

# Penentuan *Hoax* pada Artikel Politik Berbahasa Indonesia di Sosial Media dengan *Similarity Jaccard* dan Algoritma *Stemming*

Goenawan Brotosaputro<sup>[1]</sup>, Wiwin Windihastuty<sup>[2]\*</sup>, Rezza Anugrah Mutiarawan<sup>[3]</sup>  
Fakultas Teknologi Informasi<sup>[1], [2], [3]</sup>  
Universitas Budi Luhur  
Jakarta, Indonesia  
[wiwin.windihastuty@budiluhur.ac.id](mailto:wiwin.windihastuty@budiluhur.ac.id)

**Abstrak**—Pesatnya perkembangan teknologi di era globalisasi saat ini membawa pengaruh besar dalam proses pencarian informasi. Kita dapat melihat pertumbuhan yang sangat besar pada volume berita *online* yang tersedia pada jaringan *internet*, maupun pada jaringan lainnya. Salah satunya adalah sosial media yang memiliki banyak informasi yang mengenai artikel-artikel berita atau artikel-artikel informasi lainnya. Berita merupakan sumber informasi mengenai kejadian terkini yang mana dapat ditemukan pada internet dan media sosial. Saat ini berita-berita yang disebarakan terutama berita mengenai politik yang dapat mengakibatkan salah penafsiran karena berita tersebut belum tentu benar atau salah sehingga dibutuhkan pengklasifikasian artikel politik apakah termasuk dalam kategori *hoax* atau *non hoax*. *Hoax* adalah berita kebohongan yang disebar untuk memperoleh kepercayaan agar masyarakat akan merasa yakin bahwa konten tersebut benar. Dampak lain dari *hoax* dapat merugikan emosi hingga finansial masyarakat. Proses klasifikasi *hoax* menggunakan tahap *preprocessing* yang terdiri dari *tokenization* dan *stemming*. Dilanjutkan dengan proses pembobotan kata dan *jaccard similarity* hingga proses klasifikasi dengan menggunakan metode *Vector Space Model (VSM)*. Hasil evaluasi pada penelitian ini menggunakan *confusion matrix*, dimana diperoleh hasil *precision* sebesar 0,92 *recall* sebesar 0,80 dan akurasi didapatkan sebesar 87 %.

**Kata kunci** —Algoritma *Stemming*, *Stopword Removal*, *Similarity Jaccard*, *Machine Learning*, *Hoax*.

**Abstract**—The rapid development of technology in the current global era has a major influence in the process of finding information. We can see a very large growth in the volume of online news available on the internet, as well as on other networks. One of them is social media which has a lot of information about news articles or other information articles. News is a source of information about current events which can be found on the internet and social media. Currently the news that is spread is mainly news about politics which can lead to misinterpretation because the news is not necessarily true or false, so it is necessary to classify political articles as belonging to the *hoax* or *non-hoax* category. *Hoax* is false news that is spread to gain trust so that people will feel confident that the content is true. Another impact of hoaxes can be detrimental to people's emotions and finances. The *hoax* classification process uses the *preprocessing* stage which consists of *tokenization* and *stemming*. Followed by the process of weighting words and *jaccard similarity* to the classification process using the *Vector Space Model (VSM)* method. The evaluation results in this study using a

*confusion matrix*, where the precision results obtained by 0.92 recall of 0.80 and accuracy obtained by 87%.

**Keywords**— *Stemming Algorithm*, *Stopword Removal*, *Similarity Jaccard*, *Machine Learning*, *Hoax*.

## I. PENDAHULUAN

Pesatnya perkembangan teknologi informasi dan komunikasi di era global saat ini membawa pengaruh besar dalam proses pencarian informasi diberbagai belahan dunia. Semakin bertambahnya kebutuhan manusia akan informasi, maka semakin banyak pula informasi yang ada dan tersebar khususnya melalui media sosial *online* sehingga akan dengan cepat diterima pada masyarakat yang menggunakan media sosial *online*.

Media sosial *online* salah satu wadah komunikasi yang dimanfaatkan sebagai pergaulan sosial secara *online*, pada media sosial *online* para pengguna dapat menerima informasi, saling berkomunikasi, berinteraksi, *sharing*, dan berbagai kegiatan lainnya. Media sosial *online* yaitu teknologi yang dapat mengubah komunikasi menjadi dialog interaktif. Beberapa media sosial yang banyak penggunanya yaitu facebook, twitter, blog dan media sosial yang lainnya.

Media sosial adalah sarana bagi konsumen untuk berbagi informasi *text*, gambar, video satu sama lain dan berperan sebagai media penyebaran informasi dari berbagai macam informasi yang tersebar [1].

Dengan perkembangan teknologi informasi yang begitu pesat, internet berubah menjadi sumber informasi yang dengan mudah digunakan dan dapat mendorong penyebaran artikel berita, baik itu informasi berupa berita fakta maupun informasi bohong atau *hoax*. Keberadaan berita *hoax* ini memang telah membuat resah pengguna media sosial *online* pada khususnya dan hampir sebagian besar masyarakat Indonesia pada umumnya dan juga perkembangan teknologi informasi berpengaruh besar terhadap masyarakat Indonesia.

Perkembangan teknologi informasi turut serta mendorong penyebaran berita atau informasi *hoax* dengan tujuan membuat opini publik, menggiring opini, membentuk persepsi, juga

untuk bersenang-senang yang menguji kecerdasan kecermatan pengguna internet dan media sosial.

