**Skema Pendanaan : Peneliti Madya**

# PROPOSAL PENELITIAN AKSELARASI



DETEKSI BERITA HOAX PADA SOSIAL MEDIA DENGAN METODE N-GRAM DAN ALGORITMA WINNOWING

# Peneliti :

|  |  |  |  |
| --- | --- | --- | --- |
| **Ketua Peneliti** | **:** | **Dr. Ir. Goenawan Brotosaputro, S.Kom., M.Sc.** | **930008** |
| **Anggota** | **:** | **Wiwin Windihastuty, S.Kom., M.Kom** | **120077** |

**FAKULTAS TEKNOLOGI INFORMASI UNIVERSITAS BUDI LUHUR**

# JUNI 2023

**HALAMAN PENGESAHAN**

|  |  |  |
| --- | --- | --- |
| Judul | : | Deteksi Berita Hoax Pada Sosial Media Dengan Metode N-Gram dan Algoritma Winnowing. |
| Bidang Kegiatan | : | Ilmu Komputer |
| Ketua Pelaksana |  |  |
| a. Nama Lengkap | : | Dr. Ir. Goenawan Brotosaputro, S.Kom., M.Sc. |
| b. NIP/NIDN/SINTA | : | 930008 / 0324226901 / 260007 |
| c. Jabatan Fungsional | : | Lektor |
| d. Program Studi | : | Sistem Informasi |
| e. Nomor Handphone | : | 081315134242 |
| f. Alamat Email | : | goenawan.brotosaputro@budiluhur.ac.id |
| Anggota |  |  |
| a. Nama Lengkap | : | Wiwin Windihastuty, S.Kom., M.Kom |
| NIP/NIDN/SINTA | : | 120077/ 0326047001/6042810 |
| Jabatan Fungsional | : | Lektor |
| Nomor Telepon | : | 081310767472 |
| Alamat EmailMahasiswa1. Nama Nim
 | ::: | wiwin.windihastuty@budiluhur.ac.idRyo Umar Wahid 191150066 |
| Biaya Pelaksanaan Akselarasi | : | Rp 10.000.000,- |

|  |  |
| --- | --- |
| Mengetahui,Dekan Fakultas Teknologi InformasiDr. Ir. Deni Mahdiana, S.Kom., M.M., M.KomNIP. 960012 | Jakarta, 8 Juni 2023 Ketua PenelitiDr. Goenawan Brotosaputro, S.Kom.,M.Sc,NIP. 930008 |
| Menyetujui,Direktur Riset dan Pengabdian Kepada MasyarakatKrisna Adiyarta M, Ph.D. NIP. 890001 |

Tuliskan judul usulan penelitian

JUDUL USULAN

[ Deteksi Berita Hoax Pada Sosial Media Dengan Metode N-Gram dan Algoritma Winnowing

]

Ringkasan penelitian tidak lebih dari 500 kata yang berisi latar belakang penelitian, tujuan, tahapan metode penelitian, dan target luaran yang ingin dicapai.

RINGKASAN

[ Perkembangan teknologi dan informasi yang semakin pesat selain membawa dampak positif, juga membawa dampak negatif. Salah satu dampak negatif yang cukup meresahkan adalah munculnya informasi palsu atau lebih popular dikenal dengan istilah “hoax”. Penyebaran informasi hoax menimbulkan keresahan masyarakat. Media sosial sering kali dimanfaatkan sebagai tempat untuk menyebarkan berita hoax. Mudahnya penyebaran informasi melalui media sosial menjadikan penyebaran hoax semakin merajalela di dunia maya sehingga dapat menimbulkan beragam opini masyarakat. Penyebaran berita hoax juga mampu membawa pada kerancuan informasi dan kehebohan publik akan suatu informasi, bahkan dapat juga berakibat pada perpecahan suatu bangsa. Berita-berita yang tersebarkan terutama berita mengenai politik sering kali mengakibatkan salah penafsiran karena berita tersebut belum tentu benar sehingga dibutuhkan pengklasifikasian apakah termasuk dalam kategori berita hoax atau berita non hoax. Melihat beberapa permasalahan yang terjadi, peneliti bermaksud untuk membuat model deteksi yang dapat menentukan apakah suatu berita termasuk hoax atau non hoax dengan tujuan untuk memudahkan dalam menentukan atau mengkategorikan berita secara tepat yang dapat meyakinkan pembaca mengenai kebenaran subuah berita sehingga pembaca mengambil tindakan sesuai dengan isi berita Penelitian kali ini mengambil tema proses klasifikasi berita hoax menggunakan tahap preprossessing yang terdiri dari tokenization dan stemming. Dilanjutkan dengan proses pembobotan kata dengan Metode N-Gram dan Jaccard Similarity terhadap algoritma winnowing hingga proses klarifikasi dengan menggunakan metode Vector Space Model (VSM). Luaran penelitian akan dilaporkan dalam bentuk laporan kegiatan dan akan dipublikasikan ke jurnal nasional MATRIK terakreditasi SINTA 3.

