

**PERINGKASAN TEKS MULTI DOKUMEN BERITA BERBAHASA
INDONESIA MENGGUNAKAN FASTTEXT DAN
K-MEANS CLUSTERING**

SKRIPSI

Oleh:
THORIQ HARIZUL AHSAN
NIM. 19650026



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

**PERINGKASAN TEKS MULTI DOKUMEN BERITA BERBAHASA
INDONESIA MENGGUNAKAN FASTTEXT DAN
K-MEANS CLUSTERING**

SKRIPSI

Oleh :
THORIQ HARIZUL AHSAN
NIM. 19650026

Diajukan Kepada :
Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Malang
Untuk Memenuhi Salah Satu Persyaratan Dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

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

HALAMAN PERSETUJUAN

**PERINGKASAN TEKS MULTI DOKUMEN BERITA BERBAHASA
INDONESIA MENGGUNAKAN FASTTEXT DAN KMEANS
CLUSTERING**

SKRIPSI

Oleh :
THORIO HARIZUL AHSAN
NIM. 19650026

Telah Diperiksa dan Disetujui untuk Diuji
Tanggal 7 Juni 2023

Pembimbing I



Zainal Abidin, M.Kom
NIP. 19760613 200501 1 004

Pembimbing II



Dr. M. Imamudin Lc., MA
NIP. 19740602 200901 1 010

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang




Dr. Fachrul Kurniawan, ST., M. MT., IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

**PERINGKASAN TEKS MULTI DOKUMEN BERITA BERBAHASA
INDONESIA MENGGUNAKAN FASTTEXT DAN KMEANS
CLUSTERING**

SKRIPSI

Oleh :
THORIQ HARIZUL AH SAN
NIM. 19650026

Telah Dipertahankan di Depan Dewan Penguji
Dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Pada Tanggal : 22 Juni 2023

Susunan Dewan Penguji

Ketua Penguji	:	<u>Dr. Cahyo Crysdian</u> NIP. 19740424 200901 1 008	()
Anggota Penguji I:		<u>Dr. Yunifa Miftachul Arif, M.T</u> NIP. 19830616 201101 1 004	()
Anggota Penguji II	:	<u>Zainal Abidin, M.Kom</u> NIP. 19760613 200501 1 004	()
Anggota Penguji III	:	<u>Dr. M. Imamudin Lc, MA</u> NIP. 19740602 200901 1 010	()

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




Dr. Fachrul Kurniawan, ST., M. MT., IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN PENULISAN

Saya yang bertanda tangan dibawah ini:

Nama : Thoriq Harizul Ahsan
NIM : 19650026
Fakultas : Sains dan Teknologi
Program Studi : Teknik Informatika
Judul Skripsi : Peringkasan Teks Multi Dokumen Berita Berbahasa
Indonesia Menggunakan *FastText* dan *K-means Clustering*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambilalihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila di kemudian hari terbukti atau dapat dibuktikan Skripsi ini hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 13 Juni 2023
Yang membuat pernyataan,



Thoriq Harizul Ahsan
NIM. 19650026

HALAMAN MOTTO

*“Jangan Terbuai oleh kata-kata, orang bisa saja gagal atas perjuangannya.
Kalaupun ketakutanku gagal aku akan tetap berjuang, karena hanya dan satu
satunya cara meraih kesuksesan adalah berjuang”*

HALAMAN PERSEMBAHAN

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Penulis ingin mempersembahkan karya ilmiah ini kepada orang tua, keluarga, dosen, sahabat, dan semua pihak yang telah membantu secara aktif dalam menyelesaikan penelitian ini

KATA PENGANTAR

Assalamu'alaikum Wr.Wb.

Puji syukur penulis panjatkan kepada Allah SWT yang senantiasa memberikan rahmat dan kesehatan, sehingga penulis mampu menyelesaikan skripsi ini dengan baik. Penulis menyampaikan ucapan Terimakasih kepada semua pihak yang pernah terlibat langsung maupun tidak langsung dalam menyelesaikan skripsi ini, bukan hanya karena usaha keras dari penulis sendiri, akan tetapi karena adanya dukungan dari berbagai pihak. Oleh karena itu penulis berterima kasih kepada:

1. Prof. Dr. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Fachrul Kurniawan M.MT., IPM selaku Ketua Prodi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Zainal Abidin M.Kom selaku Dosen Pembimbing 1 yang telah membimbing serta memberikan arahan serta motivasi dalam penulisan skripsi dari awal hingga akhir.
5. Dr. M. Imamudin Lc, MA selaku Dosen Pembimbing 2 yang telah memberikan bimbingan, arahan serta bantuan dalam terwujudnya karya tulis skripsi ini dari awal hingga akhir.

6. Dr. Cahyo Crysdian, MCS selaku penguji I dan Dr. Yunifa Miftachul Arif, M.T selaku penguji II yang telah meluangkan waktunya untuk menguji dan dengan sabar memberi arahan dan saran dalam menyelesaikan skripsi ini.
7. Segenap civitas akademik Program Studi Teknik Informatika, dan seluruh dosen yang telah memberikan ilmu serta arahan semasa kuliah.
8. Kedua orang tua penulis, Bapak Syamsudin dan Ibu Rubiani yang telah memberikan banyak dukungan, doa serta bantuan sehingga penulis mampu menyelesaikan masa studi hingga mencapai gelar sarjana.
9. Teman-teman yang telah memberikan bantuan yang sangat besar dalam pengumpulan dataset penelitian saya, serta yang menjadi teman diskusi dan teman belajar.

Penulis sadar bahwa skripsi ini masih sangat jauh dari kata sempurna dan mungkin terdapat kesalahan di dalamnya. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun untuk mengembangkan skripsi ini agar lebih bermanfaat bagi dirinya dan pembaca pada umumnya.

Malang, 26 Juni 2023

Penulis

DAFTAR ISI

HALAMAN JUDUL	ii
HALAMAN PERSETUJUAN.....	Error! Bookmark not defined.
HALAMAN PENGESAHAN	iii
PERNYATAAN KEASLIAN PENULISAN.....	v
HALAMAN MOTTO	vi
HALAMAN PERSEMBAHAN.....	vii
KATA PENGANTAR.....	viii
DAFTAR ISI.....	x
DAFTAR GAMBAR.....	xi
DAFTAR TABEL.....	xii
ABSTRAK	xiii
ABSTRACT	xiv
المخلص	xv
BAB I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Pernyataan Masalah	5
1.3 Tujuan Penelitian	5
1.4 Batasan Masalah.....	6
1.5 Manfaat Penelitian	6
BAB II KAJIAN PUSTAKA	7
2.1 Peringkasan Teks Berita.....	7
2.2 <i>FastText</i>	8
2.3 <i>K-means Clustering</i>	14
BAB III DESAIN DAN IMPLEMENTASI	17
3.1 Data Penelitian	17
3.2 Desain Sistem.....	20
3.6 Implementasi Sistem	38
BAB IV UJI COBA DAN PEMBAHASAN.....	42
4.1 Skenario Uji Coba.....	42
4.1.1 <i>Recall</i>	43
4.1.2 <i>Precision</i>	44
4.1.3 <i>F-Score</i>	44
4.2 Hasil Uji Coba.....	44
4.3 Pembahasan.....	50
BAB V PENUTUP.....	57
5.1 Kesimpulan	57
5.2 Saran.....	58
DAFTAR PUSTAKA.	
LAMPIRAN.	

DAFTAR GAMBAR

Gambar 2. 1 Arsitektur Skipgram	11
Gambar 2. 2 Arsitektur CBOW.....	13
Gambar 2. 3 Hashing n-gram	14
Gambar 3. 1 Struktur IndoSum Dataset	17
Gambar 3. 2 <i>Dataset Indosum</i> pada <i>fold 1</i>	18
Gambar 3. 3 Dataset Multi dokumen	18
Gambar 3. 4 Hasil pembentukan label dari 30 multi dokumen.....	20
Gambar 3. 5 Desain Sistem.....	20
Gambar 3. 6 Flowchart <i>Preprocessing</i>	22
Gambar 3. 7 Flowchart <i>K-means Clustering</i>	27
Gambar 3. 8 Flowchart training model <i>FastText</i>	31
Gambar 3. 10 Implementasi Sistem	38
Gambar 3. 11 Tampilan Sistem Peringkasan Multi dokumen	39

DAFTAR TABEL

Tabel 3. 1 Hasil FastText untuk Vektor Kalimat dalam 3 dokumen	40
Tabel 4. 2 Hasil Uji Coba.....	45
Tabel 4. 3 Statistik artikel multi dokumen	48
Tabel 4. 4 Daftar kata hasil ringkasan sistem	48
Tabel 4. 5 Daftar kata hasil ringkasan manual.....	48
Tabel 4. 6 Daftar kata beririsan.....	49
Tabel 4. 7 Statistika jumlah kata ringkasan	49
Tabel 4. 8 Rekap hasil perhitungan.....	50
Tabel 4. 9 Rata-rata Skor Rouge-1 pada 30 multi dokumen.....	53

ABSTRAK

Ahsan, Thoriq Harizul.2023. **Peringkasan Teks Multi Dokumen Berita Berbahasa Indonesia Menggunakan *FastText* dan *K-means Clustering***. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (1) Zainal Abidin, M.Kom (2) Dr.M.Imamudin Lc, MA

Kata Kunci : *FastText*, *K-means Clustering*, *Multi Dokumen*, *Peringkasan Teks*

Perkembangan teknologi semakin maju sehingga berpengaruh pada persebaran berita online yang semakin meningkat. Hal ini akan menyulitkan pembaca berita dalam mencari inti informasi dari berita sehingga diperlukan cara untuk mendapatkan inti informasi dengan melakukan peringkasan teks berita multi dokumen secara otomatis. Tujuan peringkasan teks berita otomatis multi dokumen yaitu memudahkan pembaca dalam menghemat waktu saat memilah inti informasi yang relevan dari banyak berita, serta mencegah adanya redundansi maupun duplikasi informasi. Metode yang digunakan dalam meringkas adalah *FastText* dan *K-means Clustering*. Cara kerja *FastText* yaitu dengan menambahkan representasi kata dalam bentuk n-gram atau substring dari teks sehingga *FastText* dapat lebih mudah menangani kata-kata yang tidak dikenal atau kata-kata yang tidak muncul pada saat training data. Sedangkan *K-means Clustering* digunakan untuk mengelompokkan kalimat yang memiliki kedekatan. Kalimat yang memiliki kedekatan akan dipilih salah satu sebagai representasi dari ringkasan sehingga akan mengurangi redundansi dan duplikasi informasi. Hasil penelitian dengan input 30 data multi-dokumen yang memiliki 4 variasi label. Dari ke empat variasi label, ringkasan sistem mendapatkan nilai tertinggi ketika dibandingkan dengan variasi 1 dan terendah ketika dibandingkan dengan variasi 4. Dimana variasi 1 mendapat nilai rata-rata *precision* 0,674, rata-rata *recall* 0,620, dan rata-rata *f-measure* 0,637. Sedangkan pada variasi 4 mendapat nilai rata-rata *precision* 0,382, rata-rata *recall* 0,790, dan rata-rata *f-measure* 0,505.

ABSTRACT

Ahsan, Thoriq Harizul.2023. **Text Summarization Of Indonesian News Documents Using *FastText* and *K-means Clustering*** . Essay. Department of Informatics Engineering, Faculty of Science and Technology, Islamic State University of Maulana Malik Ibrahim of Malang. Supervisor: (I) Zainal Abidin, M.Kom. (II) Dr. M. Imamuddin Lc., MA.

Technological developments are increasingly advanced, so the effect of online news dissemination is increasing. This condition will make it difficult for news readers to find the core information, so a way is needed to get it using automatic multi-document text summarization for news. The purpose of automatic multi-document text summarization is to make readers easier to save time when sorting out the relevant core information from lots of news and to prevent redundancies or duplication of information. The method used in summarization is *FastText* and *Kmeans Clustering*. *FastText* works by adding word representations in the form of n-grams or substrings of the text so that *FastText* can make it easier to deal with unfamiliar words or words that do not appear during training data. At the same time, *Kmeans Clustering* is used to group sentences with proximity. Sentences that have proximity will be selected as a representation of the summary so that it will reduce redundancy and duplication of information. The research results with the input of 30 multi-document data have 4 label variations. Of the four label variations, the system summary gets the highest score compared to variation 1 and the lowest compared to variation 4. Variation 1 gets an average precision value of 0.674, an average recall of 0.620, and an average f-measure of 0.637. Whereas variation 4 gets an average precision value of 0.382, an average recall of 0.790, and an average f-measure of 0.505.

Keywords : *Text Summarization, FastText, K-means Clustering*

الملخص

الإحسان، طارق حارس. ٢٠٢٣. تلخيص نص متعدد المستندات للأخبار الإندونيسية باستخدام *FastText* و *K-means Clustering*. البحث الجامعي. قسم الهندسة المعلوماتية، كلية العلوم والتكنولوجيا بجامعة مولانا مالك إبراهيم الإسلامية الحكومية مالانج. المشرف الأول: زين العابدين، الماجستير. المشرف الثاني: د. محمد إمام الدين، الماجستير.

الكلمات الرئيسية: *K-means Clustering*، *FastText*، متعدد المستندات، تلخيص النص.

يتقدم تطور التكنولوجيا بشكل متزايد بحيث يؤثر على انتشار الأخبار عبر الإنترنت التي تتزايد. سيجعل هذا من الصعب على قراء الأخبار العثور على المعلومات الأساسية من الأخبار، لذلك هناك حاجة إلى طريقة للحصول على المعلومات الأساسية عن طريق تلخيص نص الأخبار متعدد المستندات تلقائياً. الهدف من التلخيص التلقائي للنصوص الإخبارية متعدد المستندات هو تسهيل الأمر على القراء لتوفير الوقت عند فرز جوهر المعلومات ذات الصلة من عدة الأخبار، ومنع التكرار وازدواجية المعلومات. الطريقة المستخدمة في التلخيص هي *FastText* و *K-means Clustering*. الطريقة التي يعمل بها *FastText* هي إضافة تمثيلات الكلمات في شكل سلاسل فرعية (n-gram) من النص بحيث يمكن ل *FastText* التعامل بسهولة أكبر مع الكلمات غير المعروفة أو الكلمات التي لا تظهر أثناء التدريب على البيانات. بينما يتم استخدام *K-means Clustering* لتجميع الجمل القريبة. سيتم اختيار الجمل المتقاربة كتمثيل للملخص بحيث يقلل من التكرار وازدواجية المعلومات. نتائج الدراسة مع إدخال ٣٠ بيانات متعدد المستندات تحتوي على ٤ اختلافات في التسمية. من بين أشكال التسميات الأربعة، يحصل ملخص النظام على أعلى درجة عند مقارنته بالشكل الأول وأدنى درجة عند مقارنته بالشكل الرابع. حيث يحصل الاختلاف الأول على متوسط قيمة دقة ٠.٦٧٤، واستدعاء ٠.٦٢٠، والفاء القياسي ٠.٦٣٧. بينما يحصل الشكل الرابع على متوسط قيمة دقة ٠.٣٢٨، واستدعاء ٠.٧٩٠، والفاء القياسي ٠.٥٠٥.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Berita merupakan kumpulan informasi yang berupa fakta sementara atau telah terjadi dan disebarluaskan melalui berbagai media penyiaran seperti dari mulut ke mulut, media cetak, penyiaran radio, dan internet (Rani, 2013). Berita mudah ditemukan baik secara lisan seperti dari mulut ke mulut kepada masyarakat, siaran radio dan televisi dan juga dapat secara tertulis melalui surat kabar, majalah, atau artikel *online* pada internet. Saat ini, berita *online* yang berupa artikel sangat mudah ditemukan dimana saja karena di posting dalam internet yang tidak memiliki batasan waktu sehingga dapat diakses kapan saja dan dimana saja.

Perkembangan teknologi internet saat ini semakin maju sehingga berpengaruh juga pada kendali persebaran berita *online* di internet. Sehingga pada situasi tersebut semakin meningkatnya jumlah ketersediaan data di internet setiap saat. Dengan meningkatnya jumlah persebaran berita *online* di internet maka semakin menyulitkan masyarakat atau pembaca berita dalam memahami inti informasi dari berita yang ada. Padahal dengan banyaknya jumlah berita yang ada, inti informasi yang sebenarnya tetap sama. Hal ini akan semakin membutuhkan banyak waktu bagi pembaca berita dalam memilah inti informasi yang dibutuhkan. Sehingga diperlukan cara untuk mendapatkan inti berita tersebut yaitu dengan melakukan peringkasan teks berita secara otomatis. Tujuan dari peringkasan teks otomatis yaitu untuk memperoleh inti informasi dari suatu dokumen (Hahn, 2000). Dengan penggunaan sistem peringkasan teks otomatis ini dapat memudahkan

pembaca berita dalam mendapat inti informasi tanpa harus membaca isi berita secara keseluruhan.

Jenis peringkasan teks otomatis jika dilihat dari inputnya dapat menggunakan single dokumen atau multi dokumen. Sedangkan dilihat dari outputnya dapat dikategorikan dalam bentuk ekstraktif dan abstraktif. Metode peringkasan ekstraktif membuat ringkasan secara verbal dengan mencari bagian-bagian penting dari teks. Sehingga mereka hanya bergantung pada ekstraksi kalimat dari teks aslinya. Sedangkan metode peringkasan abstraktif bertujuan menghasilkan teks baru yang lebih pendek namun tetap memberikan informasi inti dari teks aslinya (Allahyari et al., 2017).

Peringkasan teks otomatis multi dokumen adalah peringkasan teks otomatis yang jenis masukannya berasal dari banyak dokumen dari satu atau lebih sumber yang relevan dengan topik dokumen tersebut dan memiliki satu ide pokok yang sama dan bisa juga memiliki ide pokok yang berbeda (Hovy, 2005). Tujuan peringkasan teks berita otomatis multi dokumen yaitu memudahkan pembaca dalam menghemat waktu saat membaca atau memilah inti informasi yang relevan dari dokumen berita tersebut, sehingga dapat mencegah dari adanya redundansi maupun duplikasi informasi pada dokumen berita. Dalam Qur'an surah Al-Maidah 5:2 Allah *subhanahu wa ta'ala* berfirman:

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَحْلُوا شَعِيرَ اللَّهِ وَلَا الشَّهْرَ الْحَرَامَ وَلَا الْهَدْيَ وَلَا الْقَلَائِدَ وَلَا آيَاتِ الْحَرَامِ الَّتِي بَيَّنَّا كُتُبًا مِّن رَّبِّكُمْ وَرِضْوَانًا ۚ وَإِذَا حَلَلْتُمْ فَاصْطَادُوا ۚ وَلَا يَجْرِمَنَّكُمْ شَنَاٰنُ قَوْمٍ أَن صَدُّوكُمْ عَنِ الْمَسْجِدِ الْحَرَامِ أَن تَعْتَدُوا ۚ وَتَعَاوَنُوا عَلَى الْبِرِّ وَالتَّقْوَىٰ ۚ وَلَا تَعَاوَنُوا عَلَى الْإِثْمِ وَالْعُدْوَانِ ۚ وَاتَّقُوا اللَّهَ ۚ إِنَّ اللَّهَ شَدِيدُ الْعِقَابِ

“Hai orang-orang yang beriman, janganlah kamu melanggar syi'ar-syi'ar Allah, dan jangan melanggar kehormatan bulan-bulan haram, jangan (mengganggu) binatang-binatang had-ya, dan binatang-binatang qalaa-id, dan jangan (pula)

mengganggu orang-orang yang mengunjungi Baitullah sedang mereka mencari kurnia dan keridhaan dari Tuhannya dan apabila kamu telah menyelesaikan ibadah haji, maka bolehlah berburu. Dan janganlah sekali-kali kebencian(mu) kepada sesuatu kaum karena mereka menghalang-halangi kamu dari Masjidilharam, mendorongmu berbuat aniaya (kepada mereka). Dan tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan pelanggaran. Dan bertakwalah kamu kepada Allah, sesungguhnya Allah amat berat siksa-Nya.” (QS. al-Maidah :2)

Menurut tafsir Ibnu Katsir, pada ayat *وَتَعَاوَنُوا عَلَى الْبِرِّ وَالتَّقْوَى* Allah memerintahkan kepada hamba-Nya untuk saling tolong menolong dalam berbuat hal yang baik serta dilarang membantu dalam segala bentuk kebatilan atau hal-hal yang mungkar yang diharamkan oleh Allah SWT. Dalam hubungan antara manusia, berita merupakan kumpulan informasi yang sangat penting dalam mengetahui keadaan yang telah dan sedang terjadi di masyarakat. Sehingga peringkasan teks berita otomatis diharapkan dapat menjadi salah satu kebajikan untuk memudahkan masyarakat dalam mendapatkan inti informasi dalam kumpulan berita yang relevan.

