri, dan kemudian secara iteratif menggabungkan dua kluster terdekat berdasarkan metrik kesamaan yang dipilih hingga semua kata menjadi satu kluster. Secara matematis dapat dituliskan sebagai berikut.

Dimana diasumsikan terdapat data point, X_1, X_2, \dots, X_n . Dan D merupakan matriks kesamaan dari kata tersebut. Sehingga, akan dibentuk sebuah cluster baru dari tiap data point, C . Dimana $C_1 = X_1, C_2 = X_2, \dots, C_n = X_n$

Pada proses iterasi ini, dilakukan kalkulasi berdasarkan kedekatan dari cluster berdasarkan keterkaitan/skema (Müllner, 2011).

- a. Skema tunggal: Jarak antara dua klaster adalah jarak minimum antara setiap pasangan titik data dalam dua klaster.
- b. Skema lengkap: Jarak antara dua klaster adalah jarak maksimum antara setiap pasangan titik data dalam dua klaster.
- c. Skema rata-rata: Jarak antara dua klaster adalah jarak rata-rata antara semua pasangan titik data dalam dua klaster.

Pada penelitian ini kami menggunakan skema tunggal, yang dinotasikan:

$$L(C_i, C_j) = \min\{D(C_i, K), D(C_j, K)\} \quad (3-2)$$

Kemudian data dari tiap cluster C , akan diiterasi secara terus menerus sedemikian hingga akan berhenti dengan aturan;

- a. Jumlah klaster telah ditemukan: Algoritme berhenti ketika jumlah klaster yang ditentukan sebelumnya tercapai.
- b. *distance threshold*: Algoritma berhenti ketika jarak antara pasangan cluster terdekat berada di atas ambang batas yang ditentukan sebelumnya.
- c. Iterasi maksimum: Algoritme berhenti setelah jumlah iterasi yang ditentukan sebelumnya.

Pada penelitian ini kami menggunakan batas pada iterasi maksimum, sehingga ketika iterasi maksimum telah ditemukan maka iterasi akan terhenti.

3.3.5 Quality Measurement

Dalam penelitian ini, untuk menilai kualitas hasil kami menggunakan *silhouette analysis*. Metode ini merupakan teknik analisa untuk mengevaluasi hasil dari *clustering* dengan mengukur kesamaan dari tiap data point dalam tiap cluster dengan perbandingan pada cluster lain. Dalam penerapannya,

berikut adalah langkah yang akan dilakukan yakni;

Pada asumsi, hasil dari proses sebelumnya akan menghasilkan beberapa cluster. Selanjutnya pada tiap kata didalam *cluster*, dilakukan perhitungan *silhouette score*. Digunakan persamaan;

$$SilhouetteScore = \frac{(b - a)}{\max(a, b)} \quad (3-3)$$

Dimana nilai a merupakan rata rata jarak antara point ke point, yang dalam ini adalah kata, dalam *cluster* yang sama. Sedangkan untuk b , merupakan nilai jarak minimum dari point pada *cluster* yang lain. Untuk nilai dari tiap cluster tersebut memiliki rentang -1 hingga 1. Makin tinggi nilai *SilhouetteScore* artinya nilai tersebut makin sesuai.

Selanjutnya, agar lebih memudahkan, dilakukan visualisasi dalam bentuk diagram batang dengan tujuan agar lebih mudah dalam penilaian cluster. Pada diagram batang yang berisi nilai *SilhouetteScore*, klaster dapat dibedakan antara yang memiliki nilai tinggi atau rendah. Jika ditemukan nilai rata-rata yang tinggi untuk semua klaster, menunjukkan bahwa klaster berkualitas baik. Namun, jika terdapat banyak klaster dengan skor siluet rata-rata yang rendah, ini mungkin mengindikasikan bahwa klaster tersebut kurang terpisah dari klaster lain atau mengandung kata yang tidak cocok dengan kata lain dalam klaster tersebut.

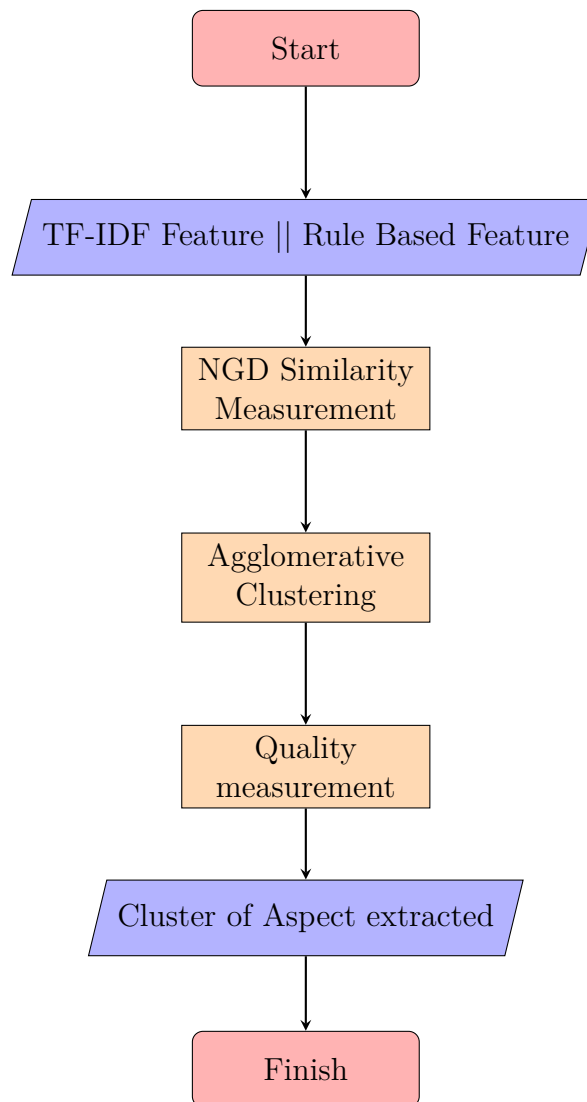
Dari penjelasan tersebut, dengan menggunakan analisis siluet. Peneliti berharap dapat memperoleh wawasan tentang kualitas pengelompokan kesamaan kata dan mengidentifikasi area yang dapat diperbaiki untuk mencapai hasil pengelompokan yang lebih baik.