]

Kata kunci maksimal 5 kata

[Hoax; Preproccessing; Vector Space Model; Tokenization; N-Gram]

Pendahuluan berisi latar belakang pemilihan topik penelitian, permasalahan yang akan diteliti, lingkup yang menjadi batasan penelitian, tujuan khusus, dan urgensi penelitian.

PENDAHULUAN [

Hoax atau penyebaran berita bohong atau juga biasa disebut fitnah sering terjadi di tahun politik seperti saat ini. Berita hoax semakin menunjukkan pengaruh dan efek negatif bagi persatuan dan kesatuan bangsa. Berita bohong atau fitnah yang telah menyebar, dimanfaatkan untuk kepentingan politik dan ekonomi dari pihak yang menghendaki kerusakan dalam bermasyarakat. Beredarnya hoax atau berita bohong telah menjadi konsumsi masyarakat sehari-hari. Hoax sering kali dianggap sebagai informasi atau berita yang benar. Masyarakat yang awam akan sumber sumber berita yang benar, sulit membedakan informasi atau berita yang diperolehnya.

Perkembangan teknologi informasi ikut serta mendorong penyebaran berita hoax dengan tujuan membuat opini publik menjadi beragam selain itu, juga untuk bersenang-senang yang menguji kecerdasan serta kecermatan pengguna internet di media sosial (Anwar, 2017).

Hoax di Indonesia semakin marak terjadi sejak pemilihan presiden 2014 dan dampaknya terlihat dengan gencarnya kampanye di media sosial. Hoax bermunculan serta saling menjatuhkan lawan politik seperti kampanye hitam. Berdasarkan laporan tahunan, Kominfo menerima konten hoax sebanyak 733 laporan di tahun 2018 pada aplikasi pesan instan. Sejak Agustus 2018 hingga 21 Januari 2019, Kominfo mendapat laporan hoax yang dikirimkan melalui WhatsApp sebanyak 43 konten. Dari hasil pemantauan Direktorat Pengendalian Aplikasi Informatika Direktorat Jenderal Aplikasi Informatika, laporan yang sering terjadi pada bulan Oktober 2018. Sebanyak 16 konten hoax disebarkan melalui platform WhatsApp. Pada bulan Agustus 2018 terdapat laporan dua konten hoax, Pada Bulan September 2018 ada lima konten hoax, sedangkan November 2018 sebanyak 8 laporan konten dan di bulan Desember 2018 sebanyak 10 laporan konten hoax. Pada tanggal 21 Januari 2019, laporan yang diterima dua laporan isi dari aplikasi milik facebook.

Kepala Biro Humas Kementrian Kominfo dalam penjelasannya, sebagai pengelola pengaduan konten negatif yang diserbarkan melalui aplikasi media online. Pada tahun 2016 terdapat 14 aduan konten dan di tahun selanjutnya jumlah aduan meningkat menjadi 281 aduan, sementara pada tahun 2018 terdapat laporan sebanyak 1440 kasus yang berkaitan dengan konten negatif.



# Gambar 1 Temuan Isu Hoac 2018 - 2019

Melihat beberapa permasalahan yang terjadi, peneliti bermaksud untuk membuat model deteksi yang dapat menentukan apakah suatu berita termasuk *hoax* atau *non hoax* dengan tujuan untuk memudahkan dalam menentukan atau mengkategorikan berita secara tepat yang dapat meyakinkan pembaca mengenai kebenaran subuah berita sehingga pembaca mengambil tindakan sesuai dengan isi berita

Berdasarkan uraian dari latar belakang diatas, dapat disimpulkan beberapa permasalahannya seperti tersebarnya berita politik, yang menyulitkan dalam menentukan kebenaran berita tersebut.

Berdasarkan identifikasi masalah, maka untuk mencapai tujuan penelitiandiatas, penelitian memfokuskan batasan penelitian ini yaitu :

1. Hanya berita politik yang menggunakan Bahasa Indonesia atau kaidah Bahasa Indonesia.
2. Pengujian data didapatkan dari website <https://turnbackhoax.id/>

Berdasarkan uraian dari latar belakang diatas, maka rumusan dari permasalahan adalah sebagai berikut :

1. Bagaimana memudahkan dalam menentukan berita yang tepat, pada penentuan label hoax dan non-hoax ?
2. Apakah dengan penggunaan metode *N-Gram* dengan *Winnowing* dan *Similarity Jaccard*

dapat memudahkan dalam pengklasifikasian berita dalam jumlah yang banyak ? Tujuan dari penelitian ini yaitu membangun aplikasi dimana dapat mengetahui informasi-

informasi berita politik dan sistem akan melakukan penentuan berita secara sistematis kedalam sebuah label atau kategori *hoax* atau *non hoax* yang telah ditentukan oleh sistem.