Peringkasan teks dapat dilakukan secara otomatis dengan memanfaatkan teknologi kecerdasan buatan dalam membuat sebuah sistem untuk melakukan peringkasan teks berita berdasarkan pola atau model yang telah diberikan. Salah satunya adalah *FastText* dimana *FastText* pernah digunakan oleh Khaerunnisa et al. (2022) sebagai ekstraksi fitur dengan pengklasifikasiannya menggunakan *Recurrent Neural Network* (RNN) dan *Gated Recurrent Unit* (GRU) dalam meringkas teks ekstraktif. Evaluasi model dilakukan dengan menggunakan metrik *f-measure ROUGE-1*, *ROUGE-2*, dan *ROUGE-L*. Hasil penelitian menunjukkan bahwa model yang menggunakan *FastText* sebagai word embedding menghasilkan

performa terbaik, dengan nilai f-measure *ROUGE-1* sebesar 70.82, *ROUGE-2* sebesar 66.525, dan *ROUGE-L* sebesar 70.

Algoritma *FastText* adalah jenis *Word Embedding* yang merupakan pengembangan model dari *Word2Vec* dan pernah dikembangkan oleh Facebook untuk memprediksi tag melalui konteks (tag ini adalah jenis teks, yang ditentukan oleh anotasi manual) (Bojanowski et al., 2017). Algoritma *FastText* digunakan dalam peringkasan teks otomatis selain karena kemampuan dalam menangani dataset yang besar dan memiliki banyak variasi kata, *FastText* juga akan menambahkan representasi kata dan kalimat ke dalam bentuk vektor dari bentuk n-gram atau substring dari teks sehingga *FastText* dapat lebih mudah menangani kata-kata yang tidak dikenal atau kata-kata yang tidak muncul pada saat training data. Setelah kalimat dapat direpresentasikan dengan baik maka representasi tersebut dapat diolah dengan mudah dalam proses-proses berikutnya.

Pendekatan lain *unsupervised* dalam peringkasan teks pernah dilakukan juga oleh Pramudi Ismi et al. (2019) dimana *TFIDF* digunakan sebagai ekstraksi fitur sedangkan *K-means Clustering* dipilih sebagai metode yang digunakan dalam peringkasan teks. Hasil yang didapat juga cukup baik dimana nilai *F1 score* dari *ROUGE-1* = 49,37%, *ROUGE-2* = 38,18% dan *ROUGE-L* = 46,87%. Hasil dari metode yang digunakan tersebut menyatakan lebih baik dari metode *unsupervised SumBasic*, *LexRank*, dan *LSA*. Dalam meringkas, kalimat yang memiliki kesamaan dapat dihilangkan untuk mengurangi redundansi. Dalam proses inilah algoritma *K-means Clustering* diperlukan. *K-means Clustering* merupakan algoritma yang

digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang saling berdekatan (*clusters*).

Dengan keberhasilan dari kedua metode tersebut dalam meringkas teks, jenis dokumen yang digunakan masih menggunakan jenis single dokumen. Sehingga dengan adanya hal tersebut, penelitian ini berusaha mengembangkan dari metode metode sebelumnya dimana sistem peringkasan teks akan dibuat menggunakan jenis multi dokumen. Sistem yang dibuat nantinya perlu dilakukan pengukuran untuk mengetahui seberapa baik performanya dalam meringkas teks multi dokumen. Pengukuran yang digunakan akan menggunakan ROUGE dimana akan diukur nilai *Precision*, *Recall*, *f-measure* pada metode *FastText* sebagai ekstraksi fitur serta *K-means Clustering* sebagai salah satu pendekatan dalam meringkas teks.

1.2 Pernyataan Masalah

Dibutuhkan sebuah sistem peringkasan teks multi dokumen berita berbahasa Indonesia menggunakan metode *FastText* dan *K-means Clustering* yang dievaluasi menggunakan ROUGE untuk mendapatkan nilai *Precision*, *Recall*, *f-measure*.

1.3 Tujuan Penelitian

Membuat sebuah sistem peringkasan teks multi dokumen berita berbahasa Indonesia menggunakan metode *FastText* dan *K-means Clustering* yang dievaluasi menggunakan ROUGE untuk mendapatkan nilai *Precision*, *Recall*, *f-measure*.

1.4 Batasan Masalah

- a. Penelitian ini menggunakan dataset kumpulan berita tunggal dari IndoSum dataset yang dibentuk ulang menjadi dataset multi dokumen (Kurniawan & Louvan, 2019).
- b. Proses pelabelan pada dataset multi dokumen dilabeli oleh 5 orang responden.
- c. Penelitian ini berfokus pada peringkasan ekstraktif.

1.5 Manfaat Penelitian

- a. Bagi pemerintah, hasil penelitian ini dapat membantu memperoleh informasi secara cepat dari berbagai sumber (multi dokumen) sehingga membantu dalam pengambilan dan perencanaan kebijakan publik yang lebih baik.
- b. Bagi pelaku bisnis, dapat membantu melakukan riset pasar dan tren industri berdasarkan berita yang diperoleh dari berbagai sumber secara cepat.
- c. Bagi masyarakat, dapat membantu meningkatkan kesadaran publik tentang perkembangan isu politik, sosial, ekonomi, dan lingkungan berdasarkan berita secara cepat.

BAB II

KAJIAN PUSTAKA

2.1 Peringkasan Teks Berita

Peringkasan teks merupakan proses menghilangkan kalimat yang dianggap tidak penting tanpa mengurangi inti informasi dalam teks dokumen tunggal atau multi dokumen. Peringkasan teks akan menghasilkan ukuran teks yang panjangnya tidak lebih dari ukuran teks aslinya (Nur Aisyah et al., 2019). Proses peringkasan teks dapat dilakukan menggunakan 2 macam pendekatan yaitu pendekatan abstraktif dan pendekatan ekstraktif. Pendekatan abstraktif berfokus pada pengambilan inti informasi pada teks sehingga menghasilkan ringkasan teks dan pendekatan ekstraktif berfokus pada pemilihan kata kunci dalam kalimat dari sumber aslinya (Abidin & Nurjanah, 2020). Berita merupakan laporan mengenai fakta sementara atau telah terjadi yang disiarkan melalui media baik cetak, internet, atau dari mulut ke mulut.

Hendrastuty & SN (2021) pernah melakukan penelitian mengenai peringkasan teks multi dokumen menggunakan algoritma genetika untuk dilakukan ekstraksi fitur teks dalam setiap kromosom yang berupa data jumlah aduan masyarakat pada pemerintah DIY. Data yang digunakan dari penelitian ini diambil dari website lapor DIY dari bulan februari 2018 sampai Juli 2020 yang berjumlah 1000 data. Pada penelitian tersebut penulis melakukan ekstraksi fitur dengan mengambil beberapa fitur yaitu posisi kalimat, kata kunci positif dan negatif, kemiripan antara masing-masing kalimat, panjang kalimat, kalimat yang mengandung kata entitas dan angka, serta jumlah dan koneksi antar kalimat itu

sendiri. Hasil pengujian pada penelitian ini didapatkan nilai rata-rata presisi 1, *Recall* 0.71, dan f-measure sebesar 0.79.

Widjanarko et al. (2018) pernah melakukan penelitian peringkasan teks multi dokumen berbahasa indonesia menggunakan metode LDA (*Latent Dirichlet Allocation*) dan *Significance sentence*. Penulis menerapkan model LDA sebagai kerangka peringkasan ekstraktif dalam penelitian ini. Uji coba penelitian ini menggunakan 5 dataset berita *online* berbahasa indonesia yang didapatkan dari *cnnindonesia.com*. Penelitian ini menggunakan *cosine similarity* untuk mengukur kinerja ringkasan. Pengujian model menggunakan kombinasi parameter alfa 0,1 dan 0,001 serta parameter beta 0,001 dan 0 yang kemudian dikombinasikan dengan kompresi tingkat 10%, 30%, dan 50% dalam proses peringkat kalimat setiap dokumen. Hasil pengujian menghasilkan hasil terbaik didapatkan dari kombinasi parameter dengan nilai alpha sebesar 0,1 dan beta sebesar 0,001 dengan tingkat kompresi 50% dan nilai *cosine similarity* sebesar 0,931.

2.2 *FastText*

FastText merupakan sebuah model representasi kata dalam vektor atau embedding kata yang dikembangkan oleh Bojanowski et al. (2017) . Model *FastText* memperhatikan aspek kedekatan makna kata dan juga kedekatan struktur kata dan kalimat (Bojanowski et al., 2017). Khaerunnisa (2022) telah melakukan penelitian tentang *FastText* dengan membandingkan kinerja tiga jenis word embedding, yaitu Word2Vec, GloVe, dan *FastText*, dalam peringkasan teks ekstraktif berita berbahasa Indonesia menggunakan metode *Recurrent Neural Network* (RNN) dan *Gated Recurrent Unit* (GRU) sebagai ekstraksi fitur dan

klasifikasi kalimat ringkasan. Evaluasi model dilakukan dengan menggunakan metrik *f-measure* ROUGE-1, ROUGE-2, dan ROUGE-L. Hasil penelitian menunjukkan bahwa model yang menggunakan *FastText* sebagai word embedding menghasilkan performa terbaik, dengan nilai *f-measure* ROUGE-1 sebesar 70.82, ROUGE-2 sebesar 66.525, dan ROUGE-L sebesar 70.288 (Khaerunnisa, et al., 2022).

Penelitian lain juga telah dilakukan sebelumnya. Amalia et al. (2020) telah melakukan penelitian tentang kinerja model *FastText* dalam klasifikasi teks untuk dokumen dalam bahasa Indonesia. Penelitian ini bertujuan membandingkan model *FastText* dengan model *tfidf*. Penelitian ini menggunakan dataset 500 dokumen berbahasa Indonesia yang terdiri dari 6 kategori, seperti politik, lingkungan, teknologi, bisnis, dan kesehatan. Model *FastText* dilatih pada dataset tersebut dan diuji pada data tes yang terpisah. Hasil penelitian menunjukkan bahwa model *FastText* dapat mencapai tingkat akurasi yang baik dalam klasifikasi dokumen dalam bahasa Indonesia, dengan akurasi mencapai 0,97. Hasil tersebut meningkat 13% dari model TFIDF yang memiliki akurasi 0,84. Hasil ini menunjukkan bahwa model *FastText* dapat digunakan untuk klasifikasi dokumen dalam bahasa Indonesia dengan kinerja yang baik.

Arsitektur model yang digunakan dalam *FastText* adalah arsitektur terdistribusi. Jadi tujuannya adalah untuk mempelajari representasi *high-dimensional dense* untuk setiap istilah kosakata. Representasi vektor yang dihasilkan memiliki dimensi yang tinggi (*high-dimensional*) dan padat (*dense*) memiliki arti bahwa setiap dimensi vektor memungkinkan mewakili konsep yang

berbeda, dan jarang ditemui nilai 0 dari dimensi tersebut, sehingga membuat representasi itu sendiri padat. Representasi harus bersifat distribusi dan juga mencoba untuk belajar dari konteks. Tujuannya untuk memperoleh representasi yang mewakili makna kata-kata yang lebih baik daripada representasi tradisional seperti *one-hot encoding* atau representasi lainnya.

FastText memiliki 2 pilihan model arsitektur yaitu Skipgram dan CBOW. Pada Skipgram, menggunakan sebuah target untuk memprediksi kata-kata konteks yang ada disekitarnya. Hal ini dilakukan dengan melewati proses *one hot encoding* dan dilanjutkan menuju lapisan *shallow neural network* yang hanya memiliki dua lapisan sederhana. Karena inputnya adalah kata yang telah dilakukan proses *one hot encoding*, sehingga *hidden layer* hanya terdiri dari satu baris dari matriks bobot *hidden input*. Tujuan adanya *neural network* adalah untuk memprediksi kata-kata yang kemungkinan muncul dalam konteks tertentu, yaitu sejumlah kata sebelum atau sesudah kata yang diberikan, dengan mempertimbangkan kata yang diberikan. Untuk skor pada tiap kata dapat dihitung dengan persamaan 2.1.

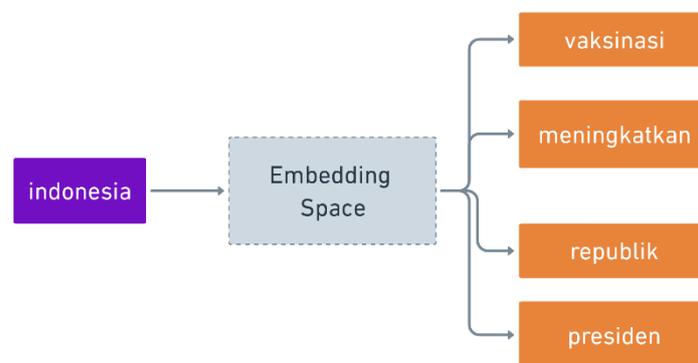
$$u = W'^T h \quad (2.1)$$

Dimana h adalah vektor pada *hidden layer* dan W adalah matriks bobot pada *hidden output*. setelah vektor representasi kata u (kata masukan) berhasil dihitung, maka distribusi *multinomial weight* untuk setiap kata konteks dalam jarak c (*window size*) dari kata masukan tersebut akan dihitung menggunakan persamaan 2.2.

$$p(w_{c,j} = w_{0,c} | w_I) = \frac{\exp u_{c,j}}{\sum_{j'=1}^V \exp u_{j'}} \quad (2.2)$$

Dimana $w_{c,j}$ adalah kata ke- j pada panel c dari *output layer*, $w_{0,c}$ adalah kata yang sebenarnya yang muncul pada posisi c dalam *output context word*. w_I merupakan satu-satunya kata masukan, dan $u_{c,j}$ adalah vektor representasi kata masukan ke- j pada panel c dari *output layer*. Sehingga dapat dilihat pada probabilitas tersebut mencoba memprediksi kata konteks yang diberikan kata masukan. Probabilitas tersebut kemudian diubah menjadi softmax. Untuk visualisasi arsitektur tersebut dapat dilihat pada Gambar 2.1.

... presiden republik indonesia meningkatkan vaksinasi ...



Gambar 2.1 Arsitektur Skipgram

Dari Gambar 2.1, misalnya diberikan sebuah kalimat “presiden republik indonesia meningkatkan vaksinasi”. Jika kata “indonesia” menjadi kata input maka akan diprediksi beberapa kata disekitar kata “indonesia” yang akan menjadi kata output. Seperti kata “vaksinasi”, “meningkatkan”, “republik”, “presiden”.

Sedangkan pada arsitektur CBOW berkebalikan dengan skipgram. Pada CBOW beberapa kata konteks akan digunakan untuk memprediksi kata target. Algoritma pada CBOW adalah dengan mengambil vektor *one hot* dari seluruh kata karena semua kata konteks nantinya akan dijadikan sebagai input. Nilai *window* k

perlu dipertimbangkan, karena jika terlalu besar bisa menghasilkan vektor one hot yang besar juga. Kemudian akan diambil nilai vektor embedding woda untuk semua kata. Untuk mendapatkan cumulative konteks maka akan dirata-ratakan. Output pada hidden layer atau h akan dibuat dengan mengikuti persamaan 2.3.

$$h = \frac{1}{C} W \cdot \left(\sum_i^c x_i \right) \quad (2.3)$$

Perlu dicatat bahwa lapisan tersembunyi adalah salah satu perbedaan utama antara skipgram dan CBOW. Sedangkan untuk perhitungan nilai skor tiap kata persamaannya hampir sama dengan Skipgram namun karena CBOW memprediksi semua kata dalam output berdasarkan konteksnya, maka u dan v untuk kolom yang berbeda (dilambangkan dengan j) perlu dihitung dengan persamaan berikut:

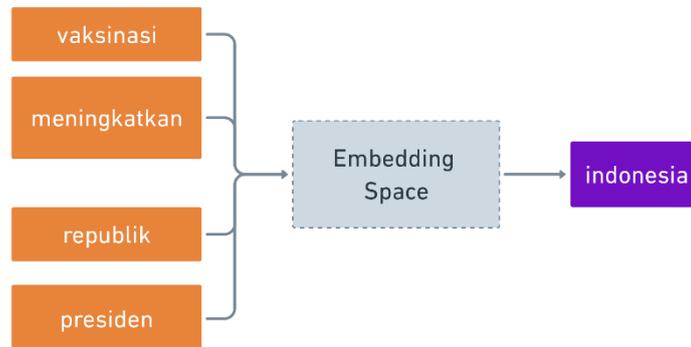
$$u_j = v_{w_j}^T \cdot h \quad (2.4)$$

Perhitungan skor perlu diubah menjadi probabilitas menggunakan softmax pada persamaan 2.5. Model perlu dilatih sehingga probabilitasnya cocok dengan probabilitas sebenarnya dari kata tersebut, yang merupakan *one hot encoding* dari kata yang sebenarnya:

$$y_j = p(w_{y_i} | w_1, \dots, w_c) = \frac{\exp(u_j)}{\sum_{j=1}^V \exp(u'_j)} \quad (2.5)$$

visualisasi arsitektur tersebut dapat dilihat pada Gambar 2.2 .

... presiden republik indonesia meningkatkan vaksinasi ...



Gambar 2.2 Arsitektur CBOV

Dari Gambar 2.2, misalnya diberikan sebuah kalimat “presiden republik indonesia meningkatkan vaksinasi”. Maka semua kata disekitar kata “indonesia” seperti kata “vaksinasi”, “meningkatkan”, “republik”, “presiden” akan dijadikan sebagai kata input untuk memprediksi kata “indonesia”.