BAB IV

CLUSTERING SIMILARITY MEASUREMENT DICTIONARY BASED

4.1 Desain Sistem

Pada tahapan ini telah diketahui *feature* yang telah didapatkan dari hasil *pre-processing* dan *feature extraction*. *feature* tersebut selanjutnya akan dijadikan input pada proses ini. Proses desain sistem percobaan pengukuran *clustering* berdasarkan *similarity measurement* dapat dilihat pada gambar 4.1.



Gambar 4.1. Desain sistem cluster NGD

4.2 Feature extraction

Untuk bentuk data awal, berbentuk csv yang telah kami jelaskan pada bab 3. Pada bagian ini, kami telah melakukan ekstraksi teks, dan melakukan pemrosesan awal mencakup pembersihan karakter tidak penting, dan pemrosesan untuk menjadi kata dasar. Langkah selanjutnya merupakan ekstraksi fitur dari data teks. Tahapan ini kami menggunakan dua metode yakni metode *TF-IDF* dan *Rule based*.

4.2.1 TF-IDF

Pada *TF-IDF*, untuk melakukan ekstraksi fitur, kami melakukan perhitungan pada *Term Frequency* (TF). TF melakukan pengukuran frekuensi setiap kata dalam dokumen. Untuk menghitung TF, kami menghitung kemunculan setiap token dalam dokumen dan membaginya dengan jumlah total token. Nilai hasil dari TF menggambarkan pengaruh relevansi berdasarkan dengan kemunculan kata dalam dokumen. Untuk perhitungan dapat dilihat pada Tabel 5.1

Tabel 4.1. Perhitungan *Term Frequency*

	barang	cepat	sesuai	bagus	kirim
Document 1	0.000000	0.000000	0.000000	0.000000	0.000000
Document 2	0.000000	0.371309	0.000000	0.000000	0.507896
Document 3	0.000000	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮
Document 1007	0.000000	0.283894	0.372946	0.000000	0.000000
Document 1008	0.000000	0.000000	0.000000	0.000000	0.000000
Document 1009	0.000000	0.000000	0.000000	0.292219	0.000000

Selanjutnya kami melakukan kalkulasi pada *Inverse Document Frequency* (IDF). IDF mengukur banyaknya tingkat kemunculan kata di seluruh korpus. Perhitungan IDF menggunakan pengambilan nilai logaritma dari rasio jumlah total dokumen dengan jumlah dokumen yang mengandung kata terpilih. Untuk perhitungan dapat dilihat pada Tabel 5.2

Selanjutnya, kami melakukan kalkulasi dari nilai dari tiap kata di seluruh dokumen menggunakan persamaan 3-1. Hasilnya dapat diketahui pada Tabel 5.3

Tabel 4.2. Perhitungan *Inverse Document Frequency*

barang	cepat	sesuai	bagus	kirim
2.713699	1.746005	2.293688	2.639591	2.388277

Tabel 4.3. Perhitungan *Inverse Document Frequency*

barang	cepat	sesuai	bagus	kirim
0.341	0.426	0.394	0.494	0.437

Pada tahapan ini, fitur dari kata pada kalimat telah didapatkan. Untuk mengurangi bias, kami mengambil maksimal nilai kemunculan DF hanya 95%. Artinya kata yang muncul pada seluruh dokumen, tidak kami ambil. Nantinya kata ini, akan kami ukur kesamaannya terhadap kata lain pada tahapan *Similarity measurement*

4.2.2 Rule based

Pada proses *feature extraction* dengan *rule based* penelitian ini menggunakan pos tagger. Pos tagger yang dimaksud menggunakan algoritma *Long Short-Term Memory* (LSTM). Model untuk ekstraksi tersebut berdasarkan pada pre trained data yang berasal dari dataset bahasa indonesia pada Universal tree bank¹.

Terdapat perbedaan proses ekstraksi pada TF-IDF dibandingkan dengan proses ekstraksi pada *rule based*. Pada *rule based*, setelah *cleansing data* akan dilanjutkan pada metode pos tagger. Pada *pos tagger* dapat diketahui kedudukan kata pada tiap kalimat. Pada penelitian ini, kedudukan kalimat yang dijadikan sebagai fitur terpilih adalah *nsubj* dan *obj*.

Pada proses selanjutnya, selanjutnya kami akan melakukan *Stemming*. Proses ini, merupakan proses perubahan kata dari kalimat menjadi kata dasar. Sehingga, imbuhan akan dihilangkan. Dan bersamaan dengan proses ini, juga dilakukan proses untuk melakukan pengecekan ada atau tidaknya kata dalam kamus bahasa indonesia. Untuk kata yang tidak baku, atau tidak tersedia dalam bahasa indonesia kami menandainya sebagai *Out of Vocabulary* (OOV). Sehingga, kata ini nantinya tidak akan digunakan dalam proses pengukuran kesamaan.

¹https://universaldependencies.org/treebanks/id_gsd/index.html

4.3 Similarity measurement

Dalam Pengukuran *Similarity measurement*, salah satu metode dalam penelitian ini adalah pengukuran menggunakan WordNet. WordNet merupakan salah satu basis data leksikal yang memetakan kesamaan kata berdasarkan hubungan semantik dalam kata. Database tersebut digunakan untuk melakukan pengukuran kesamaan kata berdasarkan fitur yang dihasilkan dalam proses sebelumnya. Pada penelitian asal, WordNet berawal dari bahasa inggris. Namun, selanjutnya terdapat penelitian yang dilakukan oleh Noor(Noor et al., 2011) untuk membuat dataset WordNet versi bahasa indonesia. Sehingga, penelitian ini pun menggunakan dataset tersebut untuk melakukan pengukuran pada kata.