Manfaat dari penelitan ini untuk memberikan keakuratan dalam menentukan berita *hoax* atau *non hoax* pada media sosial serta dapat mempermudah pemberian label atau pengkategorian *hoax* dan *non hoax* pada artikel-artikel berita politik.

]

Tinjauan pustaka menguraikan secara jelas kajian pustaka yang melandasi timbulnya gagasan dan permasalahan yang akan diteliti. Sumber pustaka mengutamakan hasil penelitian pada jurnal ilmiah yang terkini (maksimal 10 tahun terakhir).

TINJAUAN PUSTAKA [

# Klasifikasi

Klasifikasi adalah suatu proses untuk menemukan model atau fungsi yang dapat menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk bisa memperkirakan kelas dari suatu obyek (HENDRIAN, 2018). Penentuan kelas dari suatu dokumen dilakukan dengan cara membandingkan nilai probabilitassuatu sample berbeda dikelas yang satu dengan nilai probalitas suatu sampel berbeda di kelas yang lain (HENDRIAN, 2018).

Klasifikasi berasal dari kata latin *classis.* Klasifikasi adalah proses pengelompokan, artinya mengumpulkan benda atau entitas yang sama serta memisahkan benda atau entitas yang tidak sama.

Klasifikasi adalah pengelompokan yang sistematis dari jumlah obyek, gagasan, buku atau benda-benda lain kedalam kelas atau golongan tertentu berdasarkan ciri-ciri yang sama.

# Vector Space Model

*Vector Space Model* atau *Term Vector Space Model* adalah model aljabar untuk menggambarkan dokumen teks sebagai *vector identifier*. Biasanya digunakan untuk penyaringan informasi*,* penemuan informasi, *indexing* dan pemberian *ranking* yang saling relevan (Eva Y Puspaningrum, 2018).

Proses dari perhitungan metode ini adalah *indexing* dokumen, pembobotan *term* dan perhitungan sebuah kesamaan. Proses *indexing* dokumen adalah proses melalui tahapan-

tahapan dalam *text mining*. Proses selanjutnya adalah pembobotan *term* dengan algoritma TF-IDF. Proses yang terakhir yaitu perhitungan kesamaan dengan pendekatan *Jaccard* yang dinyatakan pada rumus.

# Media Sosial

Van Dijk dalam menyatakan bahwa media sosial adalah *platform* media digital yang memfokuskan pada eksistentsi pengguna yang memfasilitasi dalam beraktifitas maupun berkolaborasi, karena itu media sosial dapat dilihatsebagai media *online* yang menguatkan hubungan antara pengguna sekaligus sebagai sebuah ikatan sosial.

Meike dan Young dalam mengartikan kata media sosial sebagai konvergensi antara komunikasi personal dalam arti saling berbagi antara individu *(to be share one-to-one)* dan media publik untuk berbagi kepada siapa saja tanpa kecuali.

Menurut Boyd dalam media sosial sebagai kumpulan perangkat lunak dimana individu maupun komunitas dapat berkumpul,berbagi, berkomunikasi, dan berkolaborasi atau bermain.Media sosial memiliki kekuatan pada *user-generated content* (UGC) dimana konten dihasilkan oleh pengguna, bukan oleh editor sebagaimana di instansi media massa Pada intinya melalui sosial media dapat melakukan berbagai aktifitas dua arah dalam berbagai bentuk pertukaran, kolaborasi, dan saling berkenalan dalam bentuk tulisan, visual maupun audiovisual. Sosial media diawali dari tiga hal, yaitu *Sharing,*

*Collaborating* dan *Connecting* (Tongkotow Liedfray, 2022)*.*

# Information Retrieval

Definisi umum untuk *information Retrieval* pada dasarnya merupakan proses untuk menentukan artikel dalam koleksi yang harus ditemubalikan untuk memenuhi keperluan dari si pengguna akan suatu informasi.

*Information Retrieval* merupakan bagian dari *Computersience* yang berhubungan dengan pengembalian informasi dari artikel-artikel yang didasarkan pada isi dan konteks dari artikel-artikel itu sendiri (Ahmad Fauzi, 2018). *Information Retrieval* (perolehan informasi) merupakan suatu pencarian yang didasarkan pada suatu *query* (inputan) yang diharapkan dapat memenuhi kebutuhan dari kumpulan dokumen yang ada. Sedangkan definisi *query* dalam *Information Retrieval* merupakan sebuah formula yang digunakan untuk mencari informasi yang dibutuhkan oleh *user*, dalam bentuk yang paling sederhana (Ginabila, 2020), sebuah *query* merupakan suatu *keywords* (kata kunci) dan artikel yang menggandung *keywords* merupakan artikel yang dicari dalam *Information Retrieval*

*System*.