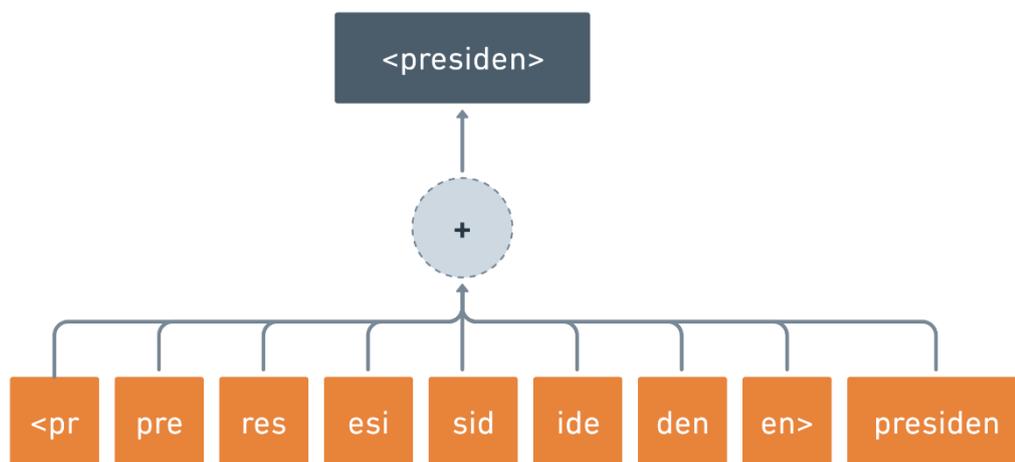
Menurut Mikolov et al. (2013), diantara dua arsitektur tersebut memiliki kekurangan dan kelebihan masing-masing. Skipgram bekerja dengan baik dengan sejumlah kecil data latih. Skipgram bekerja baik bahkan pada kata dan frasa yang langka. Sedangkan CBOV lebih cepat dalam proses latih dari pada skipgram. CBOV memiliki akurasi yang lebih tinggi pada kata dan frasa yang sering muncul.

Kedua arsitektur Skipgram dan CBOV berasal dari implementasi word embedding sebelumnya yaitu *word2vec*. *FastText* juga menggunakan arsitektur tersebut, namun perbedaannya, *FastText* menambahkan penerapan dari *subword* atau karakter n-gram sebagai fitur tambahan untuk membangun vektor kata. Hal ini diperlukan karena karakter pada n-gram menyimpan informasi yang lebih banyak. Sehingga dengan adanya karakter pada n-gram maka model *FastText*

dapat menangani kata inputan yang diluar *vocabulary* selama masih ada dalam kombinasi n-gram (Bhattacharje, 2018). Sehingga pada *FastText*, fungsi skor (u) sebelumnya sebenarnya diubah menjadi persamaan berikut:

$$s(w, c) = \sum_{g \in G_w} z_g^T V_c \quad (2.6)$$

Sehingga untuk kata “presiden” misalnya akan memiliki *collection* dari n-gram seperti yang ditunjukkan pada Gambar 2.3 .



Gambar 2. 3 Hashing n-gram

2.3 *K-means Clustering*

K-means Clustering merupakan algoritma yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang saling berdekatan (*clusters*) berdasarkan pada kesamaan fitur atau atribut yang dimiliki oleh data tersebut. Algoritma *K-means Clustering* pertama kali diusulkan oleh Stuart Lloyd (1957). James MacQueen (1967) memperkenalkan kembali algoritma *K-means Clustering* dan memberikan namanya.

Beberapa penelitian telah dilakukan sebelumnya tentang penggunaan *K-means Clustering* dalam mengatasi berbagai permasalahan. Twinandilla et al. (2018) melakukan penelitian tentang peringkasan teks berita *online* multi dokumen dengan tujuan memudahkan pembaca dalam memahami tema serta konsep berita dan meringkas waktu dalam membaca. Penulis menggunakan metode peringkasan baru dengan menggabungkan *K-means Clustering* dan LDA (*Latent Dirichlet Allocation*) - Significance Sentences untuk menghasilkan ringkasan dokumen berdasarkan topik berita. Penulis menggunakan 2 skenario uji coba, yang pertama hasil eksperimen nilai alpha terbaik yang didapat sebesar 0.001 dengan nilai ROUGE-1 sebesar 0.5545 dan untuk tingkat peringkasan terbaik didapat sebesar 30% dengan nilai ROUGE-1 yaitu 0.6118. Percobaan kedua penulis memperoleh nilai ROUGE-1 terbaik yaitu 0.61991.

Akaresti et al. (2019) juga melakukan penelitian peringkasan multi dokumen berita *online* berbasis clustering dan metode K-Means. Penulis bertujuan untuk menggabungkan sistem peringkasan dan pencarian untuk menghindari mendapatkan informasi yang sama dengan sumber yang berbeda. Penelitian ini berfokus pada penggabungan sistem peringkasan dan pencarian tanpa melakukan proses stemming pada proses *preprocessing*. Tahap pencarian dalam penelitian ini yaitu tahapan dimana sebuah query atau kata kunci dimasukkan ke dalam sistem dan sistem akan memproses output hasil ringkasan berita yang sesuai dengan kata kunci. Penulis menggunakan data yang diambil dari web berita *online* yang berbeda yaitu Kompas.com, tribunnews.com, Detik.com, Liputan6.com dengan total jumlah 200 berita. Proses pengujian sistem pencarian dilakukan dengan 3

variasi skenario terhadap jumlah dokumen hasil *cosine similarity* dengan ranking tertinggi yaitu 4, 7, dan 10 dokumen tertinggi. Hasil optimal terdapat pada 10 dokumen tertinggi yaitu skenario 3 dengan nilai *Recall* 71%, *Precision* 65.82%, dan f-measure 66.35%. Hasil optimal pengujian sistem peringkasan didapatkan dari skenario 3 dengan nilai *Recall* 37,3%, *Precision* 18% , f-measure 19.2%.

Algoritma *K-means Clustering* bekerja dengan menentukan jumlah cluster atau nilai *k* diawal. Kemudian akan diukur jarak antar data dengan *centroid*-nya. Adapun rumus jarak dapat menggunakan metode *euclidean distance* :

$$d(x_i, \mu_i) = \sqrt{\sum_{i=1}^n (x_i - \mu_i)^2} \quad (2.7)$$

Dari perhitungan jarak maka akan diketahui sebuah data masuk dalam suatu cluster tertentu berdasarkan jarak paling kecil diantara semua cluster. Nilai *centroid* untuk pertama kali akan dibentuk secara acak dan pada setiap iterasi akan diperbarui nilai baru dari *centroid* dengan mengikuti persamaan berikut.

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad (2.8)$$

μ_k = titik *centroid* dari cluster ke-K

N_k = banyaknya data pada cluster ke-K

x_q = data ke-q pada cluster ke-K

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Data Penelitian

Data yang digunakan pada penelitian ini adalah dataset multi dokumen yang dibentuk dari IndoSum Dataset. IndoSum dataset sendiri berisi kumpulan data ringkasan teks berita berbahasa Indonesia yang mengambil dari artikel berita *online* di berbagai sumber. Dataset IndoSum berasal dari *repository* (<https://github.com/kata-ai/indosum>). Dataset IndoSum disimpan dengan format *.jsonl* (json line) dimana data tersebut dibagi menjadi *5-fold corpus*. Tiap *fold* terdiri dari 3 file *.jsonl* yaitu file *train*, *dev*, dan *test*.

Tiap file dalam format *.jsonl* tersebut memiliki kumpulan objek berita. Dapat dilihat dari Gambar 3.1 pada tiap objek berita memiliki beberapa kolom yaitu “category”, “gold_labels”, “id”, “paragraphs”, “source”, “source_url”, dan “summary” (Kurniawan & Louvan, 2019).