Di Indonesia, hoax marak terjadi sejak pemilihan presiden 2014 sebagai dampak gencarnya kampanye di media sosial. Hoax bermunculan serta menjatuhkan citra lawan politik alias kampanye hitam atau kampanye negatif. Bila berdasarkan rekapitulasi tahunan, Kominfo menerima aduan konten hoax sebanyak 733 laporan sepanjang tahun 2018 di aplikasi pesan instan tersebut. Sementara bila dilihat dari Agustus 2018 sampai 21 Januari 2019, Kominfo menerima laporan hoax yang disebarakan melalui WhatsApp sebanyak 43 konten. Hasil pemantauan Direktorat Pengendalian Aplikasi Informatika Direktorat Jenderal Aplikasi Informatika, laporan terbanyak terjadi pada bulan Oktober 2018. Ada sebanyak 16 konten hoax yang disebarakan melalui platform WhatsApp. Pada bulan Agustus 2018 terdapat laporan dua konten hoax, September 2018 ada lima konten hoax, November 2018 sebanyak delapan laporan konten dan Desember 2018 sebanyak 10 laporan konten hoax. Sementara, sampai pada 21 Januari 2019 terdapat dua laporan konten misinformasi yang beredar aplikasi milik facebook itu. Informasi mengenai jumlah aduan yang masuk ke Sistem Kominfo seperti pada Gambar 1 dibawah ini;



Gambar 1 Total Aduan Konten Negatif , 2016 – 2018

Plt kepala Biro Humas Kementerian Kominfo menjelaskan, pengelola pengaduan konten negatif yang disebarakan melalui aplikasi pesan instan dimana pada tahun 2016 terdapat 14 aduan konten, di tahun selanjutnya jumlah aduan meningkat menjadi 281 aduan, sementara pada tahun 2018 sebanyak 1440 laporan yang berkaitan dengan konten negatif.

Tujuan dari penelitian ini yaitu terbentuknya sebuah aplikasi yang dimana dapat mengetahui sebuah artikel mengenai informasi-informasi berita politik dan sistem akan melakukan penentuan artikel secara sistematis kedalam sebuah label atau kategori *hoax* atau *non hoax* yang telah ditentukan oleh sistem. Manfaat dari penelitian ini memberikan ketepatan dalam kinerja menentukan artikel *hoax* atau *non hoax* pada media sosial, dan mempermudah pemberian label atau pengkategorian *hoax* dan *non hoax* pada artikel-artikel berita politik.

Pada Penelitian [2] dengan judul Metod Klasifikasi Rocchio untuk Analisa *Hoax* dengan masalah keterbatasan rentang perhatian manusia ditambah dengan informasi yang sangat

deras di media sosial dianggap menjadi penyebab *hoax* gampang menyebar. Penelitian menyebutkan, seseorang cenderung melihat “bias informasi” dan hanya menaruh perhatian, serta menyebarkan informasi yang sesuai dengan kepercayaannya. Bahkan meski infomasi tersebut palsu. Penelitian ini dilakukan menggunakan metode klasifikasi Multinomial Naïve Bayes dengan hasil pengujian menggunakan 30 artikel *hoax* dan 30 artikel *non-hoax* didapatkan akurasi Rocchio sebesar 83.501% sedangkan Multinomial Naïve Bayes sebesar 65.835 %.

Penelitian yang dilakukan [3] yang berjudul Deteksi Hoaks Pada Media Sosial Berbasis *Text Mining Classification System*. Masalah dalam penelitian ini rendahnya literasi publik terhadap pesan-pesan di media sosial. Metode yang digunakan pada penelitian ini yaitu *TextRank* dan Algoritma *Cosine Similarity* dengan hasil pengujian menggunakan 20 data didapatkan akurasi dengan *confusion matrix* sebesar 75 %.

Penelitian ini dilakukan oleh (Weddingrum *et al.*, 2018) [4] dengan judul Deteksi Kontek *Hoax* Berbahasa Indonesia Pada Media Menggunakan Metode *Levenshtein Distance*. Masalah dalam penelitian ini maraknya tren *hoax* yang meracuni pemberitaan terutama pada media sosial. Metode yang digunakan pada penelitian ini yaitu *Levenshtein Distance* dengan hasil pengujian 20 artikel *hoax* dan 20 artikel *non-hoax* didapatkan akurasi sebesar 70 %.

## II. METODE PENELITIAN

Dalam suatu penelitian memerlukan sebuah teknik atau cara untuk mencari, mendapatkan, mengkoleksi dan *me-record* suatu infromasi.

Dalam pengembangan aplikasi ini merujuk pada *waterfall model*. Ada lima tahap utama dalam metode ini. Disebut *waterfall model* karena diagram tahapan mirip dengan proses yang bertingkat. Metodologi *waterfall model* secara ringkas menurut [5], merupakan sebuah model dalam perencanaan piranti lunak. Model *waterfall* adalah sebuah contoh dari proses perencanaan, dimana semua proses kegiatan harus direcanakan dan di jadwalkan terlebih dahulu sebelum dikerjakan. Tahapan dalam *waterfall* model adalah sebagai berikut :

1. Analisis  
Tahapan ini bertujuan untuk mencari kebutuhan pengguna dan organisasi serta menganalisa kondisi yang ada (sebelum diterapkan sistem yang baru).
2. Design  
Tahapan *design* bertujuan untuk menentukan spesifikasi detail dari komponen – komponen program (manusia, perangkat keras, perangkat lunak, jaringan dan data) dan produk – produk informasi yang sesuai dengan hasil tahap analisis.
3. Implementasi  
Tahapan ini adalah tahapan untuk mendapatkan atau mengembangkan perangkat keras dan perangkat lunak (pengkodean program), melakukan pengujian, pelatihan

dan perpindahan ke sistem yang baru.