Pada Proses *Information Retreval*, seperti pada Gambar 2. menjelaskan *Information Retrieval System* itu dapat digambarkan sebagai sebuah proses untuk mendapatkan *relevant* dokument dari *collection document* yang ada melalui *query* yang diinput oleh *user* (Pratama, 2018).



# Gambar 2. Proses Informasi Retrieval

Proses yang terjadi didalam *information retrieval system* terdiri dari dua bagian utama, yaitu *indexing subsystem* dan *searching subsystem* (*matching system*). Proses *indexing* dilakukan untuk membentuk basis data terhadap koleksi dokumen yang dimasukan atau dengan kata lain, *indexing* merupakan proses persiapan yang dilakukan terhadap dokumen sehingga dokumen siap untuk diproses. Proses *indexing* sendiri terdiri dari dua proses, yaitu dokumen*indexing* dan *term indexing*. Dari *term indexing* akan dihasilkan koleksi kata yang akan digunakan untuk mengingkatkan performasi pencarian pada tahap selanjutnya (Ginabila, 2020).

*Search subsystem* (*matching*) merupakan proses perolehan informasi yangrelevan terhadap *query* yang diberikan. Tidak semua yang diambil (*retrieved*) oleh *system* merupakan dokumen yang sesuai dengan kebutuhan(*relevant*).

Adapun fungsi utama sistem perolah informasi seperti dikemukakan oleh adalah sebagai berikut :

* 1. Mengidentifikasi sumber informasi yang relevan dengan minat masyarakat pengguna yang ditargetkan.
	2. Menganalisa isi sumber informasi.
	3. Mempresentasikan pertanyaan (*query*) pengguna dengan cara tertentuyang memungkinkan untuk dipertemukan sumber informasi yang terdapatdalam basis data.
	4. Mempertemukan pernyataan pencarian dengan data yang tersimpan dalam basis data.
	5. Memperoleh informasi yang *relevant*.

# Winnowing

Algoritma winnowing merupakan salah satu algoritma yang bertujuan untuk membuat dokumen fingerprinting. Algoritma ini merupakan pengembangan dari Rabin-Karp dengan penambahan metode window. Fingerprint ini dihasilkan setelah melalui rangkaian proses sanitize, pembentukan n-gram, pembentukan hash dari n-gram, pembentukan window dan selanjutnya menghasilkan fingerprint (Eva Y Puspaningrum, 2018). Teknik hash yang digunakan ada bermacam-macam. Dalam penelitian ini menggunakan teknik rolling hash untuk membuat hash dari rangkaian n-gram. Penjelasan proses lengkapnya adalah sebagai berikut.

1. Sanitize, yaitu pembersihan/penghapusan teks dari karakter yang bukan merupakan huruf dan angka seperti tanda baca dan karakter simbol. Kemudian proses case-folding untuk mengubah huruf kapital menjadi huruf kecil. Sebagai contoh, sebuah teks “Info: Penggunaan masker bermanfaat untuk mencegah penularan virus berbahaya” akan menjadi ”info penggunaan masker bermanfaat untuk mencegah penularan virus berbahaya”.
2. Pembentukan rangkaian n-gram yang merupakan pengelompokan kata yang berdekatan dengan n panjang tiap gramnya. Rangkaian yang digunakan dalam penelitian ini adalah rangkain kata sebanyak n. Pada contoh ini

diberikan perumpamaan nilai n-gram = 5. Dari teksi diatas akan menjadi rangkaian : [info penggunaan masker bermanfaat untuk] [penggunaan masker bermanfaat untuk mencegah] [masker bermanfaat untuk mencegah penularan] [bermanfaat untuk mencegah penularan virus] [untuk mencegah penularan virus berbahaya].

1. Penghitungan Hash dari setiap gramnya. Hashing adalah pengubahan teks menjadi kunci yang panjangnya tetap sama meskipun teks yang dihasing memiliki panjang yang berbeda, kunci ini mewakili teks aslinya. Proses pengubahan menjadi nilai hash menggunakan fungsi rolling hash. 5
2. Pembagian rangkaian hash kedalam window dengan panjang n. Teknik ini mirip dengan n-gram, namun objeknya adalah rangkain hash yang menjadi beberapa window dengan panjang tiap window sejumlah n.
3. Pemilihan beberapa nilai hash untuk membentuk fingerprint dokumen. Proses ini adalah pemilihan nilai paling kecil di tiap window. Jika ada dua nilai terkecil dalam window maka dipilih nilai paling kanan (urutan terbakhir). Hasil pemilihan tersebut disimpan dalam rangkaian sehingga menghasilkan finger print dokumen.

