TUGAS AKHIR

PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

ADE RIYANI

15102002

PROGRAM STUDI S1 INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI DAN INFORMATIKA
INSTITUT TEKNOLOGI TELKOM PURWOKERTO
2019

TUGAS AKHIR

PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

IMPLEMENTATION WORD2VEC METHOD TO DETECT DOCUMENT SIMILARITY

Disusun Sebagai Salah Satu Syarat untuk Memperoleh Gelar Sarjana Komputer

ADE RIYANI
15102002

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI DAN INFORMATIKA
INSTITUT TEKNOLOGI TELKOM PURWOKERTO
2019

Lembar Pengesahan Pembimbing

PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

IMPLEMENTATION WORD2VEC METHOD TO DETECT DOCUMENT SIMILARITY

Dipersiapkan dan Disusun Oleh

ADE RIYANI
15102002

Telah Diujikan dan Dipertahankan dalam siding Ujian Tugas Akhir
Pada hari kamis, 7 Februari 2019

Pembimbing I,	Pembimbing II,
Muhammad Zidny Naf'an, S.Kom., M.Kom	Auliya Burhanuddin, S.Si., M.Kom
NIDN. 0626128801	NIDN. 0630058202

Tugas Akhir ini diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 28 Desember 2019
Dekan

Didi Supriyadi, S.T., M.Kom., ITILE
NIK. 13840016

Lembar Penetapan Penguji

PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

PENERAPAN METODE WORD2VEC UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

Dipersiapkan dan Disusun Oleh

ADE RIYANI
15102002

Tugas Akhir Telah diuji dan Dinilai Panitia Penguji Program Studi Teknik Informatika Fakultas Teknologi Industri dan Informatika Institut Teknologi Telkom Purwokerto

Pada Tanggal : 7 Februari 2019

Ketua Penguji

(Tri Ginanjar Laksana, S.Kom., M.C.S., M.Kom.)
NIDN. 0407088502

Anggota Penguji I

(Agi Prasetiadi, S.T., M. Eng.)
NIDN. 0617098802

Anggota Penguji II

(Muhammad Fajar Sidiq, S.T., M.T.)
NIDN. 0619029102

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertandatangan di bawah ini,

Nama Mahasiswa : Ade Riyani
NIM : 15102002
Program Studi : Teknik Informatika

Menyatakan bahwa Tugas Akhir dengan judul berikut:
PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN

Dosen Pembimbing Utama : Muhammad Zidny Naf'an, S.Kom.,M.Kom

Dosen Pembimbing Pendamping : Auliya Burhanuddin, S.SI., M.Kom

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Institut Teknologi Telkom Purwokerto maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan, dan penelitian Saya Sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing.
3. Dalam Karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggungjawab Saya, bukan tanggungjawab Institut Teknologi Telkom Purwokerto.
5. Pernyataan ini Saya buat dengan sesungguhnya, apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka Saya bersedia menerima Sanksi Akademik dengan pencabutan gelar yang sudah diperoleh serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Purwokerto, 28 januari 2019,
Yang Menyatakan,

(Ade Riyani)

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah SWT, karena atas berkat rahmat, taufik, hidayah dan karunia-Nya penulis dapat menyelesaikan penyusunan skripsi ini di Institut Teknologi Telkom Purwokerto tempat penulis belajar.

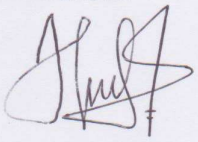
Adapun judul skripsi penelitian ini yaitu **PENERAPAN METODE *WORD2VEC* UNTUK MENDETEKSI KEMIRIPAN DOKUMEN.** Dalam menyelesaikan penulisan laporan skripsi ini tentunya tidak terlepas dari bimbingan, doa serta bantuan dari berbagai pihak. Oleh karena itu dalam kesempatan ini penulis mengucapkan terimakasih kepada:

1. Bapak Dr. Ali Rokhman, M.Si selaku Rektor Institut Teknologi Telkom Purwokerto.
2. Bapak Didi Supriyadi, S.T., M.Kom selaku Dekan Fakultas Teknik Industri dan Informatika Institut Teknologi Telkom Purokerto.
3. Muhammad Zidny Naf'an, S.Kom., M.Kom selaku ketua program studi S1 Informatika dan selaku Dosen Pembimbing pertama yang telah memberikan bimbingan dan pengarahan pada saat penyusunan Tugas Akhir.
4. Auliya Burhanuddin, S.SI., M.Kom. selaku Dosen Pembimbing kedua yang telah memberikan bimbingan dan pengarahan pada saat penyusunan Tugas Akhir.
5. Orang tua yang telah memberikan doa dan dukungan untuk penulis pada saat mengerjakan Tugas Akhir.
6. Asyhar Nurrochman yang telah menemani dan memberikan semangat untuk menyelesaikan Tugas Akhir.
7. Cuplis kucing kesayangan yang selalu menemani penulis bermain diwaktu senggang.
8. Alhamda Adisoka Bimantara sebagai konsultan pada saat program saya gagal atau eror.
9. Teman-teman mahasiswa Institut Teknologi Telkom Purwokerto yang tidak bisa disebutkan satu persatu.

Dalam penyusunan penelitian Tugas Akhir ini, penulis menyadari masih banyak kekurangan. Untuk itu, saran dan kritik pembaca untuk kesempurnaan laporan Tugas Akhir.

Akhirnya, penulis berharap semoga tugas akhir ini dapat bermanfaat dan menambah wawasan bagi pembaca.

Penulis



Ade Riyani

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemanfaatan teknologi informasi memiliki dampak pada banyak sektor, termasuk sektor Pendidikan. Dalam dunia Pendidikan, sangatlah diperlukan informasi untuk mendukung pembelajaran. Komputer dan internet sangat bermanfaat untuk pencarian informasi. Selain adanya kemudahan dalam pencarian informasi, hadirnya teknologi informasi ternyata juga memiliki dampak negatif. Salah satu dampak negatif di dunia Pendidikan adalah plagiarism oleh civitas akademika. Tindakan plagiarisme tidak hanya melibatkan siswa maupun mahasiswa, namun plagiarisme juga dapat melibatkan dosen bahkan mahasiswa yang bergelar doktor. Diungkapkan pada Kumparan, 2018 yang melaporkan bahwa Ombudsman RI menemukan plagiarisme dalam tiga karya ilmiah Rektor Universitas Halu Oleo (UHO) Kendari, Seulawesi Tenggara yang dilakukan oleh Muhammad Zamrun Firihu. Berdasarkan hasil analisis yang 30 guru besar UHO, zamrun terbukti bersalah melakukan tindakan plagiarisme dalam tiga jurnal internasional. Tingkat kesamaanya bahkan mencapai 78% bahkan lebih. Atas kasus tersebut, Ombudsman menilai ada pelanggaran berupa maladministrasi yang dilakukan Kementrian Riset, Teknologi, dan Pendidikan tinggi (Kemenristekdikti) atas pelantikan Zamrun sebagai rektor. Dari kasus tersebut, plagiat dapat merusak moral civitas akademika dalam dunia Pendidikan terutama di Indonesia.

Plagiarisme merupakan salah satu perbuatan yang melanggar kode etik dalam dunia penulisan, karena tindakan tersebut telah mengambil karya orang lain dan mengakuinya sebagai karya sendiri [1]. Selain melanggar kode etik, plagiarisme merusak moral dan mempengaruhi kualitas akan suatu dokumen sehingga keaslianya tidak dapat dipercaya. Pada tahun 2010, Direktorat Jenderal Pendidikan Tinggi telah mengeluarkan peraturan tentang cara pencegahan dan penanggulangan plagiarisme termasuk sanksi untuk dosen, mahasiswa, dan calon guru besar sekalipun. Praktek ini kerap dihubungkan dalam dunia pendidikan khususnya mahasiswa dalam melaksanakan tugasnya. Merujuk pada Permendiknas No. 17 Tahun 2010 tentang

Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi disebutkan bahwa dalam melaksanakan otonomi keilmuan dan kebebasan akademik mahasiswa, dosen, peneliti, tenaga kependidikan wajib menjunjung tinggi kejujuran dan etika akademik, terutama larangan untuk melakukan plagiat dalam menghasilkan karya ilmiah, sehingga kreativitas dalam bidang akademik dapat tumbuh dan berkembang [2]. Namun jika pengambilan suatu karangan disertai dengan mencantumkan asalnya atau nama pengarang serta judul karangan yang diambil, maka tindakan pengambilan karangan tersebut bukan merupakan plagiat. Oleh karna itu, perlu dilakukan pemeriksaan kemiripan antar dokumen, dalam hal ini adalah dokumen teks, sebagai langkah validasi keterkaitan dan hubungan antar dokumen tersebut. Sepasang kata dinyatakan mirip apabila memiliki kesamaan dari sisi makna atau konsep. Perhitungan deteksi kemiripan dilatar belakangi oleh suatu masalah yaitu mesin belum dapat menyamakan persepsi manusia dengan baik, untuk itu kemiripan semantik digunakan untuk membantu mesin memahami bahasa manusia.