#### 4. Perawatan

Tahan ini dilakukan pada saat sistem udah dioperasikan. Pada tahap perawatan ini dilakukan proses *monitoring*, evaluasi dan perubahan bila diperlukan.

Pada penelitian ini penulis menggunakan metode *Vector Space Model* dengan Algoritma Nazief dan Andriani dan fungsi kesamaan pada *Jaccard*. Metodologi penelitian mencakup metode penelitian, yaitu suatu teknik atau cara yang digunakan sehubungan dengan penelitian yang dilakukan mempunyai step – step yang sistematis. Metode penelitian menyangkut masalah pengerjaannya, yaitu suatu teknik kerja untuk dapat memahami yang menjadi tujuan penelitian yang bersangkutan. Metode penelitian meliputi prosedur penelitian dan teknik penelitian. Dalam penelitian mencoba melakukan pendekatan metode penelitian eksperimen, dengan cara menguji coba terhadap bobot kata dalam pengklasifikasian dokumen, kemudian mencari keefektifitasannya dengan basis ketepatan (*presicion*).

Sample adalah sebagian dari jumlah dan karakteristik yang dimiliki oleh populasi tersebut, ataupun bagian kecil dari anggota populasi yang diambil menurut prosedur tertentu sehingga dapat mewakili populasinya. Teknik sampling merupakan teknik pengambilan sampel, terdapat berbagai macam teknik sampling untuk menentukan sampel dapat dipakai dalam penelitian. Untuk menguji tingkat efektifitas hasil penentuan artikel dari sebuah model, maka metode pengambilan sampel pada penelitian ini menggunakan metode *Cluster Sampling*. Sampel dalam penelitian ini adalah 200 artikel yang didapatkan dari sumber sosial media online yaitu *turnbackhoax.com* sebagai media yang memvalidasi artikel *hoax* dan *non-hoax* dan beberapa dari media yang terdaftar di *dewanpers.or.id*. Artikel yang digunakan berupa teks tanpa disertakan gambar. Alasan mengapa mengambil dari sosial media *online* dikarenakan banyak informasi yang tersebar melalui media sosial dengan struktur bahasa indonesia yang digunakan. Rencana penelitian dari sampel akan dibuat dua kategori atau label yaitu *hoax* dan *non-hoax*.

Dalam suatu kegiatan penelitian dibutuhkan pengkoleksian data-data berupa informasi dan referensi untuk bahan yang dapat mendukung materi uraian dan pembahasan, hal ini menjadikan salah satu step atau tahapan yang sangatlah penting dalam melakukan kegiatan penelitian. Teknik pengumpulan data yang benar akan menghasilkan data yang mempunyai kredibilitas tinggi, namun jika pengumpulan data yang salah akan menghasilkan data yang tidak memiliki kredibilitas tinggi. Pada tahapan ini adalah artikel berita pada media digital yaitu *turnbackhoax.com* dan *dewanpers.co.id* yang merupakan situs *hoax* dan berita dengan konten berbahasa indonesia, dengan mengambil sample 200 artikel dari salah satu kategori yang berada dari sosial media *online* yaitu kategori politik. Dari 200 artikel yang telah di susun dari tiap-tiap artikel dengan tujuan untuk mempermudah pengujian kebenaran dan keakuratan.

Instrument yang digunakan dalam proses penelitian ini, berupa dokumen penunjang seperti buku, jurnal, skripsi dan tesis tentang penelitian terdahulu. Dokumen yang digunakan

untuk mengklasifikasi yang mengimplementasikan *vector space model* dengan algoritma *stemming* Nazief dan Andriani dan fungsi kesamaan *jaccard*.

Teknik analisis yang digunakan untuk penelitian ini menggunakan pendekatan Algoritma *Stemming* Nazief dan Andriani. Rancangan model pada penelitian ini dibuat memiliki proses sebagai berikut: (1) Tahap *text preprocessing*, yaitu melakukan proses *tokenizing*, *stopword removal* dan proses *stemming* menggunakan Algoritma Nazief dan Andriani. (2) Melakukan proses pencaian pembobotan pada setiap kata dari artikel tersebut dengan menggunakan fungsi kesamaan *Jaccard*. (3) Proses penghitungan nilai *similarity*, dimana untuk mencari perbedaan artikel yang dimiliki dengan *query* dengan didalamnya terdapat Algoritma Nazief dan Andriani dan fungsi kesamaan *Jaccard* untuk menghitung nilai *similarity* pada artikel.

Pengklasifikasian artikel merupakan sebuah model untuk menyusun artikel dengan ketentuan-ketentuan yang telah disepakati. Model *prototipe* ini dipilih karena mempunyai struktur yang sesuai dalam mengembangkan simulasi model pengklasifikasian. *Prototipe* klasifikasi artikel yang akan dibuat dalam penelitian ini yaitu menggunakan *web based application*.

Pada penelitian ini melakukan analisis tingkat efektifitas hasil dari klasifikasi artikel, dengan membutuhkan sebuah model klasifikasi yaitu *web based application* yang dimana dapat menerapkan *vector space model*, proses algoritma *stemming* Nazief dan Andriani dan proses pembobotan berdasarkan nilai *similarity*. Proses yang digunakan dalam penentuan artikel diantaranya yaitu: (1) Upload dokumen (2) Tokenization (3) Removing *stopword* (4) Proses *stemming* (5) Proses pembobotan kata (6) Proses perhitungan nilai kemiripan (*similarity*)

Teknik pengujian sistem yang digunakan penulis dalam penelitian ini yaitu dengan ketentuan sebagai berikut : (1) Pengujian ketepatan (*Precision*) adalah perbandingan jumlah artikel relevan yang didapatkan sistem dengan jumlah seluruh artikel yang terambil oleh sistem baik relevan maupun tidak relevan.