Dalam melakukan ekstraksi tersebut, kami menggunakan metrik *Path Similarity*. Metrik *Path similarity* melakukan pengukuran kesamaan 2 kata, berdasarkan pendeknya *graph* yang menghubungkan antara 1 kata dengan kata lainnya. Dimana semakin pendek graph tersebut, artinya kata dianggap semakin sama. Sebaliknya, semakin panjang dari *graph*, maka dianggap semakin tidak sama.

```

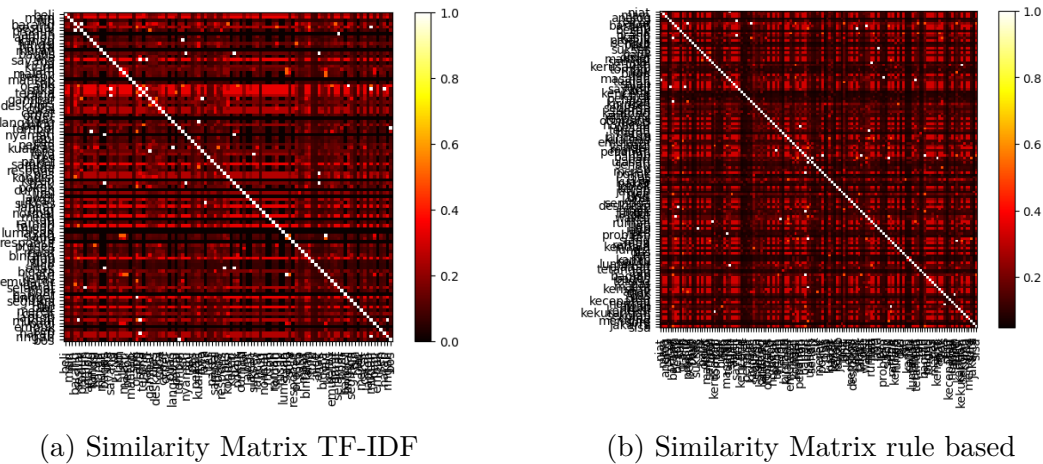
similarity_matrix = np.zeros((len(words),
    ↪ len(words)))
for i, word1 in enumerate(words):
for j, word2 in enumerate(words):
synsets1 = wn.synsets(word1, lang="ind")
synsets2 = wn.synsets(word2, lang="ind")
if len(synsets1) == 0 or len(synsets2) == 0:
similarity_matrix[i, j] = 0.0
else:
max_similarity = 0.0
for synset1 in synsets1:
for synset2 in synsets2:
similarity = synset1.path_similarity(synset2)
if similarity is not None and similarity >
    ↪ max_similarity:
max_similarity = similarity
similarity_matrix[i, j] = max_similarity

```

Gambar 4.2. Python similarity measurement using WordNet bahasa

Dalam penerapan di laboratorium, penelitian ini menggunakan

python3 dan nltk² sebagai metode untuk melakukan pengukuran *similarity measurement*. Untuk penerapannya dapat dilihat pada kode di 4.2. Pada kode tersebut, hasil dari similarity matrix ditampung dalam 1 variable yang berisi matrix dari nilai kesamaan kata. Untuk visualisasi bisa dilihat pada gambar 4.3



Gambar 4.3. Similarity Matrix

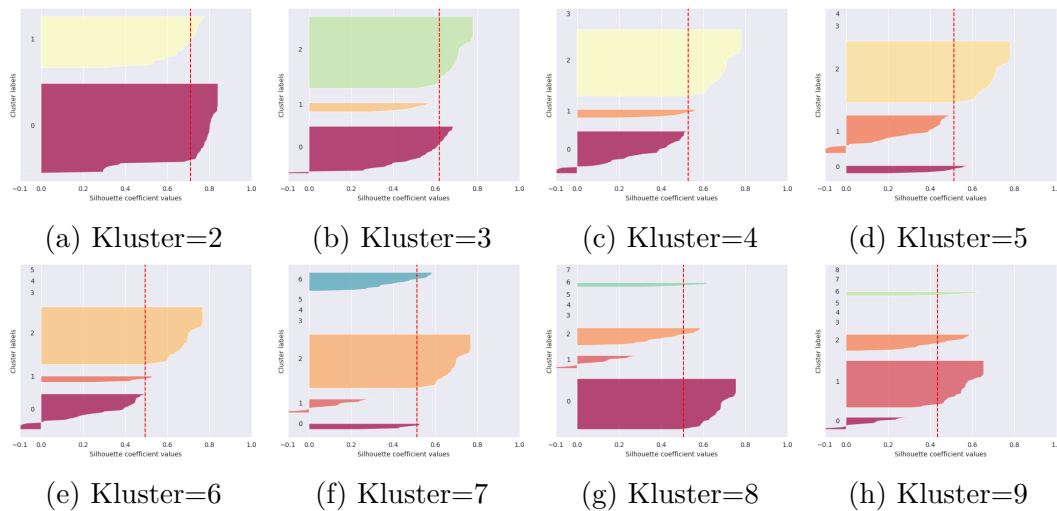
4.4 Quality Measurement

Untuk mengetahui aspek yang saling berkaitan satu sama lain, selanjutnya dilakukan proses clustering. Dalam tahapan ini *Agglomerative clustering* digunakan sebagai algoritma clustering. Hal ini sesuai dengan penelitian sebelumnya pada (Naeem et al., 2019) yang memberikan nilai cukup baik pada performa klasifikasi dokumen. Pada algoritma clustering tersebut, salah satu kriteria untuk menentukan pengelompokan akan terus dilakukan atau tidak adalah dengan pendefinisian dari total cluster yang diperlukan. Untuk menemukan nilai optimal dari total cluster yang diperlukan, penelitian ini menggunakan metrik *silhouette score*. Metrik tersebut melakukan pengukuran kohesi tiap point dalam cluster dan adhesi antara satu cluster dengan cluster lain.

Setelah proses untuk melakukan ekstraksi fitur dengan TF-IDF dan Rule based selesai. Hasil yang didapatkan berupa fitur serta nilai *similarity* dengan fitur lainnya. Selanjutnya akan dilakukan kalkulasi *silhouette score* untuk menentukan jumlah cluster yang paling tepat.

²<https://www.nltk.org/>

Clustering tiap fitur berdasarkan skor kesamaan dengan perbedaan jumlah cluster merupakan langkah awal untuk mendapatkan *silhouette score*. Pada setiap iterasi *clustering*, dilakukan perhitungan *silhouette score*. Semakin tinggi nilai tersebut, berarti bentuk clustering tercipta dengan lebih stabil. Tahapan pertama kami melakukan perhitungan pada *silhouette score* menggunakan fitur hasil ekstraksi dengan TF-IDF.