# Metode N\_Gram

Model dokumen N-gram memungkinkan kesamaan terjadi diukur berdasarkan urutan kata-kata terdekat yang tumpang tindih (frasa) daripada katakata individual. Ngram menangkap beberapa bentuk kesamaan sintaksis dan kontek kalimat antara dokumen dan menghindari kekurangan asumsi independensi kata yang membatasi model bag of word (Johnson & Zhang, 2015). N-gram pada dasarnya digunakan dalam penelitian ini untuk membedakan dan mengkategorikan dokumen berdasarkan ukuran n-gram yang serupa. Misalnya mayoritas dokumen yang sangat mirip dapat dideteksi menggunakan jumlah urutan n-gram lebih tinggi (n-gram lebih panjang) dari dokumen yang ditinjau dengan ringan. Karenanya menggunakan n-gram panjang (ukuran) tertentu bisa membantu membedakan satu kelas dokumen yang serupa dari yang lain. Namun ukuran model n-gram harus dipilih dengan cermat untuk menghindari memintas dokumen berpotensi menjiplak atau

mendeteksi dokumen dari kategori terdekat menghasilkan false positif dan penurunan kinerja. Untuk mendapatkan n-gram terbaik untuk kategori tertentu perlu menguji n-gram dengan panjang berbeda secara berurutan (Thompson, Panchev, & Oakes, 2015).

dikembangkan oleh Bobby Nazief dan Mirna Adriani pada tahun 1996 sebagai hasil penelitian internal Universitas Indonesia. Dengan menggabungkan metode pemotongan imbuhan serta pemcarian kamus yang terdiri atas kata dasar, algoritma ini menghasilkan tingkat akurasi *stemming* pada teks bahasa Indonesia.

Proses *winnowing* ini dilakukan dengan cara memotong imbuhan yang dilakukan secara rekursif (Eva Y Puspaningrum, 2018). Serta mencari data didalam kamus yang dilakukan sebelumtahap pemotongan. Kelemahan dalam algoritma ini diantaranya tingkat akurasinya tergantung dari tingkat kamus yang dimiliki.

Adapun algoritma *winnowing* bahasa Indonesia yang dikembangkan ini memilikitahapan sebagai berikut :

1. Cari kata yang akan di*stem* dalam kamus. Jika ditemukan maka kata tersebut di anggap *root word.*
2. Hapus *Inflection Suffixes* (“-lah”,”-kah”,”-ku”,”-mu” atau “-nya”). Jika imbuhan yang dihapus berupa particles (“-lah”,”-kah” dan “-tah”) maka langkah ini diulangi lagi untuk menghapus Prossesive Pronouns (“-ku”,”- mu” atau “-nya”). Contohnya, kata “mobilnyakaha” akan dipotong menjadi “mobilnya” untuk menghilangkan *particle* “-kah”, lalu dipotong kembali “mobil” yang merupakan kata dasar yang valid.
3. Hapus *Derivation Suffixes* (“-i”,”-an” atau “-kan”). Jika telah menjadi penghapusan imbuhan “-an” dan huruf terakhir dari kata tersebut adalah “-k” maka huruf “-k” juga ikut dihapus. Akhirnya *Derivation Suffixes* yang telah dihapus dikembalikan dan lanjut ke langkah 4. Misalnya terdapat kata “Merusakkan” akan dipotong menjadi “merusak”, karena kata tersebut belum merupakan kata dasar dan akan memasuki langkah berikutnya.
4. Hapus *Derivation Prefix*, ada beberapa hal yang menyebabkan algoritma

berhenti, jika *prefix* kata membentuk kombinasi awalan-akhiran yang tidak diijinkan, jika imbuhan awal yang ditemukan identik dengan imbuhan awal yang sebelumnya atau telah dilakukan pemotongan imbuhan tiga kali.Ada pun kombinasi awalan-akhiran yang tidak diizinkan. Imbuhan awalanterdiri dari atas dua jenis : Awalannya sederhana : “di-”,”ke-” atau “se-”yang dapat dihapus dengan langsung. Awalan kompleks : “te-” , “me-” , “be-” atau “pe- ” membutuhkan sebuah proses tambahan untuk menentukan tipe awalannya. Misalnya untuk kata “memperluas”, pertamakita hapus imbuhan “mem-” menjadi “perluas” kemudian diulangi lagi menjadi “luas”.