Tugas deteksi kemiripan yaitu untuk menghitung nilai kedekatan antara dua buah kata. Karena mesin hanya dapat membaca angka, maka kata-kata yang ada harus diterjemahkan dalam bentuk angka, maka kata-kata yang ada harus diterjemahkan dalam bentuk angka terlebih dahulu. Untuk mencari nilai kedekatan antara dua buah kata dapat dilakukan dengan cara menghitung nilai *vector* kata tersebut. Supaya mendapatkan nilai kemiripan dengan *vector* yaitu dengan cara menghitung perbedaan sudut antara dua buah *vector* dengan rumus *Cosine Similarity*.

Beberapa penelitian untuk mengetahui tingkat kesamaan dokumen telah dilakukan diantaranya menggunakan metode *Cosine Similarity* [3], Jaccard [4], TF-IDF [5], *Support Vector Regression* [6], *Singgle pass clustering* [7], *Levenshtein Distance* [8], dan *Latent Semantic Analysis* [9]. Berdasarkan referensi penelitian yang telah dilakukan Metode *Word2vec* belum banyak digunakan untuk kemiripan dokumen karena metode *Word2vec* baru dikembangkan pada tahun 2013. *Word2vec* merupakan salah satu metode berbasis *vector* untuk mencari kemiripan semantik sehingga penulis menggunakan metode tersebut untuk membuat sistem dan menggabungkan dengan algoritma *Cosine Similarity*. Pendekatan *Word2vec* dipilih pada penelitian tugas akhir ini sebab *Word2vec* dapat mempelajari hubungan antar

kata-kata secara otomatis dan dapat mempelajari berbagai macam *text* sebagai data latih [10]. *Word2vec* merupakan representasi kata dalam bentuk vektor yang digunakan untuk menghasilkan *word embeddings* [11]. Sedangkan *Cosine Similarity* digunakan untuk menghitung nilai kemiripan antar kalimat dan menjadi salah satu teknik untuk mengukur kemiripan teks yang popular. Kelebihan dari algoritma *Cosine Similarity* adalah tidak terpengaruh pada panjang pendeknya suatu dokumen. Hasil akhir yang akan diberikan dalam sistem ini adalah presentase nilai *similarity* dokumen yang diuji dengan dokumen pembanding sebanyak 116 abstrak yang didapatkan dari *repository* Institut Teknologi Telkom Purwokerto.

1.2 Rumusan Masalah

Rumusan masalah dari penelitian ini adalah metode *Word2vec* banyak digunakan pada teks Bahasa Inggris dan belum banyak digunakan pada teks Bahasa Indonesia terutama tentang kemiripan dokumen. Berdasarkan rumusan masalah tersebut maka munculah pertanyaan penelitian sebagai berikut:

1. Apakah metode *Word2vec* dapat digunakan untuk menghitung tingkat kemiripan pada dokumen dalam teks Bahasa Indonesia?
2. Berapa nilai kemiripan pada metode *Word2vec* dibandingkan dengan sistem plagiarism detector?

1.3 Tujuan Penelitian

Berdasarkan dekomposisi masalah dalam rumusan masalah tujuan pada penelitian ini adalah sebagai berikut.

1. Menerapkan metode *Word2vec* kedalam teks Bahasa Indonesia.
2. Mengetahui nilai kemiripan pada dokumen menggunakan metode *Word2vec*.

1.4 Manfaat

Berdasarkan tujuan penelitian yang sudah dijelaskan, adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Mempercepat proses pengoreksian dengan jumlah besar dengan cepat dan efektif.
2. Mengetahui presentase kemiripan sehingga dapat digunakan sebagai bahan pertimbangan untuk mendeteksi adanya tindakan meniru.