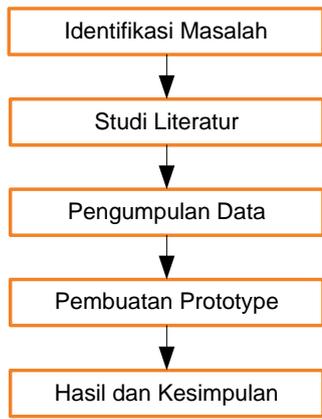
$$Precision : \frac{\text{Jumlah artikel yang relevan dengan query dan terambil}}{\text{Jumlah seluruh artikel yang terambil}}$$

(2) Pengujian kelengkapan (*Recall*) ialah perbandingan jumlah artikel yang ditetapkan sistem dengan jumlah seluruh artikel relevan yang ada dalam koleksi artikel (terambil ataupun takterambil sistem).

$$Recall: \frac{\text{Jumlah artikel yang relevan dengan query dan terambil}}{\text{Jumlah seluruh artikel relevan dan koleksi artikel}}$$

## 2.6 Langkah-langkah Penelitian

Langkah-langkah penelitian secara terstruktur dapat dilihat pada Gambar 2 berikut ini;



Gambar 2. Langkah - langkah penelitian

1. Identifikasi Masalah

Pada tahap ini adalah inialisasi dari penelitian, yaitu mencari permasalahan yang dihadapi dalam pengklasifikasian artikel berita melalui berbagai artikel berita politik. Hal ini dilakukan penelitian untuk dapat memahami kondisi permasalahan yang telah ada dan melakukan studi literatur.

2. Studi Literatur

Proses ini dilakukan dengan cara mereview dari berbagai penelitian – penelitian terdahulu yang membahas mengenai klasifikasi artikel berbahasa indonesia yang pernah dilakukan beberapa peneliti sebelumnya, oleh karena itu untuk membandingkan teks klasifikasi artikel dengan menggunakan algoritma *stemming* Nazief dan Andriani, dengan fungsi pembobotan fungsi kesamaan *Jaccard*, diharapkan dapat mengukur perbedaan dan memberikan hasil yang efisien dan valid dalam pengklasifikasian artikel berita politik.

3. Pengumpulan Data

Pengumpulan data menggunakan cara dengan mengambil artikel pada media digital yaitu *turnbackhoax.com*, dan beberapa media yang terdaftar *dewanpers.co.id* yang merupakan salah satu dari beberapa situs media digital yang memberikan artikel-artikel mengenai berita atau informasi yang dibutuhkan oleh pencari informasi. Data yang akan peneliti gunakan terdiri dari dua kategori atau label dari artikel berita politik. Pada kategori artikel politik akan di ambil sebanyak 200.3

4. Pembuatan Prototype

Penentuan artikel hoax ini adalah suatu sistem yang bertujuan mengklasifikasikan artikel teks berbahasa indonesia. Tahapan proses *text processing* di lakukang dengan pembobotan fungsi kesamaan *Jaccard* pada masing masing *word* pada setiap artiel teks, sehingga dapat menghasilkan *similarity* pada artikel

tersebut. Berdasarkan pembobotan fungsi kesamaan *Jaccard* dan metode *Vector Space Model* dengan Algoritma *Stemming* Nazief dan Andriani untuk penentuan artikel.

5. Pengujian Prototype

Penentuan artikel politik dapat diketahui pada penelitian ini telah berjalan dengan efektif dan memberikan *result output* yang diharapkan. Berdasarkan hasil pengujian masing-masing label maka dilakukan penghitungan untuk mencari akurasi dengan menggunakan fungsi kesamaan *Jaccard*. Adapun jumlah data yang akan diuji pada penelitian ini seperti pada Tabel 1 dibawah ini;

Tabel 1. Jumlah Data Artikel

Kategori Polittik (Hoax)	Kategori Politik (Non-Hoax)
100 Artikel Politik	100 Artikel Politik

6. Hasil dan Kesimpulan

Dari hasil pengujian manual mendapatkan hal-hal yang dapat mempengaruhi akurasi kebenaran tersebut, yaitu data diuji dari dua label dari media online *turnbackhoax.com*, sehingga belum tentu semua data yang diuji akan masuk kedalam label yang sesuai. Dan dari masing-masing data yang sudah di lakukan pengujian akan berpengaruh terhadap klasifikasi. Maka kesimpulan dapat peneliti ambil adalah semua data yang di ujikan telah masuk kedalam label atau kategori, namun ada artikel yang tidak sesuai dengan label atau kategori tersebut.

III. HASIL DAN PEMBAHASAN

Pengelompokan Data, Analisis

Dalam melakukan penelitian ini, artikel dokumen yang akan masuk ke dalam sistem dilakukan proses preprocessing untuk menjadikan *text* yang siap di proses, *text* ini yang berupa kata-kata atau *term* memasuki proses *stemming* menggunakan algoritma *stemming* dan melakukan pembobotan kata melalui *Vecor Space Model(VSM)* guna untuk mendapatkan akurasi dan performa model yang baik. Pada *dataset* yang akan digunakan pada penelitian ini terdiri dari satu kategori artikel *hoax* politik dengan dengan dua kelas *hoax* dan *non-hoax*, secara detail dapat dilihat pada tabel 2 dibawah ini:

Tabel 2. Kategori dataset pengujian

Kateogri	Class Hoax	Class Non-Hoax
Politik	100 Artikel	100 Artikel

Pada Tabel 2 dijelaskan bahwa artikel politik memiliki 100 Artikel berlabel *hoax* dan 100 Artikel berlabel *non-hoax*.