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14262 entries, 0 to 14261
Data columns (total 7 columns):
#   Column          Non-Null Count  Dtype
---  -
0   category        14262 non-null  object
1   gold_labels     14262 non-null  object
2   id              14262 non-null  object
3   paragraphs      14262 non-null  object
4   source          14262 non-null  object
5   source_url     14262 non-null  object
6   summary         14262 non-null  object
dtypes: object(7)
memory usage: 780.1+ KB
```

Gambar 3.1 Struktur IndoSum Dataset

Dapat dilihat pada Gambar 3. 3 berdasarkan hasil pengelompokkan objek berita, penelitian ini menghasilkan 30 berita multi dokumen dengan masing-masing berita berisi 3 hingga 5 berita single dokumen. Proses pelabelan dilakukan ulang dengan melibatkan 5 orang responden yang merupakan pembaca berita, dengan masing-masing responden diberi bobot 1. Tiap kalimat dalam multi dokumen akan di berikan nilai *true* atau *false*, dimana label tersebut menentukan kalimat tersebut masuk dalam ringkasan atau tidak. Label akhir yang dihasilkan merupakan label dengan pilihan *true* dengan jumlah 50% lebih dari jumlah bobot jawaban.

Tabel 3. 1 Variasi Bobot untuk Pembentukan Label

r1	r2	r3	r4	r5	indosum	maximum	minimum	keterangan
1	1	1	1	1	0 (Tanpa indosum)	5	2,5	variasi toleransi 1
1	1	1	1	1	1	6	3	variasi toleransi 2
1	1	1	1	1	2	7	3,5	variasi toleransi 3
1	1	1	1	1	5	10	5	variasi toleransi 4

Bobot jawaban selain berasal dari label responden juga menyertakan label Indosum dataset, dimana pembentukan label IndoSum dataset diberikan bobot yang berbeda, dengan variasi nilai 0, 1, 2, dan 5. Tabel 3.1 menunjukkan gambaran bagaimana variasi bobot akan terbentuk. Pada kolom indosum nilai 0 berarti jawaban dari label IndoSum diabaikan atau dengan kata lain hanya mempertimbangkan bobot dari responden saja. Kemudian didapati nilai 1 berarti bobot IndoSum bernilai 1, nilai 2 berarti bobot label IndoSum bernilai 2, dan nilai 5 berarti bobot label IndoSum bernilai 5. Sehingga menghasilkan nilai maksimum jawaban yang berbeda pada tiap variasinya. Sedangkan nilai minimum didapat dari 50% dari nilai maksimum yang akan digunakan sebagai acuan sebuah jawaban

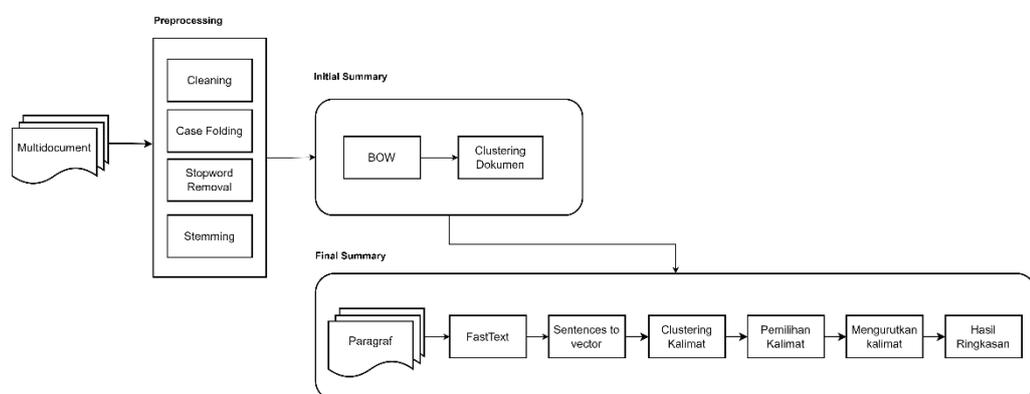
akan dikategorikan sebagai ringkasan atau tidak jika jawaban *true* melebihi dari nilai minimum tersebut. Sehingga label akhir pada dataset yang terbentuk dari 30 multi dokumen dapat dilihat pada Gambar 3.4 .

responden1	responden2	responden3	responden4	responden5	Indosum_label1	label_0	label_1	label_2	label_5
[[True, False, False, True, False, False, Fals...	[[True, False, True, True, False, False, Fals...	[[True, False, False, False, True, False, Fals...	[[True, False, False, False, True, False, Fals...	[[True, False, False, False, True, False, True...	[[True, True, True, True, False, False, Fals...	[[True, False, False, False, True, False, Fals...	[[True, False, False, False, True, False, Fal...	[[True, False, False, True, False, False, Fals...	[[True, False, True, True, False, False, Fals...
[[True, True, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, ...	[[True, True, False, True, True, False, True, ...	[[True, False, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, ...	[[True, False, True, True, True, True, True, ...	[[True, True, True, True, True, True, True, False, ...
[[True, False, True, False, True, False, True, False...	[[True, False, True, False, True, False, True, False...	[[True, False, True, False, True, False, True, ...	[[True, False, True, False, True, True, True, ...	[[True, False, True, False, True, True, True, False...	[[True, True, True, False, False, False, False...	[[True, False, True, False, True, True, True, False...	[[True, False, True, False, True, True, True, False...	[[True, False, True, False, True, True, True, False...	[[True, False, True, False, True, True, True, Fals...
[[True, True, True, True, True, True, F...	[[True, False, True, False, True, False, True, ...	[[True, False, True, False, True, False, True, ...	[[True, False, True, False, True, True, True, ...	[[True, False, True, False, True, True, True, False...	[[True, True, False, False, False, False, Fals...	[[True, False, True, False, True, True, True, ...	[[True, False, True, False, True, True, True, False...	[[True, False, True, False, True, True, True, False...	[[True, True, False, False, False, False, Fals...
[[True, False, True, True, True, True, ...	[[True, False, False, True, True, True, True, False, ...	[[True, False, True, True, True, True, True, False, ...	[[True, False, False, True, True, True, True, False, ...	[[True, False, False, True, True, True, True, False, ...	[[True, False, False, True, True, True, True, False, ...	[[True, False, True, True, True, True, True, False, ...	[[True, False, True, True, True, True, True, False, ...	[[True, False, True, True, True, True, True, False, ...	[[True, False, True, True, True, True, True, Fals...

Gambar 3.4 Hasil pembentukan label dari 30 multi dokumen

Empat variasi dari dataset dapat dari Gambar 3.4 akan menjadi empat label baru seperti pada nama kolom “label_0”, “label_1”, “label_2”, dan “label_5”.

3.2 Desain Sistem



Gambar 3.5 Desain Sistem

Dalam penelitian ini rancangan sistem yang akan dibangun dapat dilihat pada Gambar 3.5. dimana tahapan yang harus dilakukan dalam penelitian yaitu

melakukan *preprocessing* data yang meliputi proses *case folding*, *cleaning*, *stopword removal*, dan *stemming* untuk mengubah kata dari Bahasa menjadi kata dasar. Setelah dilakukan *preprocessing* akan dilanjutkan tahap *initial summary*. Tahap ini bertujuan untuk mengelompokkan berita yang memiliki kedekatan topik. Tahap ini terdiri dari dua buah proses yaitu mengubah dokumen berita menjadi vektor dengan menggunakan *Bag of Words* dan mengelompokkan dokumen menggunakan *K-means Clustering* .

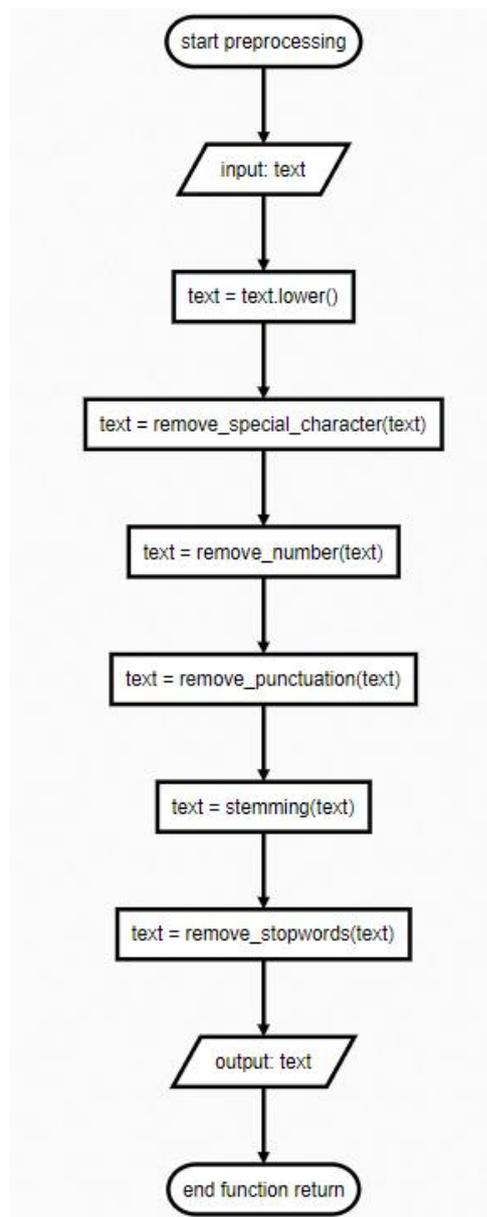
Tahapan akhir selanjutnya yaitu tahap *final summary*, pada tahap ini dilakukan proses mengubah kalimat menjadi vektor yang didapat dari hasil *training* model *FastText*, *clustering* kalimat dan memilih hasil kalimat dari hasil *clustering* kalimat sebagai representasi dari ringkasan, dan diakhiri dengan proses pengurutan menjadi sebuah kalimat sehingga menghasilkan ringkasan yang sesuai. Dalam proses pengujian data yang diambil berasal dari dataset yang telah dikumpulkan kemudian hasil ringkasan tersebut akan dilakukan evaluasi menggunakan rouge untuk menghitung tingkat akurasi dari model data.

3.3 Preprocessing

Data yang telah didapatkan perlu melalui tahap *preprocessing* untuk memperbaiki struktur inputan data. Secara umum tahapan *preprocessing* pada NLP meliputi *tokenizing*, namun karena pada penelitian ini menggunakan dataset IndoSum maka proses *tokenizing* tidak perlu dilakukan dikarenakan pada dataset IndoSum sudah dilakukan *tokenizing*.

Pada Gambar 3.6 menunjukkan bahwa tahap *preprocessing* yang pertama dilakukan yaitu *case folding* yaitu proses menyamakan data dengan mengubah

huruf kapital menjadi huruf kecil semua. Proses *casefolding* penting dilakukan untuk menghindari *case sensitive* antara huruf kapital dengan huruf kecil yang bisa menyebabkan perbedaan makna. Sehingga pada proses *casefolding* semua data inputan seperti nama kota yang awalnya huruf kapital disamakan menjadi huruf kecil semua.



Gambar 3. 6 Flowchart *Preprocessing*

Tahap kedua yaitu *cleaning* data untuk menghapus beberapa inputan yang tidak diperlukan seperti inputan email, selain itu juga menghapus karakter selain huruf seperti angka, karakter tautan, tanda baca, dan lain sebagainya. Tujuan dari *cleaning* data untuk menghilangkan data-data yang tidak diperlukan dalam proses di dalam sistem, sehingga data inputan yang tersisa hanya berupa data teks.

Tahap ketiga setelah melakukan *cleaning* yaitu melakukan proses *stemming*. Proses tersebut akan mengubah kata yang awalnya kata dalam bahasa yang sudah berbentuk token kata menjadi kata dasar. Semua jenis imbuhan yang ada seperti prefiks dan sufiks dihilangkan.

Tahap terakhir dari *preprocessing* yaitu dilakukan penghapusan beberapa kata yang tidak ada hubungannya dengan dokumen serta cenderung tidak penting seperti kata konjungsi dan preposisi. Berikut contoh hasil tahapan *preprocessing* pada Tabel 3.2 .

Tabel 3.2 *Preprocessing* Data

Keterangan	Hasil
Berita	Presiden @jokowi mengumumkan bahwa pemerintah akan meningkatkan investasi di bidang teknologi informasi. Melalui program ini, pada tahun 2021 para pengusaha teknologi akan dapat mengajukan proposal proyek melalui website resmi pemerintah http://investasi.kemenkeu.go.id . Dalam sambutannya, Presiden @jokowi menyatakan bahwa inisiatif ini akan memperkuat ekonomi digital dan meningkatkan daya saing Indonesia di bidang teknologi. Para pengusaha dapat mengikuti informasi terbaru tentang program ini melalui akun twitter resmi @kemenkeu.
<i>Case folding</i>	presiden @jokowi mengumumkan bahwa pemerintah akan meningkatkan investasi di bidang teknologi informasi. melalui program ini, pada tahun 2021 para pengusaha teknologi akan dapat mengajukan proposal proyek melalui website resmi pemerintah http://investasi.kemenkeu.go.id . dalam sambutannya, presiden @jokowi menyatakan bahwa inisiatif ini akan memperkuat ekonomi digital dan meningkatkan daya saing indonesia di bidang teknologi. para pengusaha dapat mengikuti informasi terbaru tentang program ini melalui akun twitter resmi @kemenkeu.
<i>Cleaning</i>	presiden mengumumkan bahwa pemerintah akan meningkatkan investasi di bidang teknologi informasi. melalui program ini, pada tahun para pengusaha teknologi akan dapat mengajukan proposal proyek melalui website resmi pemerintah dalam sambutannya, presiden menyatakan bahwa inisiatif ini akan memperkuat ekonomi digital dan meningkatkan daya saing indonesia di

Keterangan	Hasil
	bidang teknologi. para pengusaha dapat mengikuti informasi terbaru tentang program ini melalui akun twitter resmi .
<i>Stopword Removal</i>	presiden mengumumkan pemerintah meningkatkan investasi bidang teknologi informasi . program , pengusaha teknologi mengajukan proposal proyek website resmi pemerintah sambutannya , presiden inisiatif memperkuat ekonomi digital meningkatkan daya saing indonesia bidang teknologi . pengusaha mengikuti informasi terbaru program akun twitter resmi
<i>Stemming</i>	presiden umum perintah tingkat investasi bidang teknologi informasi program usaha teknologi laju proposal proyek website resmi perintah sambut presiden inisiatif kuat ekonomi digital tingkat daya saing indonesia bidang teknologi usaha ikut informasi baru program akun twitter resmi

3.4 Peringkasan Awal

Tahap kedua setelah dilakukan tahap *preprocessing* adalah tahap peringkasan Awal. Tujuan dari tahap ini adalah terbentuknya kelompok berita yang memiliki kedekatan satu dengan yang lain. Kelompok berita ini juga bertujuan memisahkan berita yang kurang berkaitan. Dari tujuan tersebut maka akan terbentuklah kelompok-kelompok baru dari hasil *clustering*. Kelompok-kelompok tersebut nantinya diharapkan dapat disusun menjadi sebuah paragraf.

3.4.1. *Bag of Words*

Tahap pertama dari peringkasan awal yaitu dilakukan *Bag of Words*. Pada proses *Bag of Words* ini akan mempresentasikan kemungkinan seberapa banyak kemunculan kata yang terjadi terhadap sebuah dokumen (McCallum & Kachites, 1996). Dalam proses *Bag of Words* ini nantinya akan melakukan perhitungan terhadap frekuensi kemunculan kata dalam setiap kategori dokumen. Apabila sebuah kata terdapat dalam sebuah dokumen maka akan dipresentasikan dengan nominal jumlah kemunculan kata pada dokumen. Namun apabila kata tidak ditemukan pada sebuah dokumen maka akan dipresentasikan dengan nilai 0.

Sehingga dari perhitungan nilai *Bag of Words* ini maka berita akan memiliki representasinya dalam bentuk vektor *Bag of Words*. Adapun contoh penggunaan dokumen yang akan direpresentasikan sebagai *Bag of Words*.

- a. Berita 1: "presiden indonesia meningkatkan vaksinasi covid 19 di seluruh indonesia"
- b. Berita 2: "kasus covid19 pertama kali di indonesia terjadi pada bulan maret"
- c. Berita 3: "pembuat mobil listrik akan membuka pabrik baru"

Maka akan dihitung nilai *Bag of Words* nya dengan menghitung jumlah kemunculan kata tiap dokumennya yang dapat dilihat seperti pada Tabel 3.3.

Tabel 3.3 Perhitungan Frekuensi pada BOW

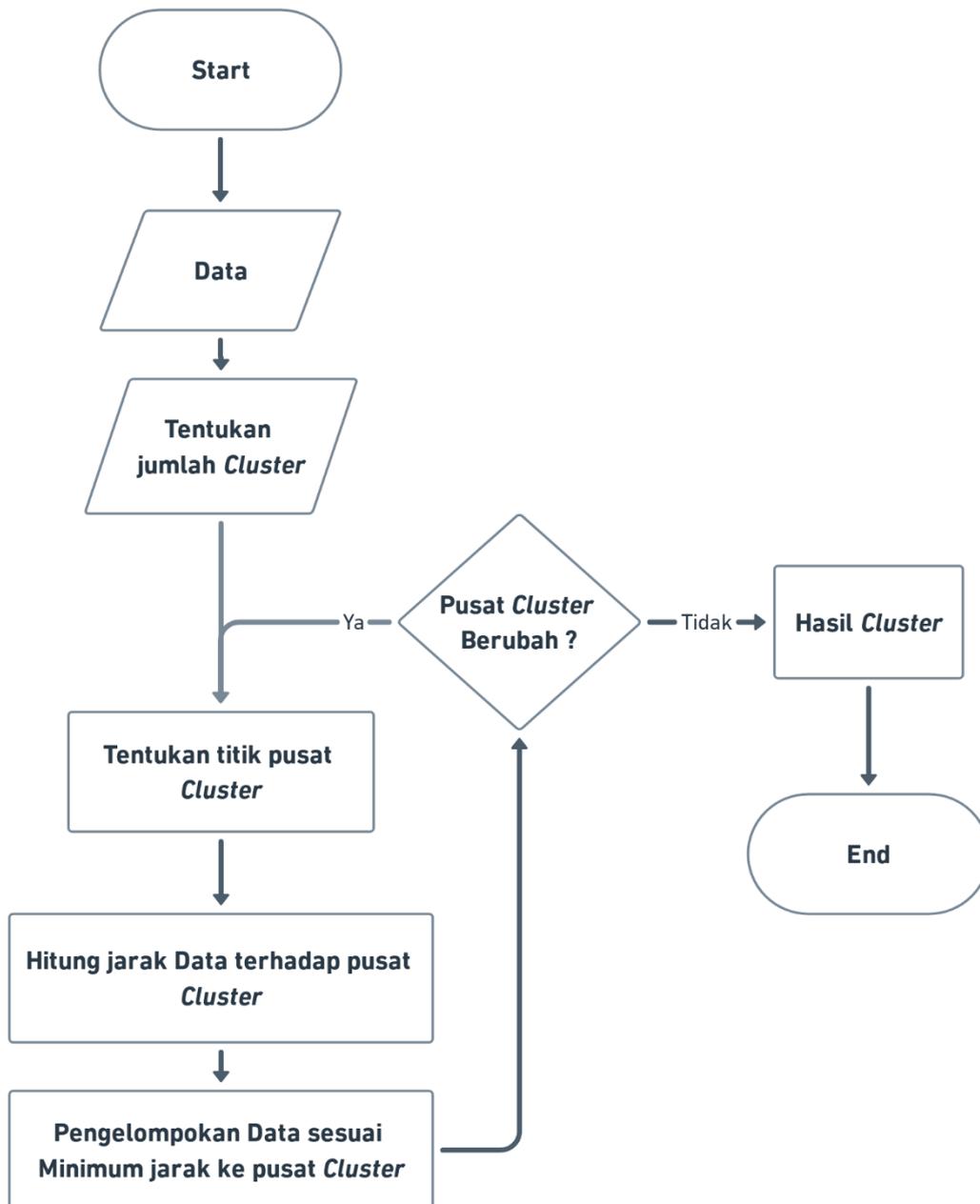
Kata	Berita 1	Berita 2	Berita 3
presiden	1	0	0
indonesia	2	1	1
meningkatkan	1	0	0
vaksinasi	1	0	0
covid19	1	1	0
di	1	1	1
seluruh	1	0	0
kasus	0	1	0
pertama	0	1	0
kali	0	1	0
terjadi	0	1	0
pada	0	1	1
bulan	0	1	0
maret	0	1	0
pembuat	0	0	1
mobil	0	0	1
listrik	0	0	1
membuka	0	0	1
pabrik	0	0	1
baru	0	0	1

Sehingga vektor *bag of words* yang dihasilkan pada tiap dokumen adalah sebagai berikut:

- a. Berita 1: [1, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- b. Berita 2: [0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]
- c. Berita 3: [0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1]

3.4.2. Clustering

Tahap selanjutnya yaitu melakukan proses *clustering* dengan mengelompokkan berita menjadi beberapa kelompok atau *cluster*. Algoritma yang digunakan untuk proses *clustering* yaitu menggunakan algoritma *K-Means*. Pada Gambar 3.7 terdapat gambaran *flowchart* mengenai proses *K-means Clustering*. Pada Gambar 3.7 juga menggambarkan alur Cara kerja algoritma *K-Means* yaitu dengan mengelompokkan objek ke dalam kelompok atau nilai k dan nilai k harus ditentukan terlebih dahulu sebelum dilakukan proses *clustering*. Selanjutnya memilih *centroid* awal dengan cara acak atau ditentukan berdasarkan beberapa titik yang dianggap representatif. Setelah itu dilakukan tahap penugasan data ke *cluster*, memperbarui *centroid*, dan perulangan proses dari penugasan data ke *cluster* hingga tidak ada perubahan pada penugasan data ke *cluster*.



Gambar 3.7 Flowchart *K-means Clustering*

Adapun contoh sederhana langkah-langkah perhitungan dalam proses *clustering* dokumen, mulai dari memasukkan data input, inialisasi *centroid* awal, menghitung jarak *centroid*, memperbarui *centroid*, hingga menghasilkan beberapa *cluster* adalah sebagai berikut:

1) Data input

Data input berasal dari dokumen yang telah diubah menjadi vektor menggunakan algoritma *Bag of Words*. Misal terdapat 3 dokumen yang telah diubah menjadi vektor yaitu:

- a. Berita 1: [1, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- b. Berita 2: [0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
- c. Berita 3: [0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1]

2) Inisiasi *Centroid* Awal

Jumlah k pada algoritma *K-means Clustering* ditentukan diawal. Misal $k = 2$. Maka akan di inisiasikan nilai *centroid* sebanyak 2 buah. Nilai *centroid* ditentukan secara acak.

- a. *Centroid* 1 : [1, 2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- b. *Centroid* 2: [0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0]

3) Menghitung jarak *centroid*

Menghitung jarak setiap data dengan setiap pusat *centroid*, dan memasukkan data ke dalam *cluster* yang memiliki jarak terdekat. Cara menghitung jarak antara dua vektor dapat menggunakan algoritma *Euclidean Distance* seperti pada persamaan 2.7 . Berikut adalah hasil perhitungan untuk setiap data yang telah dihitung jaraknya dengan *centroid* tiap data dokumen.

- a. Berita 1 Jarak ke *centroid* 1 = 1.414, Jarak ke *centroid* 2 = 4.24
- b. Berita 2 Jarak ke *centroid* 1 = 3.741, Jarak ke *centroid* 2 = 1.732
- c. Berita 3 Jarak ke *centroid* 1 = 3.316, Jarak ke *centroid* 2 = 2.449

Dari hasil perhitungan di atas, dapat dilihat bahwa kalimat pertama termasuk ke dalam *cluster* pertama sedangkan kalimat kedua dan ketiga termasuk ke dalam *cluster* 2.

4) Memperbarui *centroid*

Memperbarui pusat *centroid* untuk setiap *cluster* dengan mengambil nilai rata-rata dari setiap data dalam *cluster*. Proses memperbarui *centroid* ini menggunakan perhitungan pada persamaan 2.7. Berikut adalah hasil perhitungan untuk setiap pusat *centroid* baru:

- a. *Centroid* 1:[0.5, 1.5, 0.5, 0.5, 0.5, 1.0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