1.5 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Dokumen yang digunakan dalam penelitian ini adalah dokumen Abstrak Tugas Akhir dan laporan Praktik Kerja Lapangan.
2. Tipe dokumen yang digunakan dalam penelitian ini adalah .pdf.
3. Tidak memperhatikan kesalahan penulisan dalam dokumen.

BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Pada penelitian ini penulis menggunakan studi literatur sebagai sarana untuk kelengkapan data sekaligus untuk mempertajam masalah yang penulis kaji. Penulis telah mengkaji delapan jurnal penelitian terdahulu. Kedelapan jurnal tersebut penulis pilih bedasarkan topik dan tema yang sesuai untuk penelitian ini. Berikut merupakan penjelasan lebih lanjut.

Agung Wardhana Z. Nasution, dkk [6]. Melakukan Analisa dan implementasi perhitungan nilai semantic kemiripan kata pada ayat Al-Quran terjemahan Bahasa inggris menggunakan pendekatan *Word alignment* dan *vector semantic Word2vec* [6]. Sebagai evaluasi dari penelitian ini digunakan model *Support Vector Regression*. Hasil dari penelitian ini adalah pengujian Sistem yang menggunakan data 2017 sebagai data test dengan inputan 350 ayat terjemahan Bahasa Inggris Ibnu Katsir dan 50 ayat indeks sematik memperoleh nilai korelasi *alignment* 0,81958, nilai *Word2vec* 0,64395 dan nilai korelasi SVR yang diperoleh 0,81221. Nilai SVR ini, memiliki nilai lebih kecil 0,00737 dibandingkan *alignment* TF-IDF dikarenakan kompleksitas pada masing-masing data dan pengaruh fitur *Word2vec*. hal tersebut tidak terjadi jika data *test* menggunakan data 2016 dengan fitur tanpa *stemming* pada *Word2vec* menghasilkan nilai SVR 0,9106 lebih besar 0,0016 dari hasil *alignment* [6]. Kesamaan penelitian ini dengan penelitian penulis adalah metode yang digunakan yaitu *Word2vec*. Kekurangan dalam penelitian ini adalah kompleksitas pasangan ayat yang tidak dapat diidentifikasi oleh sistem. Sistem yang dibangun tidak sepenuhnya dapat mengidentifikasi kata-kata serapan yang berhubungan dengan AlQuran kedalam bentuk *vector*. Namun penelitian pada [11] mencantumkan alasan tersebut pada bagian analisis sehingga transparansi dalam penelitian tersebut dapat dilihat dengan jelas dan kejujuran menjadi nilai lebih dalam melaksanakan penelitiannya.

Kemal Ade Sekarwati, dkk [12]. Membangun sistem pengukuran kemiripan dokumen berbahasa Indonesia dengan menggunakan metode *Latent Semantic Analysis* (LSA) sedangkan *tools* yang digunakan untuk penelitian ini adalah *Gensim*. *Gensim* merupakan *open-source* model ruang vektor dan *toolkit* topik modeling [12].

Hasil dari penelitian ini yaitu Pengujian dokumen dilakukan terhadap dokumen yang diduga mempunyai kemiripan dengan dokumen lain. Dokumen 1 merupakan dokumen yang diuji, sedangkan dokumen 2 merupakan dokumen asli. Menghasilkan nilai persentase kemiripan 100% bagi dokumen yang sama [12].

Sugiyamta [13] membuat sistem deteksi kemiripan dokumen (*document similarity*) menggunakan Algoritma *Cosine Similarity* dan teknik mengelompokkan dokumen dengan Algoritma *Single Pass Clustering*. Hasil dari penelitian adalah pengukuran kemiripan dengan *Cosine Similarity* untuk dokumen abstrak memiliki tingkat akurasi 99%. *Single Pass Clustering* dapat membantu menemukan dokumen yang ada dalam satu *Cluster* dengan *query* yang dimasukan oleh pengguna [13].

Na'fairul Hasna [8] melakukan percobaan mendeteksi kemiripan dokumen teks menggunakan algoritma *Levenshtein Distance* sehingga dapat digunakan untuk membantu menentukan plagiarisme. Tipe dokumen yang diuji adalah .pdf .docx dan .txt. Hasil dari penelitian ini adalah pengukuran kemiripan dengan algoritma *Levenshtein Distance* menghasilkan nilai *similarity* yang tinggi yaitu 85% untuk dokumen yang tingkat kemiripannya tinggi. Sedangkan untuk dokumen yang tingkat kemiripannya rendah menghasilkan nilai *similarity* kurang dari 40% [8].