1.1.1 Proses Penelitian

Vector Space Model (VSM) digunakan pada penelitian ini, proses dari algoritma stemming dan proses pembobotan term dari kata-kata yang dihasilkan dari proses stemming dapat melakukan perhitungan. Hasil dari perhitungan jarak digunakan untuk mengetahui tingkat similarity dari dokument tersebut.

1.1.2 Pengelompokan Data

Pada penelitian ini menggunakan dataset sebagai pengujiannya yang di peroleh dari sosial media online yaitu beberapa media sosial yang terverifikasi oleh dewanpers dan turnbackhoax.id yang sebelumnya sudah diambil dengan jumlah artikel sebanyak 100 artikel hoax(berita salah) dan 100 artikel non-hoax(berita fakta).

1.1.3 Analisa Data Bahan Penelitian

Data hasil studi pustaka atau literatur pada penelitian penentuan artikel menggunakan similarity Jacard dan algoritma stemming adalah sebagai berikut:

- a. Stopword
 

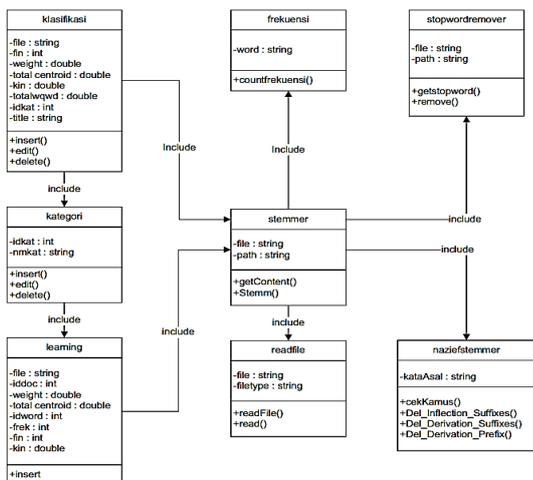
Stopword adalah kata-kata yang tidak proses dalam praprosesing. Karakteristik dari stopwords yaitu frekuensi kemunculannya tinggi, biasanya berupa kata penghubung atau kata ganti.
- b. File artikel yang akan dicek tingkat similarity
- c. Data kamus besar bahasa indonesia.
- d. Term untuk proses stemming

1.1.4 Perancangan Sistem

Proses perancangan sistem secara umum untuk pembangunan aplikasi terdiri atas beberapa tahapan, anara lain meliputi perancangan.

a. Perancangan Spesifikasi Program

Pada Gambar 3 merupakan Class Diagram yang digunakan sebagai model utama dalam memodelkan class diagram dengan konteks komponennya.

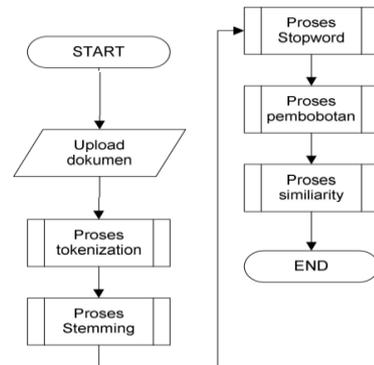


Gambar 3. Class Diagram

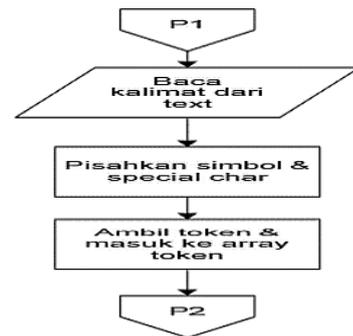
b. Perancangan Proses Sistem

Perancangan proses yang dimaksudkan adalah

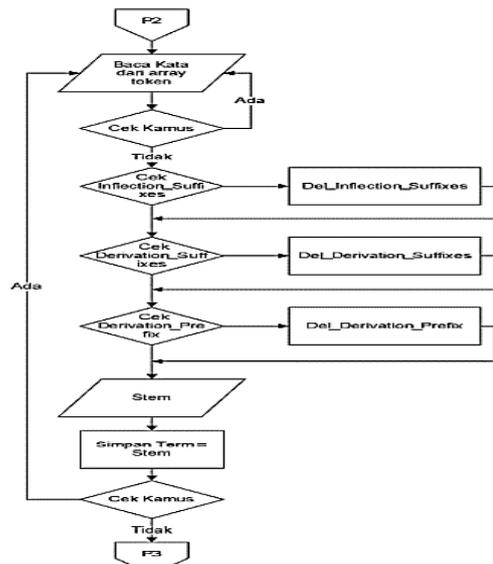
bagaimana sistem akan bekerja, proses-proses yang digunakan mulai dari user melakukan input hingga kemudian sistem menghasilkan output berupa nilai similarity menggunakan perhitungan Jaccard. Pada Gambar 3.1 – 3.6 menjelaskan proses utuh sistem klasifikasi artikel :