- b. *Centroid* 2: [0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

5) Hasil *Clustering*

Karena tidak ada perubahan pada setiap pusat *centroid*, kita dapat menganggap bahwa *clustering* sudah konvergen dan menghentikan iterasi. Sehingga hasil akhir dari *clustering* adalah:

- a. Berita 1 terdapat pada *cluster* 1
- b. Berita 2 terdapat pada *cluster* 2
- c. Berita 3 terdapat pada *cluster* 2

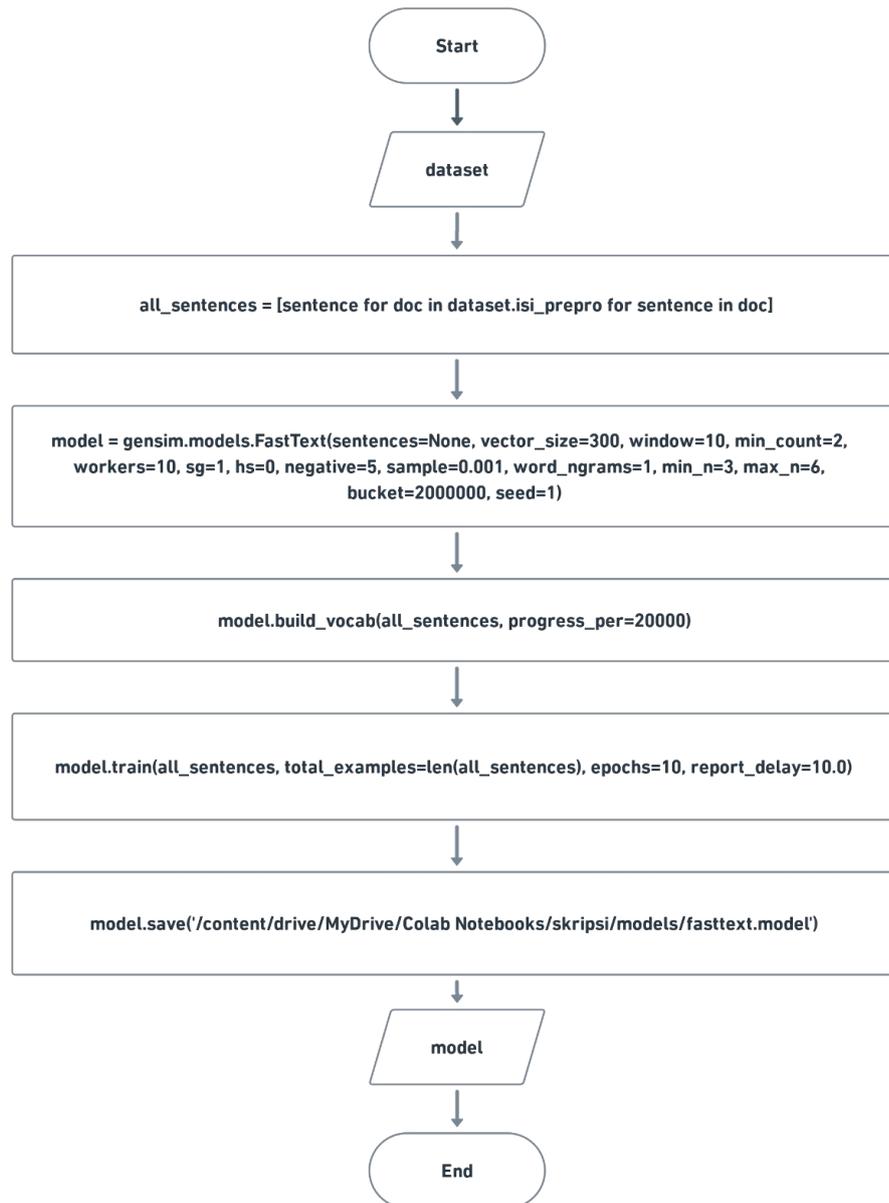
Hasil *clustering* pada tahap inilah yang akan dijadikan untuk pembentukan sebuah paragraf. Dimana anggota dalam tiap paragraf adalah dokumen yang berada pada posisi tiap *cluster*. Sehingga pada contoh sederhana perhitungan *cluster* diatas akan menghasilkan 2 paragraf hasil ringkasan. Paragraf 1 terdiri dari berita 1 sedangkan paragraf 2 terdiri dari berita 2 dan berita 3.

3.5. Ringkasan Akhir

Tahap selanjutnya yaitu dilakukan ringkasan akhir. Masukan yang dibutuhkan pada tahap Ringkasan Akhir adalah multi dokumen yang didapat dari dokumen hasil *clustering* pada tahap sebelumnya. Adapun proses selanjutnya sebelum mendapatkan hasil akhir adalah:

3.5.1. *FastText*

Setelah dilakukan *initial summary* dimana multi dokumen telah dibentuk maka selanjutnya akan dilakukan proses pembentukan model *FastText*. *Flowchart* proses *training FastText* dapat dilihat pada Gambar 3.7 . Model ini akan mengambil serta mengukur konteks dari kata kata tersebut, *FastText* menggunakan Teknik *embedding* kata yang sama dengan *word2vec* namun juga akan menambahkan representasi kata dalam bentuk *n-gram* atau *substring* dari teks. Pada penelitian ini model *FastText* dibuat dengan bantuan *library* dari *gensim*.



Gambar 3.8 Flowchart training model *FastText*

Langkah melakukan training model *FastText* menggunakan data dari tiap kalimat multi dokumen adalah sebagai berikut :

1) Menentukan Parameter Model

Tahap ini adalah awal dari pembentukan model. Pada tahap ini, akan ditentukan nilai parameter. Ada 14 parameter yang akan diinisiasi di awal. Contoh pemberian

nilai pada tiap parameter ditunjukkan pada Tabel 3.4 :

Tabel 3. 4 Parameter *FastText* pada Gensim

Parameter	Nilai	Keterangan
sentences	None	Kalimat tidak diisikan pada pembentukan parameter model agar model tidak langsung melakukan proses pelatihan data.
vector_size	300	Jumlah fitur dalam vektor kata
window	5	Ukuran berapa konteks yang akan diambil pada tiap arahnya
min_count	2	Jumlah minimal kata yang harus muncul pada seluruh kalimat
workers	10	Jumlah <i>thread</i> pelatihan
sg	0	Jika diinisiasikan nilai 0 maka arsitektur yang akan digunakan adalah arsitektur cbow. Namun jika 1 maka akan menggunakan arsitektur skipgram.
hs	0	parameter hs adalah singkatan dari hierarchical softmax. Parameter ini digunakan untuk menentukan metode yang digunakan untuk menghitung peluang kata berikutnya pada saat melatih model. Jika nilai parameter hs diset 1, maka hierarchical softmax akan digunakan dalam pelatihan model. Jika nilainya 0, metode negatif sampling akan digunakan sebagai gantinya.
negative	5	Jika nilai parameter negative diatur pada nilai yang lebih besar dari nol, maka metode negatif sampling akan digunakan selama pelatihan model. Nilai parameter ini menentukan jumlah kata yang dipilih secara acak dari kosakata sebagai kata negatif selama pelatihan. Semakin besar nilai negative, semakin banyak kata negatif yang dipilih dan semakin sedikit waktu yang dibutuhkan untuk melatih model.
sample	1e-3	Nilai parameter sample menentukan threshold untuk melakukan subsampling pada kata-kata yang sering muncul dalam korpus. Tujuan dari subsampling adalah untuk mengurangi jumlah kata yang sering muncul dalam korpus dan meningkatkan peran kata-kata yang lebih jarang muncul dalam pembentukan vektor kata. Nilai threshold biasanya diatur antara 0 dan 1. Sebagai contoh, jika kita memiliki nilai $sample=1e-3$, maka setiap kata yang muncul dalam teks dengan frekuensi yang lebih rendah dari $1e-3$ akan disubsampling selama pelatihan model.

Parameter	Nilai	Keterangan
word_ngrams	1	Secara default, nilai "word_ngrams" diatur sebagai 1, yang berarti hanya unigram (kata-kata tunggal) yang digunakan untuk membentuk vektor kata. Namun, dengan menambahkan nilai parameter "word_ngrams" menjadi 2 atau lebih, model <i>FastText</i> dapat memperluas cakupan representasi vektor kata dengan memasukkan n-gram lebih panjang ke dalam vektor kata.
min_n	3	Panjang minimum karakter n-gram yang akan digunakan untuk melatih representasi kata
max_n	6	Panjang maksimum karakter yang akan digunakan untuk melatih representasi kata.
bucket	2000000	karakter n-grams memerlukan ruang memori yang cukup besar untuk menyimpan informasi frekuensi karakter n-grams untuk setiap kata dalam dataset. Oleh karena itu, karakter n-grams di-hash ke dalam sejumlah bucket yang tetap untuk membatasi penggunaan memori oleh model. Parameter "bucket" mengontrol jumlah bucket yang digunakan oleh model <i>FastText</i> dalam proses hash karakter n-grams tersebut.

2) Pembentukan Kosakata

Pada Gensim fungsi *build_vocab()* digunakan untuk membangun kosakata atau vocabulary dari model yang akan dilatih diinisialisasikan parameter sebelumnya. Proses membangun kosakata ini melibatkan penghitungan frekuensi kemunculan setiap kata dalam dataset, serta penghapusan kata-kata yang tidak memenuhi syarat (misalnya, kata-kata yang terlalu jarang muncul dalam dataset). Setelah kosakata dibangun, model akan menetapkan representasi vektor numerik untuk setiap kata dalam kosakata. Representasi ini akan digunakan dalam proses training model, di mana model akan mempelajari bagaimana mengelompokkan dan menggabungkan

vektor kata untuk menghasilkan representasi vektor untuk kalimat atau teks yang lebih panjang.

3) Training Model

Pada Gensim fungsi *train()* digunakan untuk melatih model dengan dataset yang telah diolah sebelumnya. "*all_sentences*" pada program tersebut adalah kumpulan kalimat atau teks yang akan digunakan untuk melatih model. Sementara itu, parameter "*epochs*" digunakan untuk menentukan berapa kali model akan diberikan kesempatan untuk melihat setiap kalimat atau teks dalam dataset pada setiap iterasi. Dalam proses training model, *FastText* akan mempelajari representasi vektor untuk setiap kata dalam kosakata dan mengelompokkannya berdasarkan konteks dan makna yang serupa. Hasil akhirnya adalah setiap kata akan direpresentasikan oleh vektor numerik tertentu dalam ruang vektor yang lebih besar.

4) Menyimpan Model

Untuk menyimpan model hasil training pada *Gensim FastText*, dapat menggunakan fungsi *save()* yang disediakan oleh *Gensim*. model akan disimpan ke dalam file dengan nama "*model.bin*". Untuk memuat kembali model yang telah disimpan, dapat menggunakan fungsi *load()*.

3.5.2. Vektor kalimat

Pada tahap ini kalimat dalam tiap dokumen akan diubah menjadi vektor. Sebelum membentuk vektor kalimat maka akan diambil nilai vektor kata terlebih dahulu. Vektor kata dibentuk dari model *FastText* yang sebelumnya telah dibuat. Alur proses dari algoritma pembentukan kata menjadi vektor kalimat dibentuk dengan

cara melakukan penjumlahan vektor kata dan membagi jumlah kata dalam kalimat tersebut.

3.5.3. *Clustering* Kalimat

Tahap selanjutnya adalah clustering pada vektor kalimat. *Clustering* pada tahap ringkasan akhir berbeda dengan *clustering* pada tahap ringkasan awal. *Clustering* pada ringkasan akhir mengelompokkan kalimat untuk mendapatkan kalimat terbaik yang akan dijadikan ringkasan. Tahap ini bertujuan untuk mengelompokkan kalimat mana saja yang memiliki kedekatan jarak antar vektor. Nilai K ditentukan sebanyak 0,5 atau separuh dari panjang kalimat. Adapun tahapan dalam proses *K-means Clustering* adalah:

1) Data input

Data input berasal dari kalimat yang telah diubah menjadi vektor kalimat.

Misal terdapat 3 kalimat yang telah diubah menjadi vektor yaitu:

- a. Kalimat 1: [0.1, 0.3, 0.2, 0.1, 0.3]
- b. Kalimat 2: [0.7, 0.9, 0.6, 0.6, 0.5]
- c. Kalimat 3: [0.1, 0.2, 0.1, 0.3, 0.4]

2) Inisiasi *Centroid* Awal

Jumlah k pada algoritma *K-means Clustering* ditentukan diawal. Misal k=2. Maka akan di inisiasikan nilai *centroid* sebanyak 2 buah. Nilai *centroid* ditentukan secara acak.

- a. *Centroid* 1: [0.2, 0.1, 0.2, 0.1, 0.1]
- b. *Centroid* 2: [0.7, 0.8, 0.9, 0.6, 0.6]

3) Menghitung jarak *centroid*

Menghitung jarak setiap data dengan setiap pusat *centroid*, dan memasukkan data ke dalam *cluster* yang memiliki jarak terdekat. Cara menghitung jarak antara dua vektor dapat menggunakan algoritma *Euclidean Distance*. Berikut adalah hasil perhitungan untuk setiap data. Tiap kalimat akan dihitung jaraknya dengan *centroid*.

- a. Kalimat 1 Jarak ke *centroid* 1 = 0.36, Jarak ke *centroid* 2 = 1.2
- b. Kalimat 2 Jarak ke *centroid* 1 = 0.14, Jarak ke *centroid* 2 = 0.22
- c. Kalimat 3 Jarak ke *centroid* 1 = 0.46, Jarak ke *centroid* 2 = 1.72

Dari hasil perhitungan di atas, dapat dilihat bahwa kalimat pertama dan kedua termasuk ke dalam *cluster* 2, sedangkan kalimat ketiga termasuk ke dalam *cluster* 1.

4) Memperbarui *centroid*

Memperbarui pusat *centroid* untuk setiap *cluster* dengan mengambil nilai rata-rata dari setiap data dalam *cluster*. Berikut adalah hasil perhitungan untuk setiap pusat *centroid* baru:

- a. *Centroid* 1: [0.1, 0.25, 0.15, 0.2, 0.35]
- b. *Centroid* 2: [0.7, 0.85, 0.75, 0.6, 0.55]

5) Hasil Clustering

Karena tidak ada perubahan pada setiap pusat *centroid*, maka dapat dianggap bahwa *clustering* sudah konvergen dan menghentikan iterasi. Sehingga hasil akhir dari *clustering* adalah:

- a. Kalimat 1 terdapat pada *cluster* 1
- b. Kalimat 2 terdapat pada *cluster* 2

c. Kalimat 3 terdapat pada *cluster* 1

3.5.4. Pemilihan kalimat

Dari hasil *cluster* tersebut maka kalimat ringkasan diambil menggunakan persamaan berikut.

$$r_i = \operatorname{argMin} \left(\sqrt{\sum_{j=1}^n (x_j - c_i)^2}, x \in C \right) \quad (3.1)$$

Hasil dari perhitungan r_i merupakan kalimat ringkasan dengan jarak terdekat pada *centroid* pada *cluster* tersebut. Sehingga yang akan diambil menjadi sebuah ringkasan pada dokumen tersebut adalah r_1 hingga r_n dengan menggunakan persamaan 3.2.

$$\text{hasil ringkasan} = r_1, r_2, \dots, r_n \quad (3.2)$$

Alur dalam memilih kalimat dimana kalimat yang diambil berasal dari kalimat yang paling dekat dengan *centroid* tiap *clusternya*. Sehingga jika sebelumnya telah ditentukan nilai $K = 2$. Maka kalimat hasil ringkasan jika berjumlah sama dengan K . Yang diharapkan dari proses ini adalah adanya representasi informasi dari setiap *clusternya*. Sehingga dokumen menjadi lebih ringkas karena menghilangkan informasi yang tidak diperlukan.

3.5.5. Mengurutkan Kalimat

Proses pengurutan kalimat adalah dengan menyesuaikan urutan dalam dokumen. Jika hasil dari pemilihan kalimat adalah kalimat 3 dan 1. Maka akan

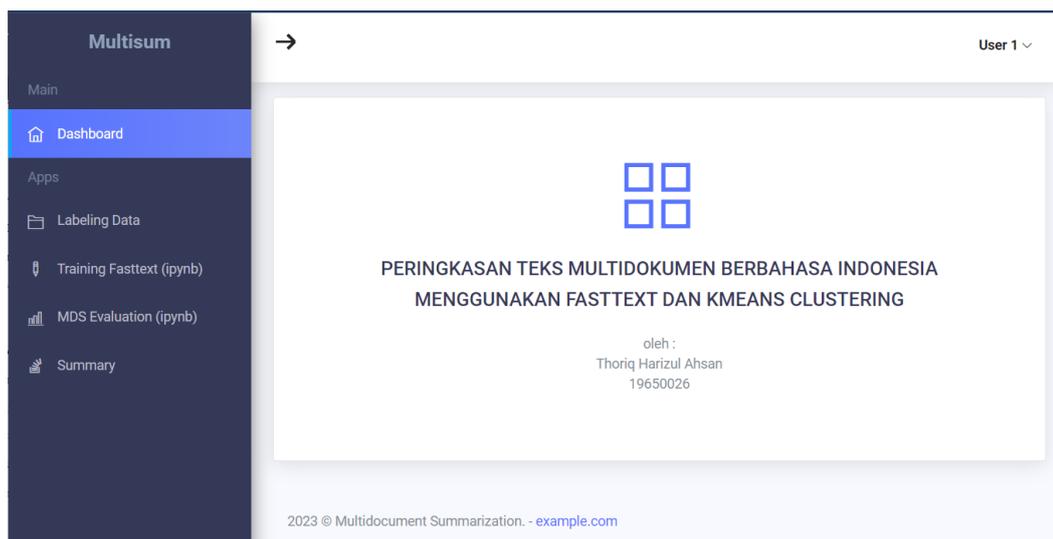
diurutkan berdasarkan urutan dokumen awalnya. Sehingga urutannya akan menjadi kalimat 1 dan 3.

3.5.6. Hasil Ringkasan

Hasil ringkasan merupakan nilai *index* yang disimpan pada sebuah list. Kemudian dari nilai index akan diambil kalimat asli dari dokumen yang belum melalui tahap *preprocessing*. Sehingga pada akhirnya kalimat ringkasan bisa dibaca dengan mudah oleh manusia.

3.6 Implementasi Sistem

Implementasi sistem peringkasan teks berita multi dokumen ditampilkan pada Gambar 3. 9 dan Gambar 3. 10.



Gambar 3. 9 Implementasi Sistem

Multidokumen Summarization

Jumlah Berita

Next

Berita 1

Judul Berita

Isi Berita

Berita 2

Judul Berita

Isi Berita

Tentukan Presentase Hasil Ringkasan yang diinginkan:

Summarize

Percentage: 0.5

Gambar 3.10 Tampilan Sistem Peringkasan Multi dokumen

Pada Gambar 3.10 dan 3.11 ini merupakan tampilan dari sistem yang dibuat sebagai *user interface*. Sistem peringkasan terdapat satu input utama yaitu berapa berita yang akan diringkaskan. Setelah dimasukkan jumlahnya maka berita

baru bisa diringkas dengan memasukkan judul dan isi tiap berita serta persentase berapa banyak hasil ringkasan yang diinginkan.

Sistem yang diimplementasikan selain pada tampilan user interface juga dapat dilihat dari vektor kalimat hasil pembentukan *FastText*. Pada tahap ini penulis telah membuat model *FastText* untuk mengubah kata menjadi vektor setelah itu mengubah vektor kata menjadi vektor kalimat berdasarkan langkah-langkah yang sudah dijelaskan pada bab sebelumnya. Selanjutnya, pada penelitian ini hasil dari proses training model *FastText* menghasilkan vektor kata yang kemudian dijumlahkan sehingga menjadi vektor kalimat. Hasil penjumlahan vektor kata menghasilkan total kalimat yang bervariasi pada setiap multi dokumen. Tabel 4.1 menampilkan contoh hasil penjumlahan vektor kata menjadi vektor kalimat untuk 5 dokumen berita sesuai dengan kategorinya.

Tabel 3.1 Hasil *FastText* untuk Vektor Kalimat dalam 3 dokumen

Judul Berita	Kalimat dari kumpulan kata	Vektor Kalimat
Berita 1	['hp', 'luncur', 'buah', 'printer', 'portable', 'cetak', 'tinta', 'nama', 'sprocket']	[-1.10004870e-02, -1.09445890e-02, 4.77426685e-02, -1.01335323e-03, ..., -5.44904312e-03, 1.70447119e-03, -1.59098655e-02, 1.07713677e-02]
	['perangkat', 'evolusi', 'kamera', 'instan', 'nama', 'pas', 'sprocket', 'in']	[-5.49651869e-03, -3.95996636e-03, 5.12697101e-02, -3.37801379e-04, ..., -1.92619748e-02, 1.43014989e-03, -1.66713875e-02, 9.41861235e-03]
	['desain', 'kerja', 'polaroid', 'snap', 'kodak', 'printomatic']	[-0.00923658, -0.00432744, 0.04932618, 0.00166672, 0.0119482, ..., 0.01160746, -0.02789092, 0.00959491, -0.01158937, 0.0033069],

	['dailysocialid', 'portal', 'berita', 'startup', 'inovasi', 'teknologi']	[-2.68852594e-03, -2.67587020e-03, 4.23590019e-02, -6.15909044e-03, ..., -2.34648604e-02, 3.88563797e-03, -2.07106620e-02, 5.01547335e-03]
['member', 'komunitas', 'startup', 'inovasi', 'dailysocialid', 'unduh', 'lapor', 'riset', 'statistik', 'putar', 'teknologi', 'cumacuma', 'berita', 'startup', 'indonesia', 'gadget']	[-3.97414854e-03, -8.70961696e-03, 4.62585203e-02, -8.87752231e-03, ..., -1.20057249e-02, 5.31037105e-03, -2.20842063e-02, 5.22575434e-03]	

Berita 2	['printer', 'portable', 'solusi', 'tarik', 'abadi', 'momenmomen', 'tangkap', 'wujud', 'fisik', 'tarik', 'kamera', 'instan']	[-6.43846393e-03, -8.80945381e-03, 4.77270596e-02, -4.68955841e-03, ..., -2.01807488e-02, 9.29071009e-03, -1.63016338e-02, 3.64672509e-03]
	['printer', 'portable', 'praktis', 'kamera', 'instan', 'hasil', 'cetak', 'bagus', 'kreativitas', 'batas', 'kamera']	[-0.00429473, -0.01262526, 0.04593062, -0.00274845, 0.01483011, ..., -0.0003984, -0.01075705, 0.00292422, -0.02166967, 0.00608309]
	['canon', 'selphy', 'salah', 'lini', 'populer', 'kategori', 'canon', 'kenal', 'model', 'selphy', 'cp']	[-0.00896886, -0.00150259, 0.04338594, 0.00806761, 0.00943622, ..., 0.01339199, -0.02402475, -0.01139156, -0.00265159, 0.00903264]
	['banding', 'cp', 'bawa', 'deret', 'signifikan', 'utama', 'tiltable', 'lcd', 'inci', 'tampil', 'antarmuka', 'mudah', 'navigasi']	[-0.0076207, -0.01387606, 0.05117436, -0.00535081, 0.02009884, ..., 0.00134514, -0.02006882, -0.0001241, -0.02015315, 0.00433696]

	['dailysocialid', 'portal', 'berita', 'startup', 'inovasi', 'teknologi']	[-0.5896456042925516, -0.2884650652607282, '...', 0.17623058954874674, -0.2884650652607282]
['member', 'komunitas', 'startup', 'inovasi', 'dailysocialid', 'unduh', 'lapor', 'riset', 'statistik', 'putar', 'teknologi', 'cumacuma', 'berita', 'startup', 'indonesia', 'gadget']	[-0.4718394731171429, -0.19495091238059103, '...', 0.17318982040160336, -0.19495091238059103]	
Berita 3	['polaroid', 'rupa', 'nama', 'populer', 'pakai', 'kamera', 'instan', 'brand', 'berpindahpindah', 'milik', 'fujifilm', 'jasa', 'jaga', 'tradisi', 'hidup', 'keluarga', 'kamera', 'instax']	[-3.42518883e-03, -1.35567300e-02, 5.01739569e-02, -3.21674161e-03, ..., -2.66185738e-02, 7.87369162e-03, -1.24964798e-02, -1.58137293e-03]
	['perangkat', 'fotografi', 'fujifilm', 'sedia', 'film', 'desain', 'frame', 'unik', 'printer', 'portable', 'cetak', 'foto', 'smartphone']	[-9.81512852e-03, -6.80393679e-03, 5.19396327e-02, -4.53754468e-03, ..., -2.44898126e-02, 8.91688559e-03, -1.34013863e-02, 3.86852492e-03]