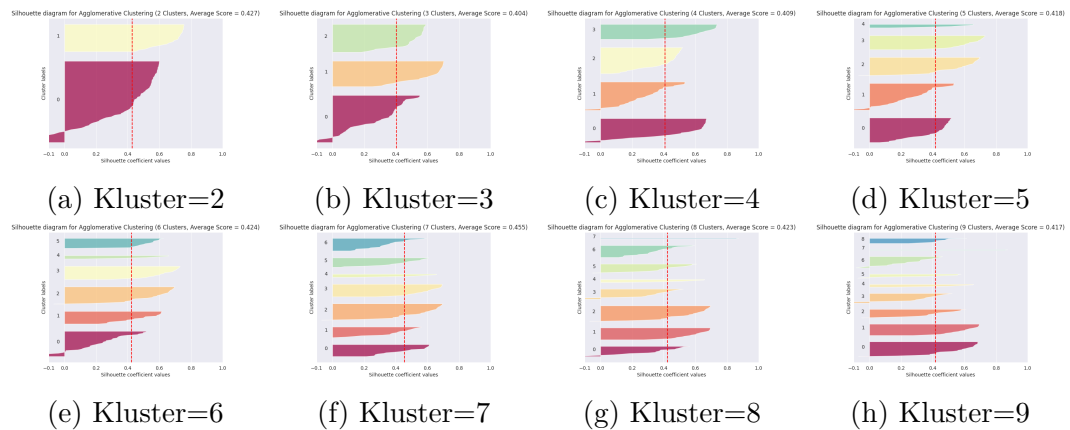


Gambar 4.4. Kalkulasi Silhouette score TF-IDF Dictionary Based

Tabel 4.4. Perhitungan *Silhouette score*

Jumlah kluster	Silhouette score
2	0.710000
3	0.619000
4	0.529000
5	0.513000
6	0.495000
7	0.513000
8	0.507000
9	0.436000

Pada Tabel 4.4 dapat diketahui bahwa, jumlah klustering terbaik untuk fitur kata menggunakan TF-IDF berdasarkan similarity measurement adalah 2 cluster. Dengan nilai Silhouette score mencapai 0.71. Selanjutnya kami melakukan perhitungan *silhouette score* pada fitur yang dihasilkan oleh *rule based*.



Gambar 4.5. Kalkulasi Silhouette score Rule Dictionary Based

Tabel 4.5. Perhitungan *Silhouette score*

Jumlah kluster	Silhouette score
2	0.427000
3	0.404000
4	0.409000
5	0.418000
6	0.424000
7	0.455000
8	0.423000
9	0.417000

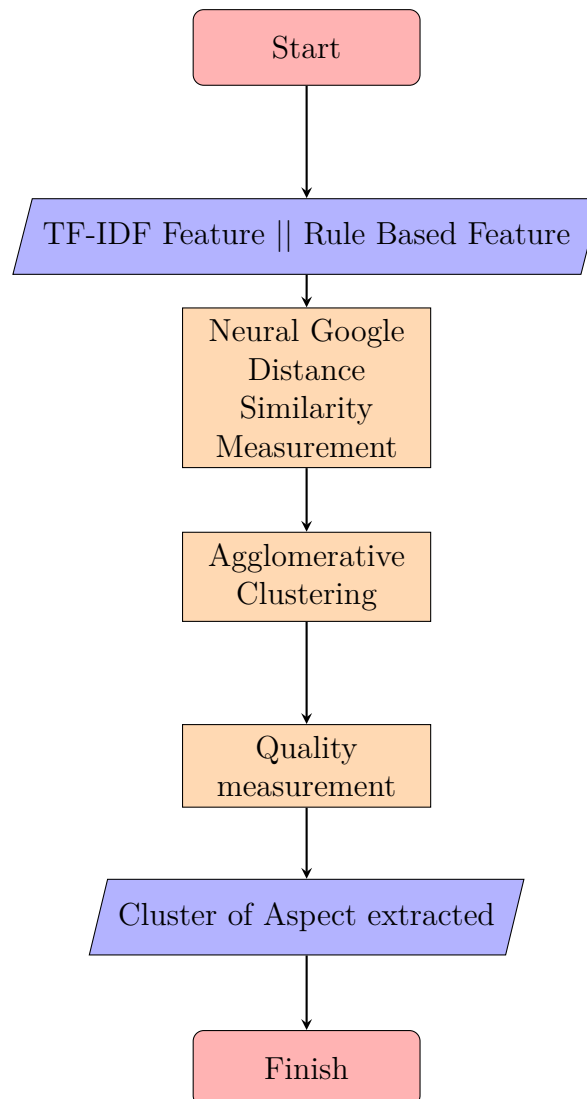
Pada Tabel 4.5 dapat diketahui bahwa, jumlah klustering terbaik untuk fitur kata menggunakan *rule based* berdasarkan similarity measurement adalah 7 cluster. Dengan nilai Silhouette score mencapai 0.455.

BAB V

CLUSTERING SIMILARITY MEASUREMENT NGD

5.1 Desain Sistem

Pada tahapan ini telah diketahui *feature* yang telah didapatkan dari hasil *pre-processing* dan *feature extraction*. *feature* tersebut selanjutnya akan dijadikan input pada proses ini. Proses desain sistem percobaan pengukuran *clustering* berdasarkan *similarity measurement* dapat dilihat pada gambar ??.



Gambar 5.1. Desain sistem cluster Dictionary Based

5.2 Feature extraction

Untuk bentuk data awal, berbentuk csv yang telah kami jelaskan pada bab 3. Pada bagian ini, kami telah melakukan ekstraksi teks, dan melakukan pemrosesan awal mencakup pembersihan karakter tidak penting, dan pemrosesan untuk menjadi kata dasar. Langkah selanjutnya merupakan ekstraksi fitur dari data teks. Tahapan ini kami menggunakan dua metode yakni metode *TF-IDF* dan *Rule based*.

5.2.1 TF-IDF

Pada *TF-IDF*, untuk melakukan ekstraksi fitur, kami melakukan perhitungan pada *Term Frequency* (TF). TF melakukan pengukuran frekuensi setiap kata dalam dokumen. Untuk menghitung TF, kami menghitung kemunculan setiap token dalam dokumen dan membaginya dengan jumlah total token. Nilai hasil dari TF menggambarkan pengaruh relevansi berdasarkan dengan kemunculan kata dalam dokumen. Untuk perhitungan dapat dilihat pada tabel 5.1

Tabel 5.1. Perhitungan *Term Frequency*

	barang	cepat	sesuai	bagus	kirim
Document 1	0.000000	0.000000	0.000000	0.000000	0.000000
Document 2	0.000000	0.371309	0.000000	0.000000	0.507896
Document 3	0.000000	0.000000	0.000000	0.000000	0.000000
⋮	⋮	⋮	⋮	⋮	⋮
Document 1007	0.000000	0.283894	0.372946	0.000000	0.000000
Document 1008	0.000000	0.000000	0.000000	0.000000	0.000000
Document 1009	0.000000	0.000000	0.000000	0.292219	0.000000