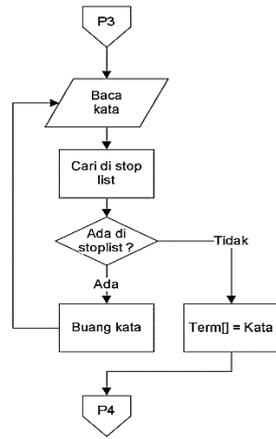
Gambar 3.1 Proses Klasifikasi Artikel



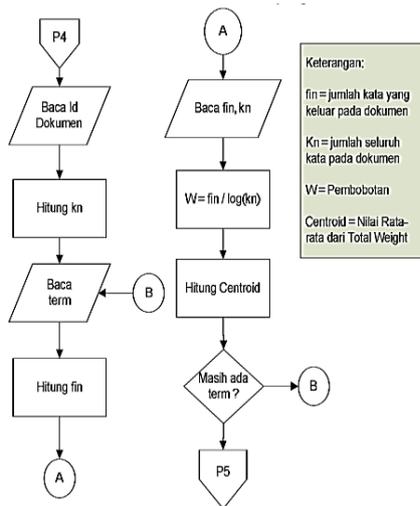
Gambar 3.2 Proses Tokenization



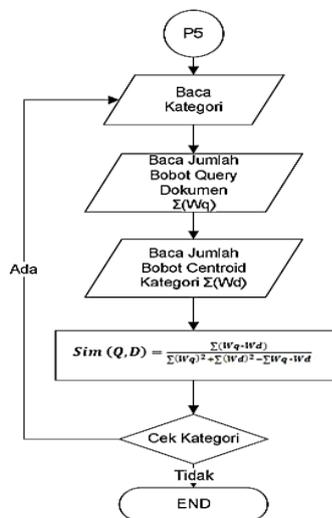
Gambar 3.3 Proses Stemming



Gambar 3.4 Proses Stopword



Gambar 3.5 Proses Pembobotan



Gambar 3.6 Proses Similarity

### 1.1.5 Pengujian Sistem

Pengujian atau testing adalah merupakan bagian yang tidak kalah penting, pengujian digunakan untuk memastikan hasil dari penelitian memenuhi tujuan yang dicapai atau tidak.

Metode pengukuran yang digunakan pada penelitian ini menggunakan metode pengujian *confusion matrix* yaitu pengujian *precision* dan *recall* dan *accuracy* dengan melakukan sepuluh kali *iteration*. Berikut formula yang digunakan pada pengukuran terdapat pada penjelasan Gambar 4 Formula Pengujian.

		Nilai sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (True Positive) <i>Correct result</i>	FP (False Positive) <i>Unexpected result</i>
	FALSE	FN (False Negative) <i>Missing result</i>	TN (True Negative) <i>Correct absence of result</i>

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Gambar 4. Formula Pengujian

Pembagian data dengan metode *holdout* yaitu membagi sejumlah data untuk *training* dan menggunakan sisanya untuk *testing*. Pembagian data menggunakan formula 70 % untuk *training* dan 30% untuk *testing*.

Pada penelitian ini pengujian dilakukan dengan menggunakan *iteration* maka metode pembagian data yang digunakan yaitu menggunakan metode *repeated holdout* pada setiap *iteration* dalam proses pengujian. Metode *repeated holdout* dilakukan dengan cara mengulangi proses dengan subsampel yang berbeda-beda pada setiap *iteration*.

Sebelum melakukan pengujian dengan *iteration* dan pembagian data dengan *repeated holdout*, penulis mempersiapkan artikel dengan jumlah keseluruhan sebanyak 200 artikel, yang dimana terdiri dari 100 dari artikel *hoax* dan 100 dari artikel *non-hoax*. Pada pengujian label *hoax* dan *non-hoax* disetiap *iteration* yang terdiri 70 artikel untuk data *training* dan 30 artikel untuk data *testing* dan artikel-artikel yang diujikan berbeda pada setiap *iteration* dengan menggunakan pembagian *repeated holdout*. Berikut penjelasan pengujian pada 1 *iteration*.

#### Iteration 1

Pada Tabel 3 merupakan hasil dari pengujian *Hoax*, sedangkan Tabel 4 Pengujian *Non-Hoax*. Tabel dibawah ini merupakan table yang menjelaskan hasil pengujian pada artikel *hoax* dan *non-hoax* pada *iteration* pertama.

Tabel 3. Pengujian Artikel Hoax Iteration 1

Hoax					
No	Artike	Nilai		Hasil	Relevan
		Non-Hoax	Hoax		
1	Hoax1	0	0.145401737	Hoax	Ya
2	Hoax5	0.020614924	0.025815217	Hoax	Ya
3	Hoax9	0.008072286	0.177811231	Hoax	Ya
4	Hoax11	0	0.022089106	Hoax	Ya
5	Hoax15	0.058237544	0.035014886	Berita	Tidak
6	Hoax19	0.038173715	0.010427494	Berita	Tidak
7	Hoax21	0.00680914	0.042578796	Hoax	Ya
8	Hoax25	0.022419514	0.077745247	Hoax	Ya
9	Hoax29	0.089517178	0.01533361	Berita	Tidak
10	Hoax31	0	0.070920884	Hoax	Ya
11	Hoax35	0	0.079298918	Hoax	Ya
12	Hoax39	0.045576875	0.038219314	Berita	Tidak
13	Hoax41	0	0.011807283	Hoax	Ya
14	Hoax45	0.022419514	0.077745247	Hoax	Ya
15	Hoax49	0.068781853	0.053666002	Berita	Tidak
16	Hoax51	0.020169371	0.055202333	Hoax	Ya
17	Hoax55	0.007321937	0.06345054	Hoax	Ya
18	Hoax59	0.071848676	0.028648903	Berita	Tidak
19	Hoax61	0.13242556	0.061363686	Berita	Tidak
20	Hoax65	0.070317104	0.071764524	Hoax	Ya
21	Hoax69	0.198717784	0.047998212	Berita	Tidak
22	Hoax71	0.267043346	0.107720875	Berita	Tidak
23	Hoax75	0.002486052	0.021229521	Hoax	Ya
24	Hoax79	0.112637232	0.069212639	Berita	Tidak
25	Hoax81	0.178229318	0.034113385	Berita	Tidak
26	Hoax85	0.147664127	0.030322195	Berita	Tidak
27	Hoax89	0.120671569	0.088872666	Berita	Tidak
28	Hoax91	0.010196978	0.08529875	Hoax	Ya
29	Hoax95	0.037204472	0.106888488	Hoax	Ya
30	Hoax99	0.01281125	0.055680547	Hoax	Ya
				Relevan	17
				NotRelevan	13