1. Untuk beberapa imbuhan *Derivation Prefix*, terdapat huruf-huruf tertentu yang mengawali kata tersebut. Karena itu dilakaukan proses *Recording*. Misalnya terdapat kata “penyendiri” pemotongan imbuhan “peny-” akan menyisakan “endiri” yang merupakan bukan kata dasar. Karena itu bedasarkan rule huruf “-s” menjadi “sendiri” dan merupakan menjadi kata dasar yang valid.
2. Jika semua langkah telah diproses dan selesai, tetapi tidak juga berhasil maka kata awal diasumsikan sebagai *root word* dan kata tersebut dikembalikan semua.

# Weight, Similarity, Centroid Pencarian Pembobotan (*Weight*)

Dalam menghitung bobot *term*, sekalipun *term frequency* banyak digunakan, namun ia hanya mendukung proporsi jumlah *Information Retrieval*, sedangkan proporsi jumlah artikel yang di temukan dandianggap relevan untuk kebutuhan pengguna akan lebih meningkat bila vektor bobot tersebut menggunakan *term* yang jarang muncul pada koleksi artikel (I Gede Iwan Supida, 2019). Tentunya *term* demikian diharapkan mampu mengelompokkan sejumlah artikelyang memuatnya, sehingga berbeda dengan seluruh anggota koleksi lainnya yang tidak memilikinya. Kriteria ini dapat di akomodasi dengan menghitung *inverse* frekuensi dokumen. Dengan digabungkan kedua metode ini yaitu konsep frekuensi kemunculan term dalam sebuah dokumen dan *inverse* frekuensi dokumen (*tf-idf*) yang mengandung kata tersebut, akan meningkatkan

proporsi jumlah dokumen yang dapat ditemukan kembali dan dianggap relevan secara sekaligus. Sehingga kriteria *term* yang paling tepat adalah *term* yang sering muncul dalam artikel secara individu namun jarang dijumpai pada artikel lainnya.

Salah satu untuk memberi bobot terhadap suatu kata adalah memberikan nilai jumlah kemunculan suatu kata ( *term frequency* ) sebagai bobot. Semakin besar kemunculan suatu kata dalam artikel akan memberikan nilai kesesuaian yang semakin besar.

Faktor lainnya yang diperhatikan dalam pemberian bobot kejarang munculan kata ( *termscarcity*) dalam koleksi. Kata yang muncul pada sedikit artikel harus di pandang sebagai kata yang lebih penting (*uncommonterm*) dari pada kata yang muncul pada banyak artikel (Annisa, 2019).

Faktor terakhirnya adalah faktor normalisasi terhadap panjang artikel. Artikel dalam koleksi memiliki karakteristik panjang yang beragam. Ketimpangan terjadi karena artikel yang panjang akan cenderung mempunyai frekuensi kemunculan kata yang besar. Sehingga untuk mengurangi ketimpangan tersebut diperlukan faktor normalisasi dalam pembobotan.

# Pencarian Kesamaan (Similarity)

Formula *jaccard* dimana nilai bobot seluruh *term* yang telah diperoleh akan dikalkulasikan dengan nilai bobot seluruh *query* yang juga telah di hitung sebelumnya (Indri Tri Hapsari, 2017).

# Pengklasifikasian Dokumen Berbasis *Centroid*

Dalam naskah ini kami mempersembahkan klasifikasi dokumen berbasis *centroid*. Didalam algoritma ini *vector centroid* dihitung untuk mengawali masing-masing nilai tolak ukur dokumen agar bisa diklasifikasikan. Didalam klasifikasi berbasis *centroid* ini, dokumen mewakili pengguna *vector space model* (Mardi Siswo utomo, 2022). Di model ini, masing-masing dokumen diwakili oleh frequensi *term* I pada dokumen (fin), dan di bandingkan dengan hasil jumlah *term* pada dokumen (kn).

Untuk lebih mudah memahami apa yang sedang dijelaskan penulis

membuat tabel beserta contoh perhitungannya terlihat pada persamaan (2.3).

# Stopword

Bahasa resmi di beberapa Negara memiliki kata sambung dan kata fungsi seperti artikel. Kata-kata ini tidaklah mempunyai arti yang penting di dalam proses pencarian informasi. Kata-kata tersebut misalnya “yang”, “di”, “pada” pada bahasa indonesia, di sebut sebagai *stopwords.* Sebuah sistem informasi retrieval selalu dipasangkan dengan sebuah *stoplist*. *Stoplist* merukapan sekumpulan kata-kata yangtidak relevan, namu sering sekali muncul. Dengan kata lain *stoplist* yaitu terdiri dari sekumpulan *stopwords*. *Stopwords removal* adalah sebuah proses dimana untukmenghilangkan kata yang tidak relevan pada hasil parsing dalam sebuah dokumen teks dengan cara membandingkannya dengan *stoplists* yang ada.