Mochammad Iran Dary [9] melakukan percobaan untuk mengetahui keterkaitan antar ayat didalam al-Quran menggunakaan metode *latent semantic analysis*. Menggunakan algoritma *Cosine Similarity* untuk membandingkan jarak dari 2 buah *vector*. Hasil dari penelitian ini adalah metode *latent semantic analysis* cukup efektif untuk mencari kesamaan antar ayat pada alquran yang berupa *short text*. Terlihat dari nilai *F-Measure* yang paling tinggi adalah 14% [9].

Radiant Victor, dkk [14] melakukan mengembangkan sebuah aplikasi yang mengimplementasikan *Cosine Similarity* yang berguna untuk mengukur kesamaan teks berdasarkan kemunculan kata-kata dalam teks tersebut dan algoritma *Smith-Waterman* yang berfungsi untuk menghitung kemiripan teks berdasarkan urutan kata [14]. Dalam proses mengubah kata-kata berbahasa indonesi menjadi kata-kata dasar (*stemming*) menggunakan algoritma Nazief-Adriani. Hasil dari penelitian ini dapat mendeteksi tingkat kemiripan teks dari sangat mirip hingga sangat tidak mirip

berdasarkan kemunculan kata di dalamnya dengan menggunakan *Cosine Similarity* dan algoritma *Smith-Waterman*.

Sedangkan Kadek Bayu, dkk [10] Membangun sebuah sistem yang dapat menentukan prioritas opini dari suatu fitur produk berdasarkan nilai kemiripan atau *similarity* suatu kata. sistem yang akan dibangun menggunakan pendekatan *Word2vec* dan pendekatan *WordNet* untuk proses klasifikasi opini dimana *Word2vec* merupakan model representasi kata dalam bentuk *vector*, dan pada penelitian ini *Word2vec* digunakan untuk melakukan klasifikasi opini pada fitur suatu produk [10]. Hasil dari penelitian ini yaitu Penerapan pendekatan *Word2vec* pada proses klasifikasi memiliki nilai akurasi yang lebih tinggi dibandingkan penerapan pendekatan *WordNet*, yaitu dengan rata-rata persentase akurasi pada *Word2vec* sebesar 43,55% sedangkan rata-rata persentase akurasi pada WordNet sebesar 41,48%, dengan selisih 2,07% [10].

Tabel 2. 1 Penelitian Terdahulu

No	Penulis	Penulis	Langkah Perancangan Sistem	Objek Penelitian	Perbedaan dengan penelitian yang akan dilakukan
1	A. W. Z. Nasution dkk (2017)	Analisis dan Implementasi Perhitungan *Semantics Similarity* Pada Ayat AL-Quran Dengan Pendekatan *Word Alignment* Berdasarkan *Support Vector Regression*	*Preprocessing alignment, contextual Similarity,* metode ekstraksi TF-IDF, *preprocessing Word2vec,* perhitungan *semantic Word2vec*, SVR	Terjemahan ayat suci Al-Quran berbahasa inggris	penelitian ini data yang digunakan berbahasa Ingris dan korpus yang dibuat juga Bahasa Ingris. Sedangkan penulis menggunakan Bahasa Indonesia
2	Sekarwati dkk (2013)	Pengukuran Kemiripan Dokumen Dengan *Tools Gensim*	*Preprocessing,* metode ekstraksi TF-IDF, *Latent Semantic Indexing,* perhitungan kemiripan *Cosine Similarity, Dice's Similarity*	Dokumen pribadi	Penelitian ini dilakukan menggunakan model *latent semantic analysis* untuk mendapatkan nilai kemiripan. Sedangkan penelitian yang dilakukan penulis menggunakan *Word2vec* dan *cosine similarity*
3	Sugiamta (2015)	Sistem Deteksi Kemiripan Dokumen Dengan Algoritma *Cosine Similarity* Dan *Single Pass Clustering*	*Preprocessing, Term indexing , term similarity* metode *cosine,* *Clutering*	550 Abstrak Skripsi	Penelitian ini melakukan pengelompokan nilai similaritas paling tinggi menggunakan *single pass clustering*. Sedangkan penelitian yang dilakukan melakukan pengelompokan dari nilai tertinggi ke terendah secara langsung