Hasil dari pengujian diatas, disimpulkan seperti dalam Tabel 4 dibawah ini;

Tabel 4. Accuracy Hasil Pengujian Artikel Hoax Iteration 1

	Nilai Prediksi	Nilai Sebenarnya	
		TRUE	FALSE
	TRUE	17	0
	FALSE	13	30
<b>Precision</b>	17/(17+0)	1.00	100%
<b>Recall</b>	17/(17+13)	0.57	57%
<b>accuracy</b>	(17+30)/(17+30+0+13)	0.78	78%

Hasil pengujian artikel hoax pada table diatas dapat dibandingkan dengan hasil pengujian pada artikel non hoax yang dapat dilihat pada Tabel 5 dibawah ini;

Tabel 5. Pengujian Artikel Non-Hoax Iteration 1

Non-Hoax					
No	Artike	Nilai		Hasil	Relevan
		Non-Hoax	Hoax		
1	Berita1	0.017881264	0	Berita	Ya
2	Berita5	0.028848609	0.016470699	Berita	Ya
3	Berita9	0.099095567	0.024866922	Berita	Ya
4	Berita11	0.11579326	0.02205727	Berita	Ya
5	Berita15	0.122298652	0.023546996	Berita	Ya
6	Berita19	0.089482137	0.021238729	Berita	Ya
7	Berita21	0.176092396	0.032629443	Berita	Ya
8	Berita25	0.138099485	0.024932012	Berita	Ya
9	Berita29	0.098140056	0.043523663	Berita	Ya
10	Berita31	0.054834924	0.028124428	Berita	Ya
11	Berita35	0.019937851	0	Berita	Ya
12	Berita39	0.158025603	0.062726837	Berita	Ya
13	Berita41	0.067437098	0.020422677	Berita	Ya
14	Berita45	0.066908758	0.037428533	Berita	Ya
15	Berita49	0.04333157	0.010620746	Berita	Ya
16	Berita51	0.038649309	0.004440996	Berita	Ya
17	Berita55	0.045423914	0.03994652	Berita	Ya
18	Berita59	0.086057367	0.017102993	Berita	Ya
19	Berita61	0.035391023	0.006487775	Berita	Ya
20	Berita65	0.121607117	0.026275558	Berita	Ya
21	Berita69	0.164812301	0.036134727	Berita	Ya
22	Berita71	0.053984576	0.017102306	Berita	Ya
23	Berita75	0.105728769	0.032538422	Berita	Ya
24	Berita79	0.130938431	0.053881498	Berita	Ya
25	Berita81	0.136057047	0.040870557	Berita	Ya
26	Berita85	0.065611811	0.023540399	Berita	Ya
27	Berita89	0.21356956	0.04042568	Berita	Ya
28	Berita91	0.115937979	0.022599502	Berita	Ya
29	Berita95	0.062889397	0.038507169	Berita	Ya
30	Berita99	0.041215826	0.016071787	Berita	Ya
				Relevan	30
				NotRelevan	0

Hasil dari pengujian Artikel Non Hoax diatas, disimpulkan seperti dalam Tabel 6 dibawah ini

Tabel 6. Accuracy Hasil Pengujian Artikel Hoax Iteration 1

	Nilai Prediksi	Nilai Sebenarnya	
		TRUE	FALSE
	TRUE	17	0
	FALSE	13	30
<b>Precision</b>	17/(17+0)	1.00	100%
<b>Recall</b>	17/(17+13)	0.57	57%
<b>accuracy</b>	(17+30)/(17+30+0+13)	0.78	78%

Jadi nilai rata-rata dari perhitungan diatas pada iteration 1:

Precision = 0.85 -> 85 %  
 Recall = 0.78 -> 78 %  
 Accuracy = 0.78 -> 78 %

Pengujian dilakukan selama 10 iterasi selama terus menerus. Hasil dari pengujian sebanyak 10 iterasi tidak ditampilkan. Dari 10 iterasi tersebut, nilai rata-rata dari perhitungan 10 iterasi seperti pada Tabel 7 dibawah ini;

Tabel 7. Nilai Rata-Rata Perhitungan Pengujian

	Precision	Recall	Accuracy
Iteration 1	0,85 -> 85	0,78 -> 78	0,78 -> 78
Iteration 2	0,85 -> 85 %	0,83 -> 83	0,83 -> 83
Iteration 3	0,82 -> 82	0,76 -> 76	0,76 -> 76
Iteration 4	0,87 -> 87	0,87 -> 87	0,87 -> 87
Iteration 5	0,88 -> 88	0,86 -> 86	0,86 -> 86
Iteration 6	0,82 -> 82	0,81 -> 81	0,81 -> 81
Iteration 7	0,81 -> 81	0,80 -> 80	0,80 -> 80
Iteration 8	0,83 -> 83	0,82 -> 82	0,82 -> 82
Iteration 9	0,81 -> 81	0,75 -> 75	0,75 -> 75
Iteration 10	0,85 -> 85	0,83 -> 83	0,83 -> 83

Dari hasil pengujian dengan 10 iterasi dapat diambil kesimpulan dari 2 (dua) sisi aspek, yaitu;

Aspek Sistem

Dari aspek sistem mengharapakan adanya sebuah server yang mendukung penyimpanan data besar untuk menyimpan berbagai data yang diperlukan untuk membangun sebuah sistem deteksi artikel.