	['brand', 'kaya', 'intelektual', 'polaroid', 'milik', 'pegang', 'saham', 'impossible', 'project', 'kecimpung', 'ranah', 'sedia', 'film', 'instan']	[-0.00595163, -0.00884553, 0.04592818, 0.00099045, 0.01913441, ..., 0.00679632, -0.024128, 0.00023122, -0.01251229, 0.00269545]

	['via', 'digital', 'trends', 'the', 'photo', 'blographer']	[-1.15355300e-02, -7.37082120e-03, 3.89672890e-02, -1.26074841e-02, ..., [-1.15355300e-02, -7.37082120e-03, 3.89672890e-02, -1.26074841e-02,
	['dailysocialid', 'portal', 'berita', 'startup', 'inovasi', 'teknologi']	[-2.68852594e-03, -2.67587020e-03, 4.23590019e-02, -6.15909044e-03, ..., -2.34648604e-02, 3.88563797e-03, -2.07106620e-02, 5.01547335e-03]
	['member', 'komunitas', 'startup', 'inovasi', 'dailysocialid', 'unduh', 'lapor', 'riset', 'statistik', 'putar', 'teknologi', 'cumacuma', 'berita', 'startup', 'indonesia', 'gadget']	[-3.97414854e-03, -8.70961696e-03, 4.62585203e-02, -8.87752231e-03, ..., -1.20057249e-02, 5.31037105e-03, -2.20842063e-02, 5.22575434e-03]

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Skenario Uji Coba

Data multi dokumen artikel berita *online* Bahasa Indonesia yang berjumlah 30 data multi dokumen akan dilakukan *preprocessing* untuk memperbaiki struktur inputan data. Setelah itu dilakukan proses perubahan dokumen menjadi vektor menggunakan *Bag of Words*. Selanjutnya dilakukan clustering pertama untuk mengelompokkan dokumen menjadi multi dokumen lalu dilakukan pembuatan model *FastText* untuk mengubah kata menjadi vektor.

Percobaan dilakukan dengan mengubah kalimat menjadi vektor kalimat berdasarkan hasil penjumlahan vektor kata pada setiap multi dokumen yang telah dikelompokkan (*cluster*) sesuai dengan kemiripan antar dokumen. Setelah itu hasil pengelompokkan kalimat pada multi dokumen akan dibandingkan dengan 4 variasi label responden dan label IndoSum yang telah dibentuk sebelumnya. 4 variasi label tersebut akan digunakan sebagai dasar dalam menentukan ringkasan manual. Perhitungan untuk membandingkan hasil ringkasan manual dari dataset dengan hasil ringkasan sistem dilakukan menggunakan metode ROUGE dengan variasi ROUGE-1. Variasi ROUGE-N akan mengukur sebanyak N kata secara berurutan.

Rouge merupakan salah satu matriks evaluasi yang umum dan sering digunakan. Rouge bekerja dengan membandingkan hasil dari teks buatan mesin dengan teks buatan manusia (Akbar, 2021). Rouge memiliki banyak ragam yang berbeda. Rouge yang populer dan sering digunakan adalah Rouge-N, Rouge-L, Rouge-S, dan Rouge-W (Lin, 2004). Abdullah pada tahun 2018 telah melakukan

penelitian mengenai peringkasan teks berita multi dokumen menggunakan pendekatan fitur berita serta informasi gramatikal (NeFGIS) dan *part of speech tagging* (POS Tagging). Penulis bertujuan untuk melakukan pembobotan kalimat sehingga dapat memilih kalimat representatif yang lebih baik dan meningkatkan kualitas hasil peringkasan teks berita. Metode evaluasi yang digunakan dalam penelitian ini menggunakan ROUGE dengan 4 varian yaitu ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-SU4. Pada percobaan sistem penulis membandingkan hasil ringkasan menggunakan metode pembobotan NeFGIS dengan metode pembobotan NeFTis (*trending issue*). Hasil pengujian didapatkan metode pembobotan NeFGIS memiliki hasil ringkasan yang paling baik dengan berdasarkan evaluasi ROUGE rata-rata nilai yang didapatkan untuk *Recall* yaitu 0.03570 (19,67%), *Precision* yaitu 0.09722 (38,13%) dan *F-Score* sebesar 0.06072 (27,26%) (Abdullah & Fatichah, 2018). Adapun perhitungan nilai *Recall*, *Precision*, dan *Rouge* sebagai berikut (Verma, 2019):

4.1.1 Recall

Recall dihitung menggunakan persamaan 4.1 dimana s adalah kalimat atau *sentence* yang merupakan elemen dari peringkasan manual (Sum_{ref}), n adalah panjang dari n -gram, $Count_{match}(gram_n)$ adalah jumlah n -gram yang muncul pada peringkasan sistem dan peringkasan manual. Sedangkan $Count(gram_n)$ adalah jumlah n -gram yang ada pada peringkasan manual.

$$Recall = \frac{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s \in Sum_{ref}} \sum_{gram_n \in s} Count(gram_n)} \quad 4.1$$

4.1.2 Precision

Precision dihitung menggunakan persamaan 4.2 dimana s adalah kalimat atau *sentence* yang merupakan elemen dari peringkasan sistem (Sum_{sys}), n adalah panjang dari n -gram, $Count_{match}(gram_n)$ adalah jumlah n -gram yang muncul pada peringkasan sistem dan peringkasan manual. Sedangkan $Count(gram_n)$ adalah jumlah n -gram yang ada pada peringkasan sistem.

$$Precision = \frac{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count(gram_n)} \quad 4.2$$

4.1.3 F-Score

F-Score dihitung dengan menggabungkan nilai *Precision* dan *Recall* dari sistem temu kembali informasi menjadi sebuah skor tunggal. Untuk nilai β yang non-negatif real ($0 \leq \beta < \infty$), *F-Score* dapat dihitung dengan rumus :

$$F_{\beta} = \frac{(1 + \beta^2)(Precision * recall)}{(\beta^2 * Precision + recall)} \quad 4.3$$

4.2 Hasil Uji Coba

Pada skenario uji coba yang telah dijelaskan pada sub bab 4.1 pengujian dilakukan dengan membandingkan hasil peringkasan manual dengan hasil peringkasan sistem. Tabel 4.2 menampilkan hasil peringkasan manual yang terdapat pada referensi dataset dan hasil peringkasan sistem berdasarkan variasi 1 tanpa menggunakan label IndoSum. untuk kategori teknologi dengan dengan kata kunci “printer”, “portable”.

Tabel 4.2 Hasil Uji Coba

<p>Konten Berita 1</p>	<p>Tahun lalu , HP meluncurkan sebuah printer portable yang dapat mencetak tanpa tinta bernama Sprocket .Baru - baru ini , perangkat tersebut telah berevolusi menjadi kamera instan , dengan nama yang pas yaitu Sprocket 2 - in - 1 .Desain dan cara kerjanya sangat mirip seperti Polaroid Snap maupun Kodak Printomatic .Bidik lewat viewfinder , tekan tombol shutter , lalu hasilnya dapat langsung dicetak di atas kertas ZINK berukuran 2 x 3 cm .Bedanya , Sprocket 2 - in - 1 masih bisa berfungsi sebagai printer biasa .Hal ini berarti pengguna masih bisa menyambungkan ponselnya via Bluetooth , mengedit foto yang diinginkan , lalu mencetaknya menggunakan perangkat ini .Yang cukup menarik , HP tidak lupa menambahkan bumbu AR pada aplikasi pendamping Sprocket di Android maupun iOS .Fitur tersebut memungkinkan pengguna untuk menyelipkan foto dan video tersembunyi pada gambar yang dicetak .Jadi ketika gambar tersebut dilihat menggunakan kamera smartphone , akan muncul sejumlah foto dan video yang diharapkan bisa membantu pengguna ‘ menghidupi ’ kembali momen berkenang yang ada pada gambar .Bersamaan dengan itu , HP juga mengumumkan Sprocket Plus .Model ini pada dasarnya sama seperti Sprocket standar yang diluncurkan tahun lalu , hanya saja hasil cetaknya 30 % lebih besar sehingga akan terasa lebih ideal untuk foto grup .HP Sprocket 2 - in - 1 saat ini sudah dipasarkan seharga \$ 160 , sedangkan Sprocket Plus sedikit lebih murah di angka \$ 150 .Keduanya hadir dalam tiga pilihan warna : putih , hitam atau merah .Sumber : The Verge .DailySocial.id adalah portal berita startup dan inovasi teknologi .Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id , mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma , dan mengikuti berita startup Indonesia dan gadget terbaru .</p>
<p>Konten Berita 2</p>	<p>Printer portable adalah solusi yang menarik bagi mereka yang ingin mengabadikan momen-momen yang mereka tangkap ke dalam wujud fisik , tapi tidak tertarik dengan kamera instan .Printer portable memang tidak sepraktis kamera instan , tapi hasil cetaknya jelas lebih bagus , dan lagi kreativitas Anda tidak terbatas oleh satu kamera saja .Canon Selphy adalah salah satu lini yang populer di kategori ini , dan Canon baru saja memperkenalkan model baru , yakni Selphy CP1300 .Dibandingkan pendahulunya , CP1300 membawa sederet pembaruan yang signifikan , utamanya adalah tiltable LCD yang lebih besar (3,2 inci) dengan tampilan antarmuka yang lebih mudah dinavigasikan .Tombol Wi - Fi terpisah memudahkan proses setup jaringan sehingga pengguna dapat segera mencetak foto yang diambil menggunakan smartphone maupun kamera .Namun masih ada fitur yang lebih menarik lagi , namanya Party Shuffle .Fitur ini memungkinkan beberapa pengguna sekaligus untuk menyambungkan perangkatnya masing-masing ke printer , lalu mencetak satu gambar yang disusun dari beberapa foto yang diambil menggunakan perangkat yang berbeda-beda itu tadi .Bagi yang hendak menciptakan photo booth dadakan di sebuah acara pesta , tersedia pula opsi untuk mencetak photo strip berukuran 2 x 6 inci .Baru juga untuk Canon Selphy CP1300 adalah kemampuan menyambung ke battery pack eksternal yang membuatnya semakin portable dibanding pendahulunya .Perangkat ini rencananya bakal dipasarkan seharga \$ 130 , dengan pilihan warna hitam atau putih .Sumber : DPReview .DailySocial.id adalah portal berita startup dan inovasi teknologi .Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id , mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma , dan mengikuti berita startup Indonesia dan gadget terbaru .</p>
<p>Konten Berita 3</p>	<p>Polaroid memang merupakan nama yang mempopulerkan pemakaian kamera instan , namun saat brand ini berpindah-pindah pemilik , Fujifilm berjasa menjaga tradisi tersebut tetap hidup lewat keluarga kamera Instax .Selain perangkat fotografi , Fujifilm turut menyediakan film dengan berbagai desain frame unik , serta printer portable untuk mencetak foto dari smartphone .Brand</p>

	<p>dan kekayaan intelektual Polaroid sendiri kini sudah menjadi milik pemegang saham terbesar Impossible Project , yang sejak tahun 2008 berkecimpung di ranah penyediaan film instan .Di bulan September kemarin , Impossible Project berubah nama jadi Polaroid Originals , dan untuk merayakan ‘ kembalinya’ brand tersebut , mereka memperkenalkan produk bernama OneStep 2 .Kembalinya Polaroid ke ranah ini sepertinya memperuncing kompetisi dengan Fujifilm , dan bertransformasi jadi sengketa setelah Polaroid beberapa kali mengirimkan surat peringatan penghentian produksi pada Fujifilm terkait produk baru mereka .Perusahaan fotografi asal Jepang itu menanggapi dengan meminta bantuan Pengadilan AS .Akar masalah perseteruan ini sebenarnya cukup sepele , yaitu terkait format film instan .Perkara ini dimulai tak lama selepas Photokina 2016 .Di acara itu , Fujifilm memperkenalkan film Instax Square berformat persegi 62×62 - milimeter , meluncurkannya di awal 2017 , berbarengan dengan kamera Instax Square SQ10 .Namun tak lama , Polaroid memperingatkan Fujifilm bahwa merek dagang film ini dimiliki oleh Polaroid .Karena Fujifilm tidak segera merespons , Polaroid mengancam buat ‘ mengambil tindakan tegas’ .Meski Polaroid yang pertama mengajukan keluhan , Fujifilm - lah yang mengubahnya jadi kasus hukum melalui pengajuan laporan ke Pengadilan Distrik Selatan kota New York minggu lalu .Dalam surat keluhan ketiga di bulan Juni 2017 , Polaroid menuntut agar Fujifilm membayar royalti / lisensi penggunaan film instan dengan format persegi .Dokumen pengadilan dari Fujifilm menyatakan : ‘ karena tidak sanggup memperoleh keuntungan lewat penjualan produk’ , Polaroid kini ‘ mencoba menciptakan pemasukan melalui sisa-sisa portofolio kekayaan intelektual Polaroid’ .Menurut Fujifilm , Instax Square tidak melanggar trademark Polaroid .Mengulik lebih dalam lagi , format bingkai persegi tak sekedar dipilih agar foto tampil unik .Di frame itulah produsen menempatkan zat - zat kimia penting buat memunculkan hasil jepretan di film .Perlu diketahui bahwa saat Fujifilm merilis film persegi , Polaroid belum memproduksi film berformat serupa setelah mereka menghentikan proses pembuatannya di tahun 2008 .Dan hingga kini , dua perusahaan itu belum menampakkan diri di pengadilan .Via Digital Trends & The Photo Blographer .DailySocial.id adalah portal berita startup dan inovasi teknologi .Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id , mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma , dan mengikuti berita startup Indonesia dan gadget terbaru.</p>
<p>Peringkasan Manual</p>	<p>Tahun lalu, HP meluncurkan sebuah printer portable yang dapat mencetak tanpa tinta bernama Sprocket. Baru - baru ini, perangkat tersebut telah berevolusi menjadi kamera instan, dengan nama yang pas yaitu Sprocket 2 - in - 1. Bedanya, Sprocket 2 - in - 1 masih bisa berfungsi sebagai printer biasa. Yang cukup menarik, HP tidak lupa menambahkan bumbu AR pada aplikasi pendamping Sprocket di Android maupun iOS. HP Sprocket 2 - in - 1 saat ini sudah dipasarkan seharga \$ 160, sedangkan Sprocket Plus sedikit lebih murah di angka \$ 150. Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id, mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma, dan mengikuti berita startup Indonesia dan gadget terbaru.</p> <p>Printer portable adalah solusi yang menarik bagi mereka yang ingin mengabadikan momen-momen yang mereka tangkap ke dalam wujud fisik, tapi tidak tertarik dengan kamera instan. Canon Selphy adalah salah satu lini yang populer di kategori ini, dan Canon baru saja memperkenalkan model baru, yakni Selphy CP1300. Tombol Wi - Fi terpisah memudahkan proses setup jaringan sehingga pengguna dapat segera mencetak foto yang diambil menggunakan smartphone maupun kamera. Bagi yang hendak menciptakan photo booth dadakan di sebuah acara pesta, tersedia pula opsi untuk mencetak photo strip berukuran 2 x 6 inci.</p>

	<p>Polaroid memang merupakan nama yang mempopulerkan pemakaian kamera instan, namun saat brand ini berpindah-pindah pemilik, Fujifilm berjasa menjaga tradisi tersebut tetap hidup lewat keluarga kamera Instax. Selain perangkat fotografi, Fujifilm turut menyediakan film dengan berbagai desain frame unik, serta printer portable untuk mencetak foto dari smartphone. Kembalinya Polaroid ke ranah ini sepertinya memperuncing kompetisi dengan Fujifilm, dan bertransformasi jadi sengketa setelah Polaroid beberapa kali mengirimkan surat peringatan penghentian produksi pada Fujifilm terkait produk baru mereka. Akar masalah perseteruan ini sebenarnya cukup sepele, yaitu terkait format film instan. Kamu bisa menjadi member komunitas startup dan inovasi DailySocial.id, mengunduh laporan riset dan statistik seputar teknologi secara cuma-cuma, dan mengikuti berita startup Indonesia dan gadget terbaru.</p>
<p>Peringkasan Sistem</p>	<p>Tahun lalu, HP meluncurkan sebuah printer portable yang dapat mencetak tanpa tinta bernama Sprocket. Desain dan cara kerjanya sangat mirip seperti Polaroid Snap maupun Kodak Printomatic. Bedanya, Sprocket 2 - in - 1 masih bisa berfungsi sebagai printer biasa. Yang cukup menarik, HP tidak lupa menambahkan bumbu AR pada aplikasi pendamping Sprocket di Android maupun iOS. Fitur tersebut memungkinkan pengguna untuk menyelipkan foto dan video tersembunyi pada gambar yang dicetak. Model ini pada dasarnya sama seperti Sprocket standar yang diluncurkan tahun lalu, hanya saja hasil cetaknya 30 % lebih besar sehingga akan terasa lebih ideal untuk foto grup. Printer portable memang tidak sepraktis kamera instan, tapi hasil cetaknya jelas lebih bagus, dan lagi kreativitas Anda tidak terbatas oleh satu kamera saja. Canon Selphy adalah salah satu lini yang populer di kategori ini, dan Canon baru saja memperkenalkan model baru, yakni Selphy CP1300. Dibandingkan pendahulunya, CP1300 membawa sederet pembaruan yang signifikan, utamanya adalah tiltable LCD yang lebih besar (3,2 inci) dengan tampilan antarmuka yang lebih mudah dinavigasikan. Tombol Wi - Fi terpisah memudahkan proses setup jaringan sehingga pengguna dapat segera mencetak foto yang diambil menggunakan smartphone maupun kamera. Bagi yang hendak menciptakan photo booth dadakan di sebuah acara pesta, tersedia pula opsi untuk mencetak photo strip berukuran 2 x 6 inci. Perangkat ini rencananya bakal dipasarkan seharga \$ 130, dengan pilihan warna hitam atau putih. Polaroid memang merupakan nama yang mempopulerkan pemakaian kamera instan, namun saat brand ini berpindah-pindah pemilik, Fujifilm berjasa menjaga tradisi tersebut tetap hidup lewat keluarga kamera Instax. Selain perangkat fotografi, Fujifilm turut menyediakan film dengan berbagai desain frame unik, serta printer portable untuk mencetak foto dari smartphone. Kembalinya Polaroid ke ranah ini sepertinya memperuncing kompetisi dengan Fujifilm, dan bertransformasi jadi sengketa setelah Polaroid beberapa kali mengirimkan surat peringatan penghentian produksi pada Fujifilm terkait produk baru mereka. Akar masalah perseteruan ini sebenarnya cukup sepele, yaitu terkait format film instan. Karena Fujifilm tidak segera merespons, Polaroid mengancam buat ‘ mengambil tindakan tegas’. Meski Polaroid yang pertama mengajukan keluhan, Fujifilm - lah yang mengubahnya jadi kasus hukum melalui pengajuan laporan ke Pengadilan Distrik Selatan kota New York minggu lalu. Dalam surat keluhan ketiga di bulan Juni 2017, Polaroid menuntut agar Fujifilm membayar royalti / lisensi penggunaan film instan dengan format persegi. Di frame itulah produsen menempatkan zat - zat kimia penting buat memunculkan hasil jepretan di film.</p>