Selanjutnya kami melakukan kalkulasi pada *Inverse Document Frequency* (IDF). IDF mengukur banyaknya tingkat kemunculan kata di seluruh korpus. Perhitungan IDF menggunakan pengambilan nilai logaritma dari rasio jumlah total dokumen dengan jumlah dokumen yang mengandung kata terpilih. Untuk perhitungan dapat dilihat pada Tabel 5.2

Selanjutnya, kami melakukan kalkulasi dari nilai dari tiap kata di seluruh dokumen menggunakan persamaan 3-1. Hasilnya dapat diketahui pada tabel 5.3

Tabel 5.2. Perhitungan *Inverse Document Frequency*

barang	cepat	sesuai	bagus	kirim
2.713699	1.746005	2.293688	2.639591	2.388277

Tabel 5.3. Perhitungan *Inverse Document Frequency*

barang	cepat	sesuai	bagus	kirim
0.341	0.426	0.394	0.494	0.437

Pada tahapan ini, fitur dari kata pada kalimat telah didapatkan. Untuk mengurangi bias, kami mengambil maksimal nilai kemunculan DF hanya 95%. Artinya kata yang muncul pada seluruh dokumen, tidak kami ambil. Nantinya kata ini, akan kami ukur kesamaannya terhadap kata lain pada tahapan *Similarity measurement*

5.2.2 Rule based

Pada proses *feature extraction* dengan *rule based* penelitian ini menggunakan pos tagger. Pos tagger yang dimaksud menggunakan algoritma *Long Short-Term Memory* (LSTM). Model untuk ekstraksi tersebut berdasarkan pada pre trained data yang berasal dari dataset bahasa indonesia pada Universal tree bank¹.

Terdapat perbedaan proses ekstraksi pada TF-IDF dibandingkan dengan proses ekstraksi pada *rule based*. Pada *rule based*, setelah *cleansing data* akan dilanjutkan pada metode pos tagger. Pada *pos tagger* dapat diketahui kedudukan kata pada tiap kalimat. Pada penelitian ini, kedudukan kalimat yang dijadikan sebagai fitur terpilih adalah *nsubj* dan *obj*.

Pada proses selanjutnya, selanjutnya kami akan melakukan *Stemming*. Proses ini, merupakan proses perubahan kata dari kalimat menjadi kata dasar. Sehingga, imbuhan akan dihilangkan. Dan bersamaan dengan proses ini, juga dilakukan proses untuk melakukan pengecekan ada atau tidaknya kata dalam kamus bahasa indonesia. Untuk kata yang tidak baku, atau tidak tersedia dalam bahasa indonesia kami menandainya sebagai *Out of Vocabulary* (OOV). Sehingga, kata ini nantinya tidak akan digunakan dalam proses pengukuran kesamaan.

¹https://universaldependencies.org/treebanks/id_gsd/index.html

5.3 Similarity measurement

Dalam Pengukuran *Similarity measurement*, salah satu metode dalam penelitian ini adalah pengukuran menggunakan NGD. Normalized Google Distance (NGD) adalah metrik yang digunakan untuk mengukur keterkaitan semantik antara kata atau konsep berdasarkan pola kemunculan sebuah kata secara beriringan dalam sebuah korpus. Korpus disini merupakan kumpulan data yang dimiliki oleh Google. Sehingga, secara natural metrik dalam NGD dihasilkan dari mesin pencari Google, yang berfungsi sebagai sumber data linguistik. Dimana nilai dihasilkan dari metrik ini mempertimbangkan keterkaitan input user dan frekuensi kemunculan penggunaan suatu kata pada internet yang *disrape* sebagai referensi.

Dalam penelitian kali ini, kami melakukan pengambilan data pada website google berbahasa indonesia dengan domain google.co.id . Sesuai dengan penjelasan pada persamaan 2-4, maka kami melakukan pengambilan data yang terdiri dari:

- Hasil pencarian kata dasar pertama dari hasil ekstraksi fitur.
- Hasil pencarian kata dasar kedua dari hasil ekstraksi fitur.
- Hasil pencarian kedua kata tersebut.

Proses ini kami lakukan manual dengan menggunakan *script* Python (5.2).

```

chrome_options = Options()
chrome_options.add_argument("--headless")
chrome_options.add_argument("--disable-
    ↪ extensions")
chrome_options.add_argument("--no-sandbox")
driver =
    ↪ webdriver.Chrome(options=chrome_options)
driver.get('https://www.google.co.id/')
search_box = driver.find_element(By.CLASS_NAME,
    ↪ 'gLFyf')
format_search = re.sub(r"\-", " ", word)
search_box.send_keys(format_search)
search_box.send_keys(Keys.ENTER)
total_count = driver.find_element(By.ID,
    ↪ 'result-stats')
pattern_result =
    ↪ re.compile(r"(1[\.,])?([0-9]{1,3}[\.,])?([0-
    ↪ 9]{1,3})([\.,][0-9]{1,3})")
result_stats =
    ↪ pattern_result.search(total_count.text)
string_number = re.sub(r"\.", "",
    ↪ result_stats.group(0).strip())
with
    ↪ open("./result/similarity_json/{}.json".format(word),
    ↪ 'w+') as f:
json.dump({
    'word': word,
    'number': string_number
}, f, indent=4)
driver.close()

```

Gambar 5.2. Python similarity measurement using WordNet bahasa

Hasil dari scrapping tersebut, selanjutnya akan dijadikan input pada persamaan 2-4. Sebagai ilustrasi, kami sertakan kalkulasi terhadap 6 data