# Pencocokan Pemprofilan TextKonsep Pencocokan Profil

The *Classic Vector* Space *Model* : Ruang *vector* menghitung kemiripan dengan mendifinisikan sebuah *vector* yang mempresentasikan setiap dokumen dan sebuah *vector* yang mempresentasikan *query* (Rina Anggeriani, 2017). Jika kata-kata dalam artikel dapat dipresentasikan dalam *vector*, sangat dimungkinkan untuk membandingkan artikel dengan *query* untuk menunjukan kemiripan konten, keduanya. Jika sebuah sebuah *query* di anggap mirip dengan sebuah artikel, koefisienkemiripan yang mengukur kemiripan antara artikel dan *query* dapat dihitung.Artikel yang kontennya paling mirip dengan konten pada *query* di anggap paling relevan.

Faktor pembobotan untuk setiap kata dalam artikel didefinisikan sebagai kombinasi ( tf-idf) *term frequency* dan *invers document frequency*. Untuk menghitung nilai bobot kata.

# Fungsi Kesetaraan

Kualitas dari sebuah penerapan penentuan artikel dapat dilihat dari tingkat relevansi data yang didapatkan. Kebanyakan menggunakan berbagai metode untuk menentukan peringkat hasil agar mampu memberikan hasil yang maksimal.

Untuk mengetahui *similarity* atau tingkat kesetaraan antara artikel yang di*input* dengan artikel yang telah ditetapkan kategorinya. Jika *similarity* artikel yang di*input* mendeteksi *similarity* artikel yang telah ditetapkan label atau kategorinya maka artikel tersebut adalah artikel yang cocok untuk masuk kedalam salah satu kategori paling mendekati *similarity*dari dokumen yang telah ditetapkan sebelumnya (Glen Hizkia Oge Mangundap, 2022). Fungsi yang digunakan pada penelitian ini adalah fungsi *JACCARD* dengan menggunakan persamaan (2.5) dan juga fungsi ini didapat dari hasil penelitian dengan judul yaitu “*Measuring the degree of similarity between object in text retrieval system*”.

# Kecerdasasn Tiruan

Kecerdasan Tiruan atau sering disebut juga *Artifical Intelegence* (AI) merupakan mesin yang belajar dari pengalaman, menyesuaikan diri dengan *input*an baru dan melakukan tugas seperti manusia (immersa.lab, 2018). Dan salahsatu cabang dari kecerdasan tiruan adalah pembelajaran mesin (*Machine Learning*).

Machine Learning

Secara umum ‘*Machine Learning’* adalah salah satu disiplin ilmu dari *computerscience* yang mempelajari bagaimana membuat komputer atau mesin itu mempunyai kecerdasan. Agar mempunyai kecerdasan, komputer atau mesin harus dapat belajar. *Learning* sendiri mempunyai arti menambahpengetahuan atau memahami atau menguasai dengan belajar, mengikuti intruksi atau melalui pengalaman.

Pelatihan Mesin

Dalam melakukan pembelajaran mesin, maupun teknik-teknik atau metode-metode pelatihan yang dapat dipelajari oleh mesin sehingga mesin bisa belajar dan memiliki kecerdasan (Ni Komang Widyasanti, 2018).

]

Metode atau cara untuk mencapai tujuan yang telah ditetapkan meliputi tahapan-tahapan penelitian, peubah yang diamati/diukur, rancangan penelitian, serta teknik pengumpulan dan analisis data.

METODE

[

Pada penelitian ini, digunakan metode *N-Gram* dengan Algoritma *Winnowing* dan fungsi kesamaan pada *Jaccard.* Metodologi mencakup suatu teknik atau cara yang digunakan sehubungan dengan penelitian yang dilakukan, yang mempunyai step –step yang sistematis. Metode penelitian meliputi prosedur penelitian dan teknik penelitian.

Dalam penelitian dicoba untuk melakukan pendekatan ekperimen, dengan cara menguji coba terhadap bobot kata dalam pengklasifikasian dokumen, kemudian mencari keefektifitasannya dengan basis ketepatan.

# Metode Pemilihan Sampel

Sample adalah bagian dari jumlah dan karakteristik yang dimiliki oleh suatu populasi atau sebagian kecil dari anggota populasi yang diambil menurut procedur sehingga dapat mewakili populasinya. Teknik sampling merupakan teknik pengambilan sampel, terdapat berbagai macam teknik sampling untuk mentukan sample yang dapat dipakai dalam penelitian.