Aspek Manajerial

Module sistem yang dihasilkan pada penelitian ini diharapkan dapat diterapkan pada berbagai jenis sistem untuk deteksi artikel lainnya pada sebuah institusi.

Aspek Penelitian Lanjut

Para peneliti selanjutnya dapat menggunakan hasil penelitian ini untuk bahan referensi pada penelitian sejenisnya dan dapat menjadi model awal yang dapat dikembangkan dan menghasilkan penelitian lanjut yang lebih baik.

IV. KESIMPULAN

Sesuai dengan pembahasan, studi pustaka, tinjauan penelitian dan metode penelitian dalam pengklasifikasian artikel *hoax* pada sosial media dengan *similarity Jacard* dan algoritma *stemming*, maka kesimpulan yang dapat di ambil yaitu sebagai berikut: (1) Dapat memudahkan pengklasifikasian artikel yang kurang tepat, yang bisa menimbulkan *miss information* bagi para pencari informasi. (2) Dengan metode *Vector Space Model* dengan *stemming* dan *similarity Jaccard* dapat memudahkan dalam pengklasifikasian artikel dalam jumlah banyak dan menghasilkan nilai *precision* sebesar 0.92 *recall* sebesar 0,80 dan akurasi sebesar 87 %

Saran untuk kedepan, diuraikan sebagai berikut; (1) Pada penelitian selanjutnya diharapkan dalam menggunakan data training lebih diperbanyak, sehingga nilai kemiripan lebih tepat, semakin banyak data *training* yang dipelajari maka sistem akan mempelajari kata-kata yang biasanya terdapat pada dokumen tersebut. (2) Pada instansi-instansi terkait, yang berkaitan dalam penyebaran informasi artikel, aplikasi pengklasifikasian ini juga dapat digunakan untuk artikel-artikel yang lainnya seperti artikel kesehatan, artikel bencana alam dan sebagainya.

REFERENCES

[1] Maxmanroe. 2017. Pengertian Media Sosial Secara Umum, Fungsi, Tujuan, Jenis Sosmed. *Maxmanroe*. [Online] 2017. [Dikutip: 21 Juli 2019.] <https://www.maxmanroe.com/vid/teknologi/internet/pengertian-media-sosial.html#>.

[2] Afriza, A. (2019) ‘Metode Klasifikasi Rocchio untuk Analisis Hoax Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis’, *Metode Klasifikasi Rocchio untuk Analisis Hoax Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis*, (February), pp. 0–10. doi: 10.29244/jika.5.1.1-10.

[3] TAMMAM, A. G. (2018) ‘Deteksi Hoaks Pada Media Sosial Berbasis Text Mining Classification System’, *Deteksi Hoaks Pada Media Sosial Berbasis Text Mining Classification System*.

[4] Weddiningrum, F. G. et al. (2018) ‘Deteksi Konten Hoax Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance’, *Deteksi Konten Hoax Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance*.

[5] Roger S. Pressman, Ph.D. 2012. *Rekayasa Perangkat Lunak*. s.l. : Andi , 2012. 9789792931044.

[6] Annisa, Nurul, D. (2017) ‘Implementasi Algoritma Vector Space Model Dalam Pencarian E-Book’, (October 2014).

[7] E. Garcia, P. (2006) ‘The Classic TF-IDF Vector Space Model’, *The Classic TF-IDF Vector Space Model*, (1), pp. 1–15.

[8] Irmawati (2017) ‘Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model’, *Jurnal Ilmiah FIFO*, IX(1), pp. 74–80. doi: 10.1093/clinids/13.Supplement\_1.S109.

[9] Lutfi and Musodo, K. A. (2013) ‘Trigram Pada Vector Space Model Dengan Pembobotan Dalam Dokumen Dan Fungsi Kesetaraan Cosine’, *Trigram Pada Vector Space Model Dengan Pembobotan Dalam Dokumen Dan Fungsi Kesetaraan Cosine*.

[10] Maulina, D. et al. (2018) ‘Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document’, *Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document*, 2(1), pp. 35–40.

[11] Prasetyo, A. R. and Adikara, P. P. (2018) ‘Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor’, *Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor*, 2(12), pp. 7466–7473.

[12] Immersa.lab. 2018. Pengertian Artificial Intelligence. *immersa.lab*. [Online] February 15, 2018. [Cited: Juli 23, 2019.] <http://www.immersa-lab.com/pengertian-artificial-intelligence.htm>.

[13] Nasrullah, Rulli. 2015. *Media Sosial; Perspektif Komunikasi, Budaya, dan Sosioteknologi*. Bandung : Simbiosis Rekatama Media, 2015.

[14] Purbolaksono, Arfianto. 2017. Anti Hoax. *theindonesianinstitute*. [Online] Januari 01, 2017. [Cited: Juli 21, 2019.] <https://www.theindonesianinstitute.com/anti-hoax/>.