yang ada. Data tersebut bisa dilihat pada Tabel 5.4

Tabel 5.4. Perhitungan *Silhouette score*

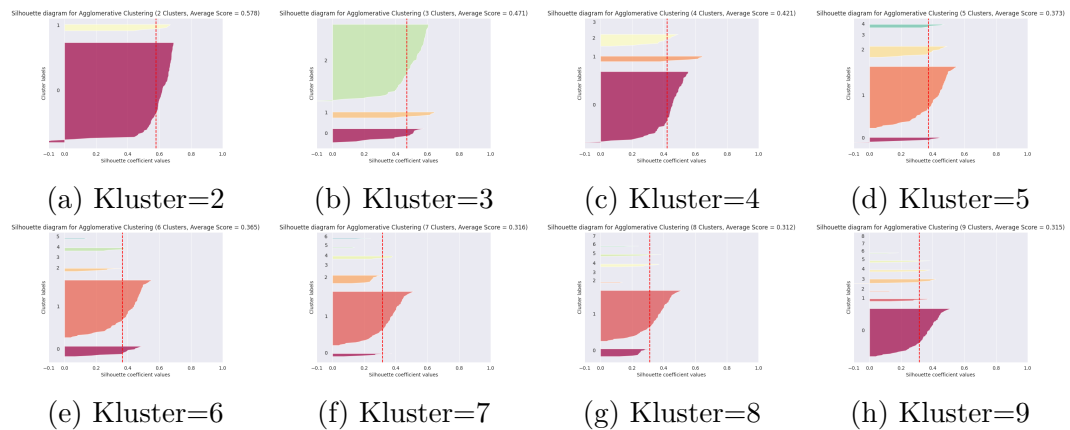
	beli	main	barang	sesuai	produk	aman
beli	0.000000	0.235292	0.345649	0.055684	0.171176	0.549360
main	0.155376	0.000000	0.251211	0.091465	0.220353	0.247286
barang	0.353742	0.251211	0.000000	0.184825	0.260589	0.434123
sesuai	0.055684	0.091465	0.184825	0.000000	0.208541	0.398956
produk	0.299384	0.220353	0.260589	0.208541	0.000000	0.497562
aman	0.549360	0.247286	0.434123	0.398956	0.497562	0.000000

5.4 Quality Measurement

Untuk mengetahui aspek yang saling berkaitan satu sama lain, selanjutnya dilakukan proses clustering. Dalam tahapan ini *Agglomerative clustering* digunakan sebagai algoritma clustering. Hal ini sesuai dengan penelitian sebelumnya pada (Naeem et al., 2019) yang memberikan nilai cukup baik pada performa klasifikasi dokumen. Pada algoritma clustering tersebut, salah satu kriteria untuk menentukan pengelompokan akan terus dilakukan atau tidak adalah dengan pendefinisian dari total cluster yang diperlukan. Untuk menemukan nilai optimal dari total kluster yang diperlukan, penelitian ini menggunakan metrik *silhouette score*. Metrik tersebut melakukan pengukuran kohesi tiap point dalam kluster dan adhesi antara satu kluster dengan kluster lain.

Setelah proses untuk melakukan ekstraksi fitur dengan TF-IDF dan Rule based selesai. Hasil yang didapatkan berupa fitur serta nilai *similarity* dengan fitur lainnya. Selanjutnya akan dilakukan kalkulasi *silhouette score* untuk menentukan jumlah kluster yang paling tepat.

Clustering tiap fitur berdasarkan skor kesamaan dengan perbedaan jumlah cluster merupakan langkah awal untuk mendapatkan *silhouette score*. Pada setiap iterasi *clustering*, dilakukan perhitungan *silhouette score*. Semakin tinggi nilai tersebut, berarti bentuk clustering tercipta dengan lebih stabil. Tahapan pertama kami melakukan perhitungan pada *silhouette score* menggunakan fitur hasil ekstraksi dengan TF-IDF.

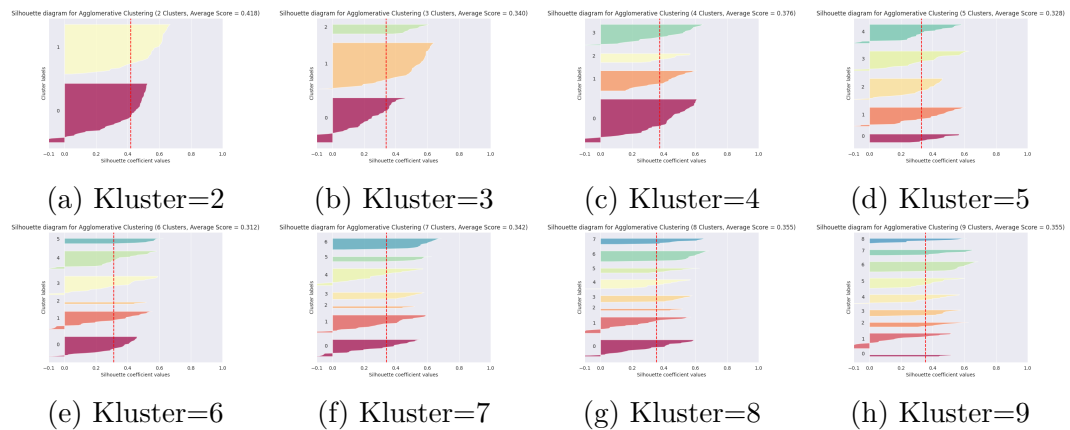


Gambar 5.3. Kalkulasi Silhouette score TF-IDF Menggunakan NGD

Tabel 5.5. Perhitungan *Silhouette score*

Jumlah kluster	Silhouette score
2	0.578000
3	0.471000
4	0.421000
5	0.373000
6	0.365000
7	0.316000
8	0.312000
9	0.315000

Pada Tabel 5.5 dapat diketahui bahwa, jumlah klustering terbaik untuk fitur kata menggunakan TF-IDF berdasarkan similarity measurement menggunakan *NGD* adalah 2 cluster. Dengan nilai Silhouette score mencapai 0.578. Selanjutnya kami melakukan perhitungan *silhouette score* pada fitur yang dihasilkan oleh *rule based*.



Gambar 5.4. Kalkulasi Silhouette score Rule NGD Based

Tabel 5.6. Perhitungan *Silhouette score*

Jumlah kluster	Silhouette score
2	0.418000
3	0.340000
4	0.376000
5	0.328000
6	0.312000
7	0.342000
8	0.355000
9	0.355000

Pada Tabel 5.6 dapat diketahui bahwa, jumlah klustering terbaik untuk fitur kata menggunakan *rule based* berdasarkan similarity measurement adalah 2 cluster. Dengan nilai Silhouette score mencapai 0.41.

BAB VI

PEMBAHASAN

6.1 Pembahasan

Pada bagian ini, dilakukan perbandingan hasil dari data antara *similarity measurement* menggunakan NGD dan Dictionary based. Proses analisa pada data ini menggunakan *silhouette analysis*, dimana acuan dari penilaian tersebut menggunakan *silhouette score*. Semakin tinggi nilai dari *silhouette score* maka, nilai hubungan tiap node dalam sebuah kluster akan semakin relevan.

Percobaan kali ini kami menganalisa dengan melakukan pengelompokan pada data (*clustering*) menggunakan *agglomerative clustering*. Pada *agglomerative clustering* ini kami melakukan pengukuran pada jarak antar titik dengan menggunakan *manhattan distance*. Kami melakukan analisa pada rentang cluster antara 1 hingga 9. Pada masing masing cluster, kami lakukan *silhouette analysis*.