No	Penulis	Penulis	Langkah Perancangan Sistem	Objek Penelitian	Perbedaan dengan penelitian yang akan dilakukan
4	N. Ariyani dkk (2016)	Aplikasi Pendeteksi Kemiripan Isi Teks Dokumen Menggunakan Metode *Levenshtein Distance*	*Preprocessing,* Metode *Levenshtein Distance, term similarity*	Dokumen Pribadi	Penelitian ini menggunakan *Levenshtein Distance* untuk mendapatkan nilai *similarity* dari dokumen. Data yang digunakan dokumen berplagiat yang diambil dari artikel/berita. Sedangkan penelitian yang akan dilakukan menggunakan *cosine similarity* untuk mendapatkan nilai kemiripan
5	M. Dary (2017)	Analisis dan Implementasi *Short Text Similarity* dengan Metode *Latent Semantic Analysis* Untuk Mengetahui Kesamaan Ayat al-Quran	*Preprocessing, Text Preprocessing, Term Document,* Implementasi *Latent Semantic Analysis, Reconstructed matrix,* evaluasi	Terjemahan ayat suci Al-Quran berbahasa Indonesia	Penelitian ini menerapkan *latent semantic analysis* untuk mencari kesamaan antar ayat pada al-Quran yang berupa *short text*. Sedangkan penelitian yang akan dilakukan menggunakan dokumen abstrak.
6	R. Imbar dkk (2014)	Implementasi *Cosine Similarity* dan Algoritma *Smith-Waterman* untuk Mendeteksi Kemiripan Teks	*Preprocessing, Term Similarity Cosine Similarity, Term Similarity* algoritma *Smith Waterman*	Dokumen Pribadi	penelitian ini dapat mendeteksi tingkat kemiripan teks dari sangat mirip hingga sangat tidak mirip berdasarkan urutan kata pembentuknya dengan menggunakan algoritma *Smith-Waterman*. Sedangkan penelitian yang akan dilakukan tidak berdasarkan ututan kata pembentuknya.

No	Penulis	Penulis	Langkah Perancangan Sistem	Objek Penelitian	Perbedaan dengan penelitian yang akan dilakukan
7	B Surarso dkk (2016)	Analisa Performa Metode *Cosine* dan *Jacard* Pada Pengujian Dokumen	*Preprocessing, Indexing,* Hitung Similaritas metode *Jaccard* dan *cosine*, *klastering*	Abstrak Skripsi	penelitian ini membandingkan bandingkan performa dari metode *Cosine* dan *Jaccard* untuk menguji tingkat kemiripan dokumen dalam bentuk abstrak
8	I.Bayu dkk (2017)	Klasifikasi Opini Pada Fitur Produk Berbasis *Graph*	*Dataset*, Ekstraksi, Klasifikasi *Word2vec* dan *WordNet*	Data review produk dari *paper*"*Mining and Summarizing Customers Reviews*	Penelitian ini menganalisis opini yang berupa opini positif atau negatif secara otomatis. Sedangkan penelitian yang akan dilakukan mendeteksi kemiripan teks dalam dokumen.

2.2 Landasan Teori

2.2.1 Plagiarisme

Plagiarisme merupakan tindakan kriminal yang sering terjadi dalam dunia akademis. Plagiarisme itu sendiri berasal dari kata latin "Plagiarus" yang berarti penculik dan "Plagiare" yang berarti mencuri. Jadi, secara sederhana plagiat berarti mengambil ide, kata-kata, dan kalimat seseorang dan memposisikannya sebagai hasil karyanya sendiri atau mengunakan ide, kata-kata, dan kalimat tanpa mencantumkan sumber dimana seorang penulis mengutipnya [1].

Menurut Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 17 Tahun 2010 mengatakan bahwa plagiat adalah perbuatan sengaja atau tidak sengaja dalam memperoleh atau mencoba memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat dan memadai [15]. Dalam Kamus Besar Bahasa Indonesia (2008) Plagiat adalah

pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat) sendiri"[16].