Tabel 4.3 menampilkan statistik dokumen artikel berita dengan membandingkan hasil peringkasan manual variasi 1 dengan hasil peringkasan sistem untuk berita multi dokumen index 1, 2, dan 3 dengan kategori teknologi.

Tabel 4.3 Statistik artikel multi dokumen

Statistik	Jumlah kata	Jumlah kalimat
Berita 1	255	16
Berita 2	244	13
Berita 3	382	22
Ringkasan manual	309	16
Ringkasan sistem	378	21

Berdasarkan Tabel 4.3 hasil ringkasan manual dan hasil ringkasan sistem akan dibandingkan kemudian dihitung nilai *Recall*nya menggunakan variasi ROUGE-1. ROUGE-1 merupakan total jumlah kata pada ringkasan sistem yang dianggap cocok atau sama dengan ringkasan manual variasi 1.

Tabel 4.4 Daftar kata hasil ringkasan sistem

meluncurkan	yang	hp	portable	berbarengan	printer
tahun	lalu	seperti	mencetak	tanpa	tinta
bernama	sprocket	polaroid	dan	cara	kerjanya
sangat	mirip	dapat	Desain	snap	maupun
kodak	printomatic	bedanya	sprocket	2	in
1	masih	bisa	berfungsi	sebagai	printer
...
format	tampil	persegi	kamera	meluncurkannya	di
awal	2017	sebuah	dengan	milimeter	instax
square	sql0	mengulik	lebih	dalam	lagi
berformat	bingkai	persegi	tak	sekedar	dipilih
agar	Foto	6262	unik		

Tabel 4.4 menampilkan contoh daftar kata pada hasil ringkasan sistem untuk index berita 1 dengan kategori teknologi. Tabel 4.5 menampilkan contoh daftar kata pada hasil ringkasan manual variasi 1 untuk index berita 1 dengan kategori teknologi.

Tabel 4.5 Daftar kata hasil ringkasan manual

portable	sprocket	telah	menjadi	sebuah	printer
----------	----------	-------	---------	--------	---------

tahun	yang	dapat	mencetak	kamera	tinta
bernama	lalu	baru	baru	ini	perangkat
tersebut	hp	berevolusi	meluncurkan	tanpa	instan
dengan	nama	yang	pas	Kamu	sprocket
...
instan	menjadi	format	persegi	yaitu	bisa
dengan	mengunduh	komunitas	cumacuma	startup	inovasi
Teknologi	member	indonesia	riset	dan	statistik
seputar	dailysocialid	secara	dan	dan	mengikuti
berita	dan	laporan	startup	gadget	terbaru

Tabel 4.6 menampilkan daftar kata yang saling beririsan pada hasil ringkasan manual variasi 1 dan sistem untuk *index* berita 1 dengan kategori teknologi.

Tabel 4.6 Daftar kata beririsan

bernama	sprocket	dapat	baru	sebuah	printer
portable	yang	hp	instan	tanpa	perangkat
tahun	lalu	mencetak	baru	ini	tinta
tersebut	menjadi	kamera	meluncurkan	dengan	yang
yaitu	bedanya	2	in	1	sprocket
...
agar	masalah	polaroid	berita	cukup	dan
akar	format	film	instan	startup	di
perseteruan	2017	Terkait	film	dengan	dan
persegi	bisa	startup	dan	inovasi	dailysocialid
format	teknologi	dan	sebenarnya	dalam	Sepele

Jumlah kata pada Tabel 4.4, Tabel 4.5, Tabel 4.6 ditampilkan pada Tabel 4.7 sebagai berikut.

Tabel 4.7 Statistika jumlah kata ringkasan

Kata unik	Jumlah
Ringkasan manual	309
Ringkasan sistem	378
Kemunculan yang sama pada ringkasan manual dan ringkasan sistem	223

Nilai ROUGE didapatkan dari perbandingan antara jumlah kata pada ringkasan sistem dengan jumlah kata pada ringkasan manual. Perhitungan nilai ROUGE-1 berdasarkan nilai pada Tabel 4.7 ditampilkan sebagai berikut.

$$ROUGE - 1 = \frac{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s \in Sum_{ref}} \sum_{gram_n \in s} Count(gram_n)} = \frac{223}{309} = 0,589$$

Selanjutnya dilakukan juga perhitungan *Precision* untuk mengetahui persentase jumlah kata yang benar pada hasil ringkasan sistem. Kemudian juga dilakukan perhitungan F1 untuk mengukur performa efisiensi algoritma yang digunakan dalam peringkasan teks otomatis multi dokumen. Perhitungan *Precision* dan F1 ditunjukkan sebagai berikut.

$$Precision = \frac{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count_{match}(gram_n)}{\sum_{s \in Sum_{sys}} \sum_{gram_n \in s} Count(gram_n)} = \frac{223}{378} = 0,721$$

$$F_1 = \frac{(1 + \beta^2)(Precision * recal)}{(\beta^2 * Precision + recall)} = \frac{(1 + 1^2)(0,589 * 0,721)}{(1^2 * 0,589 + 0,721)} = 0.649$$

Rekap hasil perhitungan ringkasan artikel ditunjukkan pada Tabel 4.8

Tabel 4. 8 Rekap hasil perhitungan

	Total Kalimat			Evaluasi		
	Input	Manual	Sistem	Recall	Precision	F1
Tabel 4.2	61	29	25	0.589	0.721	0.649

4.3 Pembahasan

Pada penelitian ini menggunakan *ROUGE* dalam mendapatkan nilai *Recall*, *Precision*, dan *f-measure*. Nilai *Recall* digunakan untuk mengukur nilai relevansi antara hasil ringkasan sistem dengan ringkasan manusia. Semakin relevan hasil ringkasan sistem yang dihasilkan terhadap ringkasan manusia, maka

semakin tinggi juga nilai *Recall*nya. Jika nilai *Recall* mencapai nilai maksimum atau mendapatkan nilai 1 berarti seluruh ringkasan manusia akan masuk dalam ringkasan sistem. Sebaliknya jika nilai *Precision* mencapai nilai maksimum atau 1 maka seluruh ringkasan sistem akan masuk kedalam ringkasan manusia.

Informasi inti dari berita direpresentasikan oleh ringkasan manusia. Sehingga nilai *Recall* akan lebih sering dipertahankan untuk mencapai atau mendekati nilai maksimum dengan tujuan agar seluruh ringkasan manusia masuk dalam ringkasan sistem. Jika ringkasan manusia telah masuk dalam ringkasan sistem maka kata ataupun kalimat diluar ringkasan manusia dapat dianggap sebagai informasi tambahan atau pendukung. Sedangkan nilai *f-measure* disini merupakan bentuk keseimbangan antara nilai *Recall* dan *Precision*. Sehingga performa dari sistem dikatakan berhasil dalam meringkas atau tidak juga bisa dinilai dari nilai *f-measure*.

Pada Tabel 4.8 ditunjukkan bahwa nilai *Recall* dari sistem peringkasan teks berita otomatis multi dokumen ini menghasilkan nilai sebesar 0,589. Nilai *Recall ROUGE* tertinggi untuk variasi 1 terdapat pada multi dokumen berita *index 2* dengan nilai 0,765. Dan nilai *Recall ROUGE* terendah untuk variasi 1 pada multi dokumen berita terdapat pada indeks ke 20 dengan nilai 0,391. Nilai *Recall ROUGE* tertinggi untuk variasi 2 terdapat pada multi dokumen berita *index 6* dengan nilai 0,876 . Dan nilai *Recall ROUGE* terendah untuk variasi 2 pada multi dokumen berita terdapat pada indeks ke 20 dengan nilai 0,41. Nilai *Recall ROUGE* tertinggi untuk variasi 3 terdapat pada multi dokumen berita *index 6* dengan nilai 0,876. Dan nilai *Recall ROUGE* terendah untuk variasi 3 pada multi dokumen

berita terdapat pada indeks ke 20 dengan nilai 0,439. Nilai *Recall ROUGE* tertinggi untuk variasi 4 terdapat pada multi dokumen berita *index* 4 dengan nilai 0,925 . Dan nilai *Recall ROUGE* terendah untuk variasi 4 pada multi dokumen berita terdapat pada indeks ke 29 dengan nilai 0,6.

Nilai *Precision* pada dokumen berita di Tabel 4.8 memperoleh nilai sebesar 0,721. Nilai *Precision* tertinggi untuk variasi 1 terdapat pada dokumen berita *index* ke 28 dengan nilai 0,857. Nilai *Precision ROUGE* terendah untuk variasi 1 pada dokumen berita *index* 3 dengan nilai 0,3. Nilai *Precision* tertinggi untuk variasi 2 terdapat pada dokumen berita *index* ke 26 dengan nilai 0,752. Nilai *Precision ROUGE* terendah untuk variasi 2 pada dokumen berita *index* 3 dengan nilai 0,06. Nilai *Precision* tertinggi untuk variasi 3 terdapat pada dokumen berita *index* ke 26 dengan nilai 0,831. Nilai *Precision ROUGE* terendah untuk variasi 3 pada dokumen berita *index* 3 dengan nilai 0,208. Nilai *Precision* tertinggi untuk variasi 4 terdapat pada dokumen berita *index* ke 20 dengan nilai 0,731. Nilai *Precision ROUGE* terendah untuk variasi 4 pada dokumen berita *index* 3 dengan nilai 0,173.

Nilai *f-measure* pada dokumen berita Tabel 4.8 memperoleh nilai sebesar 0,649. Nilai *f-measure* tertinggi untuk variasi 1 terdapat pada dokumen berita *index* 6 dengan nilai 0,791. Nilai *f-measure ROUGE* terendah untuk variasi 1 pada dokumen berita *index* 3 dengan nilai 0,402. Nilai *f-measure* tertinggi untuk variasi 2 terdapat pada dokumen berita *index* 17 dengan nilai 0,757. Nilai *f-measure ROUGE* terendah untuk variasi 2 pada dokumen berita *index* 3 dengan nilai 0,111. Nilai *f-measure* tertinggi untuk variasi 3 terdapat pada dokumen berita *index* 28 dengan nilai 0,73. Nilai *f-measure ROUGE* terendah untuk variasi 3 pada dokumen

berita *index* 3 dengan nilai 0,33. Nilai *f-measure* tertinggi untuk variasi 4 terdapat pada dokumen berita *index* 20 dengan nilai 0,798. Nilai *f-measure ROUGE* terendah untuk variasi 4 pada dokumen berita *index* 3 dengan nilai 0,289.

Tabel 4.9 Rata-rata Skor Rouge-1 pada 30 multi dokumen

Variasi Label	<i>Precision</i>	<i>Recall</i>	<i>F Measure</i>
Variasi 1	0.674	0.620	0.637
Variasi 2	0.526	0.689	0.580
Variasi 3	0.583	0.679	0.612
Variasi 4	0.382	0.790	0.505

Nilai rata-rata hasil akhir dari sistem yang telah dibuat dengan menggunakan 30 multi dokumen ditunjukkan pada Tabel 4.9 dimana didapati nilai rata-rata menggunakan variasi 1 yaitu nilai rata-rata *Precision* 0.674, rata-rata *Recall* 0.620, dan rata-rata *f-measure* 0,637. Selanjutnya, nilai rata-rata menggunakan variasi 2 yaitu nilai rata-rata *Precision* 0.526, rata-rata *Recall* 0.689, dan rata-rata *f-measure* 0.580. Nilai rata-rata menggunakan variasi 3 yaitu nilai rata-rata *Precision* 0.583, rata-rata *Recall* 0.679, dan rata-rata *f-measure* 0.612. Nilai rata-rata menggunakan variasi 4 yaitu nilai rata-rata *Precision* 0.382, rata-rata *Recall* 0.790, dan rata-rata *f-measure* 0.505.

Berdasarkan nilai rata-rata setiap variasi yang digunakan dalam pengujian, hasil terbaik didapatkan dari hasil peringkasan manual menggunakan variasi 1, yaitu nilai rata-rata *Precision* 0,674, rata-rata *Recall* 0,620, dan rata-rata *f-measure* 0,637. Hal ini disebabkan oleh beberapa alasan. Pertama ketika dilihat kembali bagaimana ringkasan manusia ini terbentuk maka hasil dari ringkasan manusia memiliki perbedaan terhadap 4 variasi ketika menggabungkan responden dengan dataset IndoSum. Label *true* yang berasal dari IndoSum memiliki jumlah yang

relatif lebih sedikit jika dibandingkan dengan label dari responden. Kedua disebabkan oleh sistem yang belum bisa menentukan jumlah ringkasannya sendiri sehingga berpengaruh sekali terhadap perbedaan jumlah kata pada ringkasan manusia dan ringkasan sistem.

Dari hasil penelitian yang telah dilakukan, sistem peringkasan teks berita berbahasa Indonesia multi dokumen dapat digunakan untuk kepentingan dalam bidang *Natural Language Processing (NLP)*. Sistem peringkasan dibangun dengan tujuan untuk mendapatkan inti informasi dan esensi dari sebuah berita, dengan adanya ringkasan berita maka akan lebih mudah dalam memahami isi dari berita tersebut. Berdasarkan hasil pengujian sistem yang telah dibangun dan diimplementasikan, pengguna dapat langsung mendapatkan ringkasan berita dari banyak dokumen sekaligus (multi dokumen) secara otomatis hanya dengan waktu yang relatif lebih singkat dibandingkan meringkas dengan cara manual.

Sistem peringkasan teks otomatis ini selaras dengan konsep muamalah dalam Islam yaitu *muamalah ma'a Allah* dan *muamalah ma'a an-nas*. Pada hadis dari Abu Hurairah ra, Nabi Muhammad SAW, bersabda :

عَنْ أَبِي هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ عَنِ النَّبِيِّ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ قَالَ مَنْ نَفَسَ عَنْ مُؤْمِنٍ كُرْبَةً مِنْ كُرْبِ الدُّنْيَا ، نَفَسَ اللَّهُ عَنْهُ كُرْبَةً مِنْ كُرْبِ الْآخِرَةِ ، وَمَنْ يَسِّرْ عَلَى مُعْسِرٍ ، يَسِّرَ اللَّهُ عَلَيْهِ فِي الدُّنْيَا وَالْآخِرَةِ ، وَمَنْ سَتَرَ مُسْلِمًا ، سَتَرَهُ اللَّهُ فِي الدُّنْيَا وَالْآخِرَةِ ، وَاللَّهُ فِي عَوْنِ الْعَبْدِ مَا كَانَ الْعَبْدُ فِي عَوْنِ أَخِيهِ ،

“Barangsiapa yang melepaskan satu kesusahan seorang mukmin, pasti Allah akan melepaskan darinya satu kesusahan pada hari kiamat. Barangsiapa yang menjadikan mudah urusan orang lain, pasti Allah akan memudahkan urusannya di dunia dan akhirat. Barangsiapa yang menutupi aib seorang muslim, pasti Allah akan menutupi aibnya di dunia dan di akhirat. Allah senantiasa menolong hamba-Nya selama hamba-Nya suka menolong saudaranya.”(Hadits Shahih: Diriwayatkan oleh Muslim no 2699).

Menolong kesusahan orang lain merupakan tindakan yang menerapkan konsep *muamalah ma'a an-nnas*. Dimana dalam sistem yang dibuat dapat membantu atau menolong pengguna dalam memperoleh informasi. Hal ini juga selaras dengan konsep *muamalah ma'a Allah* Saling tolong menolong dalam kebaikan merupakan bentuk kepatuhan manusia kepada *Allah subhanahu wa ta'ala* yang Maha Kuasa. Sistem peringkasan teks multi dokumen otomatis ini diharapkan dapat membantu urusan orang lain sebagai bagian menjalankan perintah Allah *subhanahu wa ta'ala* dalam kebaikan dan takwa.

Selain ayat diatas konsep *muamalah ma'a Allah* juga terdapat pada salah satu permasalahan dimana apabila terdapat satu isu atau topik berita namun banyak sekali bahan bacaannya, maka sistem peringkasan teks berita otomatis dapat menyortir isi yang penting dari berbagai bacaan berita (multi dokumen) tanpa menghilangkan informasi yang penting. Pada era modern saat ini banyak sekali berita yang muncul dengan berbagai judul yang terkadang tidak relevan dengan isinya yang disebut sebagai *clickbait*, sehingga banyak sekali masyarakat terkecoh dalam membaca konten berita tersebut. Dalam islam sendiri sangat dianjurkan untuk melakukan *tabayyun* yang bermakna untuk mencari fakta yang sebenarnya dari memahami sebuah informasi yang didapatkan. Jika judul berita dapat menyesatkan pembaca maka sistem peringkasan teks berita dapat meringkas inti informasi dari berita tersebut berdasarkan isi artikelnya bukan berdasarkan judulnya. Sebagaimana firman Allah *subhanahu wa ta'ala* dalam surah Al-Hujurat ayat 6 dalam Al-Qur'an yang berbunyi:

يَا أَيُّهَا الَّذِينَ ءَامَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَأٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهْلَةٍ فَتُصْحَبُوا عَلٰى مَا فَعَلْتُمْ نَادِمِينَ

“Hai orang-orang yang beriman, jika datang kepadamu orang fasik membawa suatu berita, maka periksalah dengan teliti agar kamu tidak menimpakan suatu musibah kepada suatu kaum tanpa mengetahui keadaannya yang menyebabkan kamu menyesal atas perbuatanmu itu.” (Al-Hujurat ayat 6)