Tabel 6.1. Hasil Sillhouette score *Dictionary based*

Kluster	SS TF-IDF DB	SS RB DB
2	0.710000	0.578000
3	0.619000	0.471000
4	0.529000	0.421000
5	0.513000	0.373000
6	0.495000	0.365000
7	0.513000	0.316000
8	0.507000	0.312000
9	0.436000	0.315000

Tabel 6.1 menunjukkan hasil dari *analysis silhouette* pada metode similarity menggunakan dictionary based. Pada tabel tersebut, kami menemukan bahwa. Jumlah tertinggi nilai *silhouette score* ada pada kluster dengan jumlah 2. Nilai tertinggi ini ditunjukkan oleh kedua model feature

extraction yang kami gunakan, yakni pada TF-IDF *based* maupun pada *Rule based*. Pada *Dictionary based* nilai tertinggi adalah 0,71. Dan Nilai tertinggi pada ekstraksi menggunakan *rule based*, adalah 0,57.

Selanjutnya, kami menggunakan yang sama untuk melakukan analisis pada *similarity measurement* dengan menggunakan NGD. Yakni dengan rentang jumlah kluster antara 1 hingga 9 kluster. Hasil percobaan tersebut, dapat dilihat pada tabel 6.2.

Tabel 6.2. Hasil Sillhouette score *NGD*

Kluster	SS TF-IDF NGD	SS RB NGD
2	0.427000	0.418000
3	0.404000	0.340000
4	0.409000	0.376000
5	0.418000	0.328000
6	0.424000	0.312000
7	0.455000	0.342000
8	0.423000	0.355000
9	0.417000	0.355000

Berbeda dengan analisa pada model *dictionary based*, yang mana hasil tertinggi memiliki jumlah kluster yang sama. Pada pengukuran *similarity* dengan NGD kluster terbaik memiliki jumlah yang beragam. Pada *feature extraction* dengan TF-IDF kluster terbaik memiliki jumlah kluster 7, dengan nilai 0,45. Sedangkan pada penggunaan *rule based* kluster dengan nilai tertinggi tetap pada kluster dengan jumlah 2 dengan nilai 0,41.

Dari hasil pengamatan pada 2 metode *similarity measurement*, yakni *dictionary based* dan *rule based*. Selanjutnya, dilakukan perbandingan pada hasil yang didapat. Perbandingan tersebut dapat ditemukan pada Tabel 6.3.

Tabel 6.3. Perbandingan antara Sillhouette score *Dictionary based* dan *NGD*

Kluster	SS TF-IDF DB	SS RB DB	SS TF-IDF NGD	SS RB NGD
2	0.710000	0.578000	0.427000	0.418000
3	0.619000	0.471000	0.404000	0.340000
4	0.529000	0.421000	0.409000	0.376000
5	0.513000	0.373000	0.418000	0.328000
6	0.495000	0.365000	0.424000	0.312000
7	0.513000	0.316000	0.455000	0.342000
8	0.507000	0.312000	0.423000	0.355000
9	0.436000	0.315000	0.417000	0.355000

Pada Tabel 6.3 dapat diketahui bahwa jumlah kluster tidak berpengaruh terhadap nilai sillhouette score. Sebagai contoh ada pada sillhouette score tertinggi pada *rule based* menggunakan NGD yang memiliki nilai 0.45. Begitu juga pada kluster 7 pada TF-IDF *dictionary based* mengungguli kluster dengan jumlah 6.

Selain itu pada rentang percobaan NGD nilai yang memiliki rentang yang lebih pendek, yakni tertinggi 0,45 dan terendah 0,31. Berbeda dengan rentang pada *dictionary based*, yang memiliki rentang 0,71 hingga 0,31.

BAB VII

KESIMPULAN

7.1 Kesimpulan

Berdasarkan penelitian tentang ekstraksi aspek eksplisit pada komentar produk di toko elektronik ini, menggunakan TF-IDF dan *rule based* sebagai metode untuk ekstraksi fitur. Sedangkan, untuk similarity measurement menggunakan NGD dan *dictionary based*. Dalam Penelitian ini digunakan 1200 kalimat review yang berasal dari internet. Pada *raw data* tersebut, selanjutnya kami melakukan tahapan *Preprocessing*. Pada tahapan ini, kami melakukan beberapa tahapan mencakup *cleaning*, *stopword removal*, dan normalisasi data. *Feature Extraction* merupakan tahapan selanjutnya dalam penelitian ini, dibagian ini kami menggunakan TF-IDF dan *Rule Based*. Hasil dari *Feature extraction* adalah kata yang dianggap sesuai, kemudian dilakukan kalkulasi kesamaan katanya pada tahapan *Similarity Measurement*. Selanjutnya untuk analisa digunakan *silhouette score analysis* untuk menilah jumlah kluster terbaik.

Pada penggunaan *Similarity Measurement* menggunakan *dictionary based*. Kami menemukan bahwa, Jumlah tertinggi nilai *silhouette score* ada pada kluster dengan jumlah 2. Nilai tertinggi ini ditunjukkan oleh kedua model *feature Extraction* yang kami gunakan, yakni pada TF-IDF *based* maupun pada *Rule based*. Pada *Dictionary based* nilai tertinggi adalah 0,71. Dan Nilai tertinggi pada ekstraksi menggunakan *rule based*, adalah 0,57.

Berbeda dengan analisa pada model *dictionary based*, yang mana hasil tertinggi memiliki jumlah kluster yang sama. Pada pengukuran *similarity* dengan NGD, kluster terbaik memiliki jumlah yang beragam. Pada *feature extraction* dengan TF-IDF kluster terbaik memiliki jumlah kluster 7, dengan nilai 0,45. Sedangkan pada penggunaan *rule based* kluster dengan nilai tertinggi tetap pada kluster dengan jumlah 2 dengan nilai 0,41.

Dari hal tersebut, kami dapat menyimpulkan bahwa jika mengacu pada *Silhouette score*, maka ekstraksi fitur dengan TF-IDF memiliki hasil yang lebih baik dibandingkan dengan *rule based*. Dimana nilai *silhouette score* mencapai 0,71 dibandingkan dengan *Rule based* yang hanya mencapai 0,5.