a. Jenis plagiarisme berdasarkan aspek yang dicuri

Dalam jenis plagiarismee ini terdiri dari 4 kategori yaitu plagiarisme ide, isi, kata, kalimat, paragraph, dan plagiarisme total. Plagiarisme ide sering dihubungkan dengan laporan hasil penelitian replikatif. Penelitian replikatif adalah penelitian yang secara garis besar mengulang penelitian orang lain, dengan maksud untuk menambah data, menguji hipotesis apakah hasil yang sudah ditemukan dalam suatu populasi berlaku pula untuk populasi yang lain [1]. Plagiarisme isi fabrikasi dan falsifikasi. Fabrikasi adalah tindakan membuat data yang tidak ada menjadi ada. Sedangkan falsifikasi adalah mengubah data dengan maksud agar sesuai dengan yang dikehendaki oleh peneliti tersebut [17]. Plagiarismee kata, kalimat, paragraph merupakan jenis plagiarismee yang mengubah sebagian kecil dalam sebuah karya tulis, seperti menganti sebagian kalimat atau paragraf atau meniru seluruh isinya dengan bahasa yang berbeda [1]. Plagiarismee total merupakan plagiarismee yang dibuat tanpa merubah total isi dari penelitian tersebut dan sama persis dengan karya orang lain. plagiarismee ini merupakan plagiarismee yang paling berata [1].

b. Klafisikasi berdasarkan proporsi atau persentasi kata, kalimat, paragraf yang dibajak

Plagiarismee ringan: <30%

Plagiarismee sedang: 30-70%

Plagiarismee berat atau total : >70%

(angka-angka tersebut tentu dibuat secara arbitrer berdasarkan "kepantasan" tanpa dasar kuantitatif yang definitif) [1].

2.2.2 Dokumen

Menurut Kamus Besar Bahasa Indonesia (KBBI) menyebutkan dokumen adalah sesuatu yang tertulis atau tercetak yang dapat dipergunakan sebagai bukti atau keterangan. Dokumen merupakan salah satu hal yang sangat penting karena merupakan sumber informasi yang diperlukan oleh suatu instansi, organisasi, atau Negara [16]. Tanpa dokumen kita akan kehilangan data-data yang diperlukan untuk

kegiatan kantor/organisasi masa yang akan datang. Jenis-jenis format dokumen adalah sebagai berikut:

a. DOC (.doc)

File Doc merupakan dokumen pengolah kata yang diciptakan oleh program Microsoft Word. Jenis *file* ini bisa berisi susunan teks, image, tabel, grafik, *chart*, format halaman, dan pengaturan cetak. Setiap *file* DOC selalu diikuti dengan ekstensi.doc [18]. Jenis *file* DOC secara otomatis tercipta ketika membuat dokumen pada program MS Word dan menyimpannya. Ekstensi .doc memang tidak selalu muncul di akhir nama *file*. Namun, untuk mengetahui identitas atau jenis *file* tersebut, anda bisa melihat kolom *Type* dimana pada kolom tersebut diinformasikan jenis *file* adalah Microsoft Word Document [18]

b. DOCX (.docx)

File DOCX dirancang untuk membuat dokumen yang kontennya dapat diakses dengan mudah. Hal ini dimaksudkan bahwa sebuah dokumen dengan format DOCX bisa memiliki format *file* yang beragam. Sebagai contoh, misalnya dokumen teks disimpan menggunakan *file plain text* dan dokumen gambar disimpan sebagai *file* gambar individual dalam *file* DOCX. Jenis *file* ini memiliki ekstensi .docx [18].

Berbeda dengan *file* DOC yang menyimpan dokumen data dalam *file* binary tunggal, *file* DOCX diciptakan menggunakan format Open XML yang menyimpan dokumen sebagai sebuah kumpulan *file* dan folder yang terpisah dalam paket kompresi zip. *File* DOCX berisi *file* XML dan tiga folder, yakni *docProps*, *Word*, dan *_rels* yang menjaga dokumen properties, konten, dan hubungan di antara *file*. *File* DOCX sebenarnya merupakan kumpulan *file* XML yang dikompres. *File* XML dapat ditampilkan secara individual dengan mengubah ulang nama *file* .docx ke dalam .zip dan memperlakukan *file* tersebut sebagai *file* .zip [18].

c. TXT (.txt)

File TXT merupakan standar dokumen teks yang berisi rangkaian teks yang tidak terformat. Jenis *file* ini diakui oleh sembarang program pengolah kata. *File* jenis ini umumnya dihasilkan oleh sebuah *software* atau program yang

bernama *Notepad* [18]. Jenis *file* TXT disebut juga sebagai *file* plain *text*. Berbeda dengan *file* RTF, pada *file* TXT tidak didukung oleh format teks seperti pengaturan *style* teks, baik tebal, miring, ataupun garis bawah dan pengaturan *style font*, baik jenis maupun ukuran *font* untuk teks tertentu [18].