Menurut tafsir Ibnu Katsir, Allah *subhanahu wa ta'ala* memerintahkan kita semua (orang mukmin) untuk meneliti kembali berita yang diterima oleh orang yang tidak dikenal, barangkali berita tersebut berasal dari orang-orang yang fasik. Memahami konteks beritanya sama saja dengan mengikuti jejaknya (orang fasik itu). Sedangkan Allah *subhanahu wa ta'ala* sudah melarang kaum mukmin untuk mengikuti jalan orang-orang fasik (orang yang rusak). Begitu pula dengan persebaran berita *online* yang marak terjadi belum dapat dipastikan kebenarannya, karena jika sembarangan membaca berita tanpa memahami inti dari konteks berita maka akan terjerumus dalam informasi dan pemahaman yang salah. Dengan adanya sistem peringkasan teks multi dokumen berita berbahasa Indonesia ini dapat diambil pelajaran bahwa untuk memahami isi berita dapat dilakukan dengan cara mengambil kalimat yang memiliki kedekatan jarak dengan *centroid*-nya dimana kalimat yang dipilih merupakan representasi dari ringkasan. Sehingga informasi dalam suatu kumpulan berita tidak akan terduplikasi dan tidak adanya redundansi serta dapat meminimalisir terjadinya kesalahpahaman.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan pengujian sistem Peringkasan Teks Multi Dokumen Berita berbahasa Indonesia Menggunakan *FastText* dan *K-means Clustering* dengan 30 data multi dokumen menghasilkan dimana didapati nilai rata-rata menggunakan variasi 1 yaitu nilai rata-rata *Precision* 0.674, rata-rata *Recall* 0.620, dan rata-rata *f-measure* 0,637. Selanjutnya, nilai rata-rata menggunakan variasi 2 yaitu nilai rata-rata *Precision* 0.526, rata-rata *Recall* 0.689, dan rata-rata *f-measure* 0.580. Nilai rata-rata menggunakan variasi 3 yaitu nilai rata-rata *Precision* 0.583, rata-rata *Recall* 0.679, dan rata-rata *f-measure* 0.612. Nilai rata-rata menggunakan variasi 4 yaitu nilai rata-rata *Precision* 0.382, rata-rata *Recall* 0.790, dan rata-rata *f-measure* 0.505. Dapat disimpulkan bahwa sistem dapat melakukan peringkasan dengan baik dan cukup akurat dibuktikan dengan nilai rata-rata *Recall* yang tinggi dimana recall terendah mencapai 0.620 pada variasi 1 dan recall tertinggi sebesar 0,790 pada variasi 4 sehingga disimpulkan bahwa hasil peringkasan cukup relevan dan sesuai dengan data inputnya serta dengan keberhasilan sistem yang dibuat juga akan menjadi sarana bentuk *muamalah ma'a Allah* dan *muamalah ma'a an-nnas* yang mana telah sesuai pada Qur'an surah Al-Maidah ayat 2, Al-Hujurat ayat 6, dan hadis dari Abu Hurairah ra yang diriwayatkan oleh Muslim no 2699.

5.2 Saran

Dalam pengembangan sistem peringkasan teks otomatis berbahasa Indonesia, terdapat beberapa saran yang dapat diterapkan untuk mencapai hasil yang lebih maksimal. Berikut saran kepada peneliti selanjutnya :

- a. Menambahkan suatu algoritma yang berfungsi untuk menentukan jumlah k atau jumlah cluster secara otomatis.
- b. Inisiasi *centroid* awal dilakukan secara acak sehingga perlu dilakukan suatu pendekatan dalam mendapatkan nilai *centroid* awal yang lebih optimal.

DAFTAR PUSTAKA

- Abdullah, M. Z., & Faticah, C. (2018). Peringkasan multi-dokumen berita berdasarkan fitur berita dan part of speech tagging. *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, 4(2), 56–66. <https://doi.org/10.26594/register.v4i2.1251>
- Abidin, Z., & Nurjanah, E. (2020). Sistem Peringkasan Teks Otomatis Multi Dokumen Kliping Artikel Berita Gempa menggunakan Metode Tf-Idf. *Jurnal Teknologi Informasi Dan Komunikasi*, 13(1), 2252–4517.
- Akaresti, A. K., Fauzi, M. A., & Abdurrachman Bachtiar, F. (2019). *Peringkasan Multi-Dokumen Berbasis Clustering pada Sistem Temu Kembali Berita Online Menggunakan Metode K-Means* (Vol. 3, Issue 1). <http://j-ptiik.ub.ac.id>
- Akbar, G. (2021). *Automatic Text Summarization Berita Bahasa Indonesia Menggunakan Metode Attentional Encoder Decoder*.
- Allahyari, M., Pouriyyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey. *ArXiv Preprint ArXiv, 2022-March*. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>
- Amalia, A., Sitompul, O. S., Nababan, E. B., & Mantori, T. (2020). An Efficient Text Classification Using FastText for Bahasa Indonesia Documents Classification. *International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. https://doi.org/10.1162/tacl_a_00051/1567442/tacl_a_00051.pdf
- Hahn, U. (2000). *Automatic Summarization*.
- Hendrastuty, N., & SN, A. (2021). Text Summarization in Multi Document Using Genetic Algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 15(4), 327. <https://doi.org/10.22146/ijccs.66026>
- Hovy, E. (2005). *Automated text summarization*.
- Khaerunnisa, Z., Ashari, A., & Sari, Y. (2022). *PERBANDINGAN PENGGUNAAN WORD EMBEDDING WORD2VEC, GLOVE, DAN FASTTEXT DALAM PERINGKASAN TEKS EKSTRAKTIF BERITA BERBAHASA INDONESIA MENGGUNAKAN*. Universitas Gadjah Mada.
- Kurniawan, K., & Louvan, S. (2019). IndoSum: A New Benchmark Dataset for Indonesian Text Summarization. *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, 215–220. <https://doi.org/10.1109/IALP.2018.8629109>

- Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–84.
- McCallum, & Kachites, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. In <http://www.cs.cmu.edu/~mccallum/bow>.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*. <http://arxiv.org/abs/1301.3781>
- Nur Aisyah, K., Anggraini, S., & Agus Zainal Arifin, dan. (2019). Peringkasan Teks Multi-Dokumen Berdasarkan Metode Sentence Extraction dan Word Sense Disambiguation. *NJCA* , 4(1).
- Pramudi Ismi, D., Ardianto, F., Dahlan, A., & Selatan, J. R. (2019). Peringkasan Ekstraktif Teks Bahasa Indonesia dengan Pendekatan Unsupervised Menggunakan Metode Clustering. *CYBERNETICS*, 3(02), 90–99.
- Rani, N. L. R. M. (2013). Persepsi Jurnalis dan Praktisi Humas terhadap Nilai Berita. *Jurnal ILMU KOMUNIKASI*, 10.
- Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. (2018). Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) - Significance Sentences. *Procedia Computer Science*, 135, 663–670. <https://doi.org/10.1016/j.procs.2018.08.220>
- Widjanarko, A., Kusumaningrum, R., & Surarso, B. (2018). Multi Document Summarization for the Indonesian Language Based on Latent Dirichlet allocation and Significance sentence. *International Conference on Information and Communications Technology (ICOIACT)*, 520–524.

LAMPIRAN

LAMPIRAN

Lampiran 1 Data 30 data multi dokumen

Index	Judul berita
0	Mati Dalam Operasi Antinarkotika Filipina Pro, Perang Duterte Lawan Narkoba Dipergencar, Filipina Diserang Isu Pelanggaran Ham, Rodrigo Duterte Diselidiki Mahkamah Internasional
1	Hp Luncurkan Kamera Instan Sprocket In , Jajal Printer Portable Canon Selphy Cp, Polaroid Dan Fujifilm Bersengketa Akibat Format Sq
2	Arkavidia Itb Akan Fokus Bahas Perkembangan Sta, Visio Incubator Umumkan Startup Terpilih Untuk , Clapham Startupfest Segera Hadir Kembali Di M
3	Asus Xiu Laptop Multimedia Gaming Ready, Asus Rilis Laptop Gim Seharga Motor Sport, Asus Ramu Notebook Rog Gvik
4	Tesla Akan Tanam Teknologi Otonom Di Semua Mobil L, Tiongkok Siapkan Jutaan Dolar Untuk Ganggu Pasar M
5	Asia Pacific Predator League Resmi Dimulai, Acer Nitro Spin Laptop Gaming In Terbaru, Philips Sedang Garap Monitor K
6	Angin Meningkatkan Peran Perempuan Di Dunia Teknol, Rayakan Agustus Go Jek Sumbang Rp Miliar Ke
7	Juni Tesla Model X Masuk Indonesia, Bmw Punya Seri Listrik Untuk Lawan Tesla Model , Tesla Model Mulai Diproduksi
8	Asus Xiu Laptop Multimedia Gaming Ready, Asus Rilis Laptop Gim Seharga Motor Sport, Monitor Gaming Terbaru Asus Andalkan Layar Curved, Asus Ramu Notebook Rog Gvik, Amd Luncurkan Prosesor Ryzen Untuk Laptop, Mouse Gaming Msi Clutch Gm
9	Dua Smartphone Zenfone Baru Tawarkan Solusi Bersel, Asus Diam Diam Rilis Zenfone Selfie Lite Di Mala, Samsung Galaxy A Bidik Generasi Millennial, Asus Zenfone Selfie Lite Tiba Di Filipina
10	Pencarian Korban Longsor Ponorogo Terkendala Medan, Orang Masih Hilang Akibat Longsor Di Ponorogo
11	Polisi Temukan Potongan Tangan Di Lokasi Carok, Carok Massal Di Madura Tewaskan Tiga Orang, Polisi Ringkus Pelaku Carok Massal Di Madura
12	Smartphone Flip Premium Samsung W Resmi Dirili, Samsung Luncurkan Flip Phone Android W, Microsoft Office Versi Android Kini Sudah Tersedia
13	Vivo X Plus Ud Akhirnya Diluncurkan, Vivo V Plus Tiba Di India, Vivo Rilis Teaser Varian X Mars Edition
14	Nintendo Targetkan Untuk Memproduksi Dua Juta Unit, Nintendo Switch Konsol Game Paling Ditunggu Hadir , Nintendo Switch Sudah Terjual Juta Unit, Nintendo Akan Kembali Produksi Nes Classic Edition, Beli Game Nintendo Switch Di Eshop Dengan Paypal, Nintendo Umumkan Jadwal Rilis Game Baru Untuk Swit, Penjualan Nintendo Switch Sukses Melampaui Pencapa
15	Seagate Keluarkan Harddisk Kapasitas Tb, Corsair Gelar Kompetisi Modifikasi Case Komputer T, Menjajal Hard Disk Eksternal Lacie Dji Copilot
16	Google Assistant Kini Bisa Kenali Lagu Seperti Sha, Deezer Luncurkan Songcatcher, Apple Resmi Akuisisi Aplikasi Penebak Judul Lagu S
17	Jokowi Siapkan Regulasi Mobil Listrik, Presiden Jokowi Mengaku Deg Degan Jajal Mobil List
18	Jokowi Hormati Putusan Ma Soal Perda, Jokowi Resmikan Tol Becakayu Proyek Mangkrak Dua D, Resmikan Tol Jokowi Bertolak Ke Ujung Selatan Pula, Presiden Jokowi Resmikan Tol Pertama Di Sumatera S, Kredit Ultra Mikro Untuk Ratakan Kesejahteraan, Pembangunan Infrastruktur Jadi Kunci Indonesia Men

Index	Judul berita
19	Militer Korsel Ancam Balas Dendam Serangan Korut, Di Tengah Ancaman Korut Jepang As Latihan Maritim
20	Tujuh Suku Papua Ingin Dialog Dengan Freeport, Jokowi Untung Kita Punya Menteri Susi Yang Tenggel, Mike Pence Kunjungi Indonesia April
21	Pemerintah Naikkan Cukai Rokok Persen, Ylki Kenaikan Harga Rokok Terlalu Kecil Harusnya J, Sri Mulyani Resmikan Harga Rokok Naik Mulai Janu, Cukai Resmi Naik Harga Jual Eceran Rokok Bakal Mel
22	Pertamina Realisasikan Wilayah Bbm Satu Harga, Iri Bbm Satu Harga Pgn Ingin Harga Gas Sama Rata, Esdm Bbm Satu Harga Bukan Masalah Untung Rugi Tapi, Salurkan Bbm Satu Harga Pertamina Butuh Rp T, Bbm Satu Harga Masuk Ke Daerah Terpencil, Dpr Dukung Kebijakan Energi Untuk Rakyat
23	Kisruh Impor Beras Dpr Minta Pemerintah Validasi D, Bps Aktivitas Ekspor Impor Beras Melambung Hingga , Pengumuman Impor Beras Seret Harga Beras Turun Rp, Enggar Impor Beras Ton Tanpa Apbn, Sekjen Partai Komunis Vietnam Bahas Zona Perdagang
24	Gara Gara Pemilu As Harga Emas Dunia Turun, Akhir Pekan Rupiah Dan Harga Komoditas Warnai Gera, Euro Melemah Usai Catalonia Ingin Pisah Dari Spanyol, Hari Ini Harga Emas Turun Rp Menjadi Rp , Emas Tertekan Penguatan Saham Dan Dolar As
25	Anies Enggan Beberkan Rencana Atasi Banjir Jakarta, Banjir Karena Penyempitan Kali Anies Minta Warga G, Antisipasi Banjir Kiriman Pemprov Dki Siagakan Sel, Pemprov Dki Berencana Kaji Ulang Intensif Pasukan , Demi Selamatkan Jakarta Dari Banjir Anies Rela Lak
26	Gunung Sinabung Kembali Erupsi, Ketua Dpr Minta Sarana Di Kawasan Sinabung Mulai D, Gunung Sinabung Semburkan Awan Panas, Gunung Sinabung Meletus
27	Dirut Pt Garam Terancam Tahun Penjara, Polri Tangkap Dirut Pt Garam
28	Sri Mulyani Yakin Ekonomi Kuartal Iv Tumbuh Tembus, Sri Mulyani Kestabilan Inflasi Adalah Kunci, Indonesia Masih Rendah Ketimbang Negara Lain, Sri Mulyani Ekonomi Tumbuh Persen Tahun Ini
29	Polri Luncurkan Aplikasi Peringatan Dini Gunung Ag, Gunung Agung Bali Meletus, Gunung Agung Meletus Lagi, Gubernur Pastika Ajak Warga Tidak Panik Hadapi Sta, Banyuwangi Belum Terdampak Abu Vulkanis Gunung Agung

Lampiran II Rekap hasil peringkasan variasi 1

index	jumlah kalimat input	jumlah kalimat sistem	jumlah kalimat manual	jumlah kata sistem	jumlah kata manual	jumlah kata_ overlap	<i>Precision</i>	<i>Recall</i>	f1
0	61	25	29	438	583	344	0,785	0,59	0,674
1	51	21	16	378	309	223	0,59	0,722	0,649
2	58	25	21	518	455	348	0,672	0,765	0,715
3	120	48	21	830	409	249	0,3	0,609	0,402
4	41	18	15	289	310	200	0,692	0,645	0,668
5	71	28	24	552	583	344	0,623	0,59	0,606
6	32	14	15	374	435	320	0,856	0,736	0,791
7	52	20	20	320	491	223	0,697	0,454	0,55
8	213	83	52	1521	1230	869	0,571	0,707	0,632
9	94	37	28	666	701	384	0,577	0,548	0,562
10	25	10	9	169	210	122	0,722	0,581	0,644
11	42	18	16	346	374	252	0,728	0,674	0,7
12	42	18	12	308	261	159	0,516	0,609	0,559
13	57	24	19	428	381	267	0,624	0,701	0,66
14	132	51	41	970	1029	688	0,709	0,669	0,688
15	52	21	16	357	341	205	0,574	0,601	0,587
16	51	19	15	275	294	164	0,596	0,558	0,576
17	39	17	13	248	273	200	0,806	0,733	0,768
18	97	36	27	827	771	497	0,601	0,645	0,622
19	27	12	10	245	265	159	0,649	0,6	0,624
20	35	14	14	219	353	138	0,63	0,391	0,483
21	82	33	22	701	591	434	0,619	0,734	0,672
22	103	44	36	895	973	607	0,678	0,624	0,65
23	91	37	24	655	594	411	0,627	0,692	0,658
24	78	29	26	573	599	448	0,782	0,748	0,765
25	86	35	33	542	715	415	0,766	0,58	0,66
26	36	17	19	278	464	224	0,806	0,483	0,604
27	30	11	14	227	368	180	0,793	0,489	0,605
28	74	31	38	545	814	467	0,857	0,574	0,687
29	56	24	25	491	686	382	0,778	0,557	0,649

Lampiran II Rekap hasil peringkasan variasi 2

index	jumlah kalimat input	jumlah kalimat sistem	jumlah kalimat manual	jumlah kata sistem	jumlah kata manual	jumlah kata_overlap	Precision	Recall	f1
0	61	25	17	438	345	229	0,523	0,664	0,585
1	51	21	13	378	235	188	0,497	0,8	0,613
2	58	25	14	518	321	252	0,486	0,785	0,601
3	120	48	7	830	74	50	0,06	0,676	0,111
4	41	18	11	289	229	171	0,592	0,747	0,66
5	71	28	15	552	339	241	0,437	0,711	0,541
6	32	14	9	374	259	227	0,607	0,876	0,717
7	52	20	14	320	366	183	0,572	0,5	0,534
8	213	83	31	1521	770	616	0,405	0,8	0,538
9	94	37	19	666	486	302	0,453	0,621	0,524
10	25	10	7	169	154	111	0,657	0,721	0,687
11	42	18	14	346	328	234	0,676	0,713	0,694
12	42	18	9	308	196	132	0,429	0,673	0,524
13	57	24	11	428	221	164	0,383	0,742	0,505
14	132	51	29	970	716	503	0,519	0,703	0,597
15	52	21	10	357	184	119	0,333	0,647	0,44
16	51	19	12	275	230	142	0,516	0,617	0,562
17	39	17	10	248	193	167	0,673	0,865	0,757
18	97	36	21	827	578	397	0,48	0,687	0,565
19	27	12	6	245	147	91	0,371	0,619	0,464
20	35	14	14	219	323	133	0,607	0,412	0,491
21	82	33	15	701	427	316	0,451	0,74	0,56
22	103	44	26	895	732	517	0,578	0,706	0,636
23	91	37	16	655	444	307	0,469	0,691	0,559
24	78	29	19	573	444	347	0,606	0,782	0,682
25	86	35	20	542	465	320	0,59	0,688	0,636
26	36	17	15	278	350	209	0,752	0,597	0,666
27	30	11	11	227	303	146	0,643	0,482	0,551
28	74	31	24	545	563	394	0,723	0,7	0,711
29	56	24	18	491	493	352	0,717	0,714	0,715

Lampiran II Rekap hasil peringkasan variasi 3

index	jumlah kalimat input	jumlah kalimat sistem	jumlah kalimat manual	jumlah kata sistem	jumlah kata manual	jumlah kata_overlapse	<i>Precision</i>	<i>Recall</i>	f1
0	61	25	20	438	418	281	0,642	0,672	0,657
1	51	21	17	378	296	232	0,614	0,784	0,688
2	58	25	15	518	345	261	0,504	0,757	0,605
3	120	48	15	830	218	173	0,208	0,794	0,33
4	41	18	12	289	247	184	0,637	0,745	0,687
5	71	28	18	552	412	279	0,505	0,677	0,579
6	32	14	9	374	259	227	0,607	0,876	0,717
7	52	20	14	320	366	183	0,572	0,5	0,534
8	213	83	34	1521	806	634	0,417	0,787	0,545
9	94	37	22	666	538	327	0,491	0,608	0,543
10	25	10	9	169	181	122	0,722	0,674	0,697
11	42	18	16	346	379	259	0,749	0,683	0,714
12	42	18	13	308	272	171	0,555	0,629	0,59
13	57	24	12	428	243	184	0,43	0,757	0,548
14	132	51	32	970	762	536	0,553	0,703	0,619
15	52	21	12	357	223	152	0,426	0,682	0,524
16	51	19	14	275	254	155	0,564	0,61	0,586
17	39	17	14	248	250	181	0,73	0,724	0,727
18	97	36	23	827	630	418	0,505	0,663	0,574
19	27	12	6	245	147	91	0,371	0,619	0,464
20	35	14	18	219	401	176	0,804	0,439	0,568
21	82	33	17	701	479	354	0,505	0,739	0,6
22	103	44	26	895	732	517	0,578	0,706	0,636
23	91	37	19	655	520	350	0,534	0,673	0,596
24	78	29	20	573	462	364	0,635	0,788	0,703
25	86	35	26	542	550	356	0,657	0,647	0,652
26	36	17	19	278	392	231	0,831	0,589	0,69
27	30	11	11	227	303	146	0,643	0,482	0,551
28	74	31	27	545	598	417	0,765	0,697	0,73
29	56	24	21	491	544	371	0,756	0,682	0,717

Lampiran II Rekap hasil peringkasan variasi 4

index	jumlah kalimat input	jumlah kalimat sistem	jumlah kalimat manual	jumlah kata sistem	jumlah kata manual	jumlah kata_overlap	Precision	Recall	f1
0	61	25	13	438	260	207	0,473	0,796	0,593
1	51	21	11	378	162	136	0,36	0,84	0,504
2	58	25	11	518	212	185	0,357	0,873	0,507
3	120	48	11	830	165	144	0,173	0,873	0,289
4	41	18	8	289	133	123	0,426	0,925	0,583
5	71	28	8	552	183	124	0,225	0,678	0,337
6	32	14	8	374	150	118	0,316	0,787	0,45
7	52	20	9	320	150	114	0,356	0,76	0,485
8	213	83	19	1521	353	312	0,205	0,884	0,333
9	94	37	12	666	250	180	0,27	0,72	0,393
10	25	10	6	169	117	102	0,604	0,872	0,713
11	42	18	9	346	204	173	0,5	0,848	0,629
12	42	18	11	308	191	129	0,419	0,675	0,517
13	57	24	11	428	178	130	0,304	0,73	0,429
14	132	51	22	970	465	373	0,385	0,802	0,52
15	52	21	9	357	163	116	0,325	0,712	0,446
16	51	19	11	275	155	112	0,407	0,723	0,521
17	39	17	8	248	124	87	0,351	0,702	0,468
18	97	36	13	827	363	226	0,273	0,623	0,38
19	27	12	4	245	98	80	0,327	0,816	0,466
20	35	14	10	219	182	160	0,731	0,879	0,798
21	82	33	9	701	211	188	0,268	0,891	0,412
22	103	44	16	895	398	320	0,358	0,804	0,495
23	91	37	14	655	305	252	0,385	0,826	0,525
24	78	29	15	573	319	268	0,468	0,84	0,601
25	86	35	16	542	289	249	0,459	0,862	0,599
26	36	17	15	278	214	132	0,475	0,617	0,537
27	30	11	6	227	130	113	0,498	0,869	0,633
28	74	31	13	545	236	208	0,382	0,881	0,533
29	56	24	15	491	325	195	0,397	0,6	0,478