7.2 Saran

Penelitian ini dapat menjadi salah satu landasan awal untuk melakukan analisa dari pengaruh komentar pada produk di toko online. Pengembangan selanjutnya yang dapat dilakukan adalah melakukan rangkingan kata dari seperti menggunakan *Textrank*, maupun metode *tree*. Selain itu, dengan meningkatnya perkembangan teknologi penggunaan metode atau algoritma lain dapat digunakan untuk melakukan penelitian terhadap kesamaan kata.

DAFTAR PUSTAKA

- Al-Bukhari, M., et al. (1978). Sahih al-bukhari.
- Asgari, O., Weise, A., Dubard Barbosa, S., & Martinez, L. F. (2022). The Effect of Electronic Word-of-Mouth (eWOM) on Consumer Ratings in the Digital Era. In *Advances in Digital Marketing and eCommerce* (pp. 267–273). Springer.
- Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, 52, 201–213.
- Bancken, W., Alfarone, D., & Davis, J. (2014). Automatically detecting and rating product aspects from textual customer reviews. *Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD, 1202*, 1–16.
- Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, 10(5), 2780–2788.
- Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3), 370–383.
- Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, 78(4), 1175–1184.
- Fransiska, S., Rianto, R., & Gufroni, A. I. (2020). Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. *Scientific Journal of Informatics*, 7(2), 203–212.
- Gerdt, S.-O., Wagner, E., & Schewe, G. (2019). The relationship between sustainability and customer satisfaction in hospitality: An explorative investigation using eWOM as a data source. *Tourism Management*, 74, 155–172.
- Hazarika, B., Chen, K., & Razi, M. (2021). Are numeric ratings true representations of reviews? A study of inconsistency between reviews and ratings. *International Journal of Business Information Systems*, 38(1), 85–106.
- Hidayanto, A. N., Ovirza, M., Anggia, P., Budi, N. F. A., & Phusavat, K. (2017). The roles of electronic word of mouth and information

- searching in the promotion of a new e-commerce strategy: A case of online group buying in Indonesia. *Journal of theoretical and applied electronic commerce research*, 12(3), 69–85.
- Horn, F., Arras, L., Montavon, G., Müller, K.-R., & Samek, W. (2017). Exploring text datasets by visualizing relevant words. *arXiv preprint arXiv:1707.05261*.
- Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. *AAAI*, 4(4), 755–760.
- Indonesia, T. R. K. B. (2008). Kamus Bahasa Indonesia. *Jakarta: Pusat Bahasa Departemen Pendidikan Nasional*, 725.
- Iqbal, F., Fung, B. C., Debbabi, M., Batool, R., & Marrington, A. (2019). Wordnet-based criminal networks mining for cybercrime investigation. *Ieee Access*, 7, 22740–22755.
- Karve, S., Shende, V., & Hople, S. (2019). Semantic relatedness measurement from Wikipedia and WordNet using modified normalized google distance. *Data Analytics and Learning: Proceedings of DAL 2018*, 143–154.
- Kobayashi, N., Inui, K., & Matsumoto, Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 1065–1074.
- Lee, S., Lee, S., & Baek, H. (2021). Does the dispersion of online review ratings affect review helpfulness? *Computers in Human Behavior*, 117, 106670.
- Li, H., Mao, H., & Wang, J. (2021). Part-of-speech tagging with rule-based data preprocessing and transformer. *Electronics*, 11(1), 56.
- Lin, Y.-S., Jiang, J.-Y., & Lee, S.-J. (2013). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, 26(7), 1575–1590.
- Liu, K., Xu, L., & Zhao, J. (2014). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on knowledge and data engineering*, 27(3), 636–650.
- Luo, Z., Huang, S., & Zhu, K. Q. (2019). Knowledge empowered prominent aspect extraction from product reviews. *Information Processing & Management*, 56(3), 408–423.

- Mao, X., Huang, S., Li, R., & Shen, L. (2020). Automatic keywords extraction based on co-occurrence and semantic relationships between words. *IEEE Access*, 8, 117528–117538.
- Marcińczuk, M., Gniewkowski, M., Walkowiak, T., & Będkowski, M. (2021). Text document clustering: Wordnet vs. TF-IDF vs. word embeddings. *Proceedings of the 11th Global Wordnet Conference*, 207–214.
- Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, 41(17), 7764–7775.
- Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.
- Naeem, A., Rehman, M., Anjum, M., & Asif, M. (2019). Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm. *Current Science*, 117(6), 1045–1053.
- Noor, N. H. B. M., Sapuan, S., & Bond, F. (2011). Creating the open wordnet bahasa. *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, 255–264.
- Orkphol, K., & Yang, W. (2019). Word sense disambiguation using cosine similarity collaborates with Word2vec and WordNet. *Future Internet*, 11(5), 114.
- Parolin, C. F., & Boeing, R. (2019). Consumption of experiences in boutique hotels in the context of e-WOM. *Tourism & Management Studies*, 15(2).
- Popovski, G., Kochev, S., Korousic-Seljak, B., & Eftimov, T. (2019). FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction. *ICPRAM*, 12, pp–915.
- Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, 28–37.
- Rana, T. A., & Cheah, Y.-N. (2016). Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, 46(4), 459–483.
- Rana, T. A., & Cheah, Y.-N. (2017). A two-fold rule-based model for aspect extraction. *Expert systems with applications*, 89, 273–285.

- Rashel, F., Luthfi, A., Dinakaramani, A., & Manurung, R. (2014). Building an Indonesian rule-based part-of-speech tagger. *2014 International Conference on Asian Language Processing (IALP)*, 70–73. <https://doi.org/10.1109/IALP.2014.6973521>
- Ray, P., & Chakrabarti, A. (2022). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. *Applied Computing and Informatics*, 18(1/2), 163–178.
- Rushd, I. (n.d.). Bid? yatu al-Mujtahid wa Nih? yatu al-Muqtashid. *Beirut: D.*
- Saumya, S., Singh, J. P., & Kumar, A. (2021). A Machine Learning Model for Review Rating Inconsistency in E-commerce Websites. In *Data Management, Analytics and Innovation* (pp. 221–230). Springer.
- Shihab, M. Q. (2009). *Tafsir Al-Mishbah (Vol.2)*.
- Tala, F. (2003). A study of stemming effects on information retrieval in Bahasa Indonesia.
- Wei, C.-P., Chen, Y.-M., Yang, C.-S., & Yang, C. C. (2010). Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management*, 8(2), 149–167.
- Wibowo, H. A., Prawiro, T. A., Ihsan, M., Aji, A. F., Prasajo, R. E., Mahendra, R., & Fitriany, S. (2020). Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation. *2020 International Conference on Asian Language Processing (IALP)*, 310–315.
- Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 481–492.