d. Pdf (.pdf)

File Pdf atau *Portable Document Format* merupakan standar untuk distribusi elekronis yang dibuat oleh Adobe. *File* ini menyediakan *layout*, tipografi, gambar *bitmap*, transparasi, dan gambar *vector* dengan hasil yang sangat baik. PDF juga mempunyai kualitas dan presisi *layout* untuk proses yang baik dan sangat ideal untuk distribusi secara *online*. *File* PDF merupakan format yang sangat umum digunakan pada *World Wide Web* [18].

2.2.3 Text Mining

Text mining adalah proses penemuan akan informasi atau *trend* baru yang sebelumnya tidak terungkap dengan memproses dan menganalisis data dalam jumlah besar. Dalam menganalisa sebagian atau keseluruhan *unstructured text*, *text mining* mencoba untuk mengasosiasikan satu bagian *text* dengan yang lainnya berdasarkan aturan aturan tertentu. Hasil yang di harapkan adalah informasi baru atau "*insight*" yang tidak terungkap jelas sebelumnya [19].

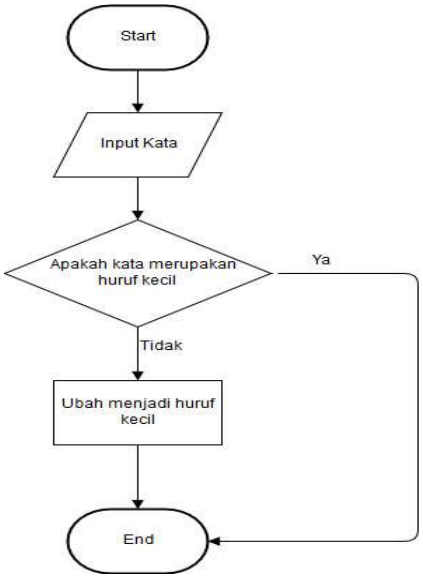
Seperti halnya data mining, *text mining* juga menghadapi masalah yang sama, termasuk jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data "*noise*." Berbeda dengan data *mining* yang utamanya memproses *structured* data, data yang digunakan *text mining* pada umumnya dalam bentuk *unstructured*, atau minimal semi-*structured*, *text*. Akibatnya, *text mining* mempunyai tantangan tambahan yang tidak ditemukan dalam data *mining*, seperti struktur *text* yang complex dan tidak lengkap, arti yang tidak jelas dan tidak standard, dan bahasa yang berbeda ditambah translasi yang tidak akurat. Dikarenakan structured data ditujukan agar mudah di proses komputer secara *automatic*, *pre-process* data di data *mining* jauh lebih mudah dilakukan dari pada *unstructured text* [19]. *Text* di ciptakan bukan untuk di gunakan oleh mesin, tapi untuk dikonsumsi manusia langsung. Karena itu, pada umumnya "*Natural Language Processing*" digunakan untuk memproses *unstructured text*.

2.2.4 Text Processing

Struktur data yang baik dapat memudahkan proses komputerisasi secara otomatis. Pada *text mining*, informasi yang akan digali berisi informasi-informasi yang strukturnya sembarang [20]. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk proses dalam data *mining*, yang biasanya akan menjadi nilai-nilai numerik. Proses ini sering disebut *text preprocessing* [21]. Setelah data menjadi data terstruktur dan berupa nilai numerik maka data dapat dijadikan sebagai sumber data yang dapat diolah lebih lanjut. Berberapa proses yang dilakukan adalah sebagai berikut:

a. *Case folding*

Case folding merupakan proses pertama dari rangkaian *preprocessing dokumen*. Dalam proses ini mengubah semua huruf dalam dokumen menjadi huruf kecil. Hanya huruf a sampai dengan z yang diterima [21]. Gambar 2.1 menunjukan *flowchart* dari proses *case folding.*



Gambar 2. 1 Tahapan proses *case folding*.

b. *Tokenizing*

Tahap *Tokenizing* adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya [20]. Karakter selain huruf akan dianggap *delimiter* dan akan dihilangkan atau dihapus untuk proses mendapat kata-kata penyusun teks. Gambar 2.2 menunjukan *flowchart* tahapan *tokenizing*.