Untuk menguji tingkat efektifitas hasil penentuan artikel dari model, maka metode pengambilan sampel menggunakan metode *Cluster Sampling.* Sampel dalam penelitian ini sebanyak 200 artikel yang diperoleh dari sosial mediaonline. Turnbackhoax.com adalah media yang memvalidsi artikel *hoax* dan *non- hoax* dari media yang terdaftar di dewanpers.or.id. Artikel yang digunakanberupa teks tanpa disertakan gambar. Artikel diambil dari sosial media *online* karena banyak informasi yang tersebar melalui media sosial. Penelitian akan dibuat dua kategori atau label yaitu *hoax* dan non-*hoax*.

# Metode Pengumpulan Data

Penelitian membutuhkan data berupa informasi dan referensi sebagai sumber yang dapat mendukung materi uraian dan pembahasan, sebagai tahapan yang 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. 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 disusun dari tiap-tiap artikel dengan tujuan untuk mempermudah pengujian kebenaran dan keakuratan.

# Instumentasi

Instument yang digunakan dalam proses penelitian, adalah dokumen penunjangseperti buku, jurnal, skripsi dan tesis tentang penelitian terdahulu. Dokumen yang digunakan untuk mengklasifikasi yang mengimplementasikan *n-gram* dengan algoritma *winnowing* dan fungsi kesamaan *jaccard*.

# Teknik Analisis, Prototipe Model dan Pengujian Data Teknik Analisis

Teknik analis yang digunakan dengan menggunakan pendekatan Algoritma *winnowing*.

Rancangan model pada penelitian ini dibuatmemiliki proses sebagai berikut :

1. Tahap text preprocessing, yaitu melakukan proses tokenizing, stopword removal dan proses stemming menggunakan Algoritma Nazief dan Andriani.
2. Melakukan proses pencarian pembobotan pada setiap kata dari artikel tersebutdengan

menggunakan fungsi kesamaan Jaccard.

1. Proses penghitungan nilai similarity, dimana untuk mencari perbedaan artikel yang dimiliki dengan query dengan didalamnya terdapat Algoritma winnowing dan fungsi kesamaan Jaccard untuk menghitung nilai similarity pada artikel.

# Prototipe Model

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

Analisis tingkat efektifitas dari klasifikasi artikel dengan membutuhkan sebuah model klasifikasi yaitu *web based application* yang dimana dapat menerapkan *n-gram*, proses algoritma *winnowing* dan proses pembobotan berdasarkan nilai *similarity.* Proses yangdigunakan dalam penentuan artikel diantaranya yaitu :

* 1. *Upload* dokumen
	2. *Tokenization*
	3. *Removing stopword*
	4. Proses *wionning*
	5. Proses pembobotan kata
	6. Proses perhitungan nilai kemiripan (*similarity*)

# Teknik Pengujian Sistem

Teknik pengujian sistem yang digunakan sebagai berikut : Pengujian ketepatan (*Precision*)

Pengujian ketepatan (*Precision*) adalah perbandingan jumlah artikel relevan yang didapatkan sistem dengan jumlah seluruh artikel yang terambil oleh sistem baik relevan maupun tidak relevan.

*Precission* : Jumlah artikel yang relevan dengan *query* dan terambil Jumlah seluruh artikel yang terambil

.... (i)

Pengukuran kelengkapan (*Recall*)

Pengujian kelengkapan (*Recall*) ialah perbandingan jumlah artikel yang ditetapkan sistem dengan jumlah seluruh artikel relevan yang ada dalam koleksi artikel (diambil atau tidak diambil sistem).

*Recall* : Jumlah artikel yang relevan dengan *query* dan terambil Jumlah seluruh artikel relevan dan koleksi artikel

.... (ii)

# Langkah-langkah Penelitian

Langkah-langkah penelitian dapat dilihat pada struktur Gambar 3. di bawah ini

Identifikasi Masalah

Studi Literatur

Pengumpulan Data

Pembuatan Prototype

Hasil dan Kesimpulan

# Gambar 3. Langkah - langkah penelitian

Identifikasi Masalah

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

Studi Literatur

Proses ini dilakukan dengan cara mereview dari berbagai 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 algorima *winnowing*, dengan fungsi pembobotan fungsi kesamaan *Jaccard*, diharapkan dapat mengukur perbedaan dan memberikan hasil yang efisien dan validdalam pengklasifikasian artikel berita politik.

Pengumpulan Data

Pengumpulan data menggunakan cara dengan mengambil artikel pada media digital yaitu turnbackhoax.com, dan beberapa media yang terdaftar dewanpers.co.idyang 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 dariartikel berita politik. Pada kategori artikel politik akan diambil sebanyak 200 data. 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 masingmasing *word* pada setiap artiel teks, sehingga dapat menghasilkan *similarity* pada artikel tersebut. Berdasarkan pembobotan fungsi kesamaan *Jaccard* dan metode *N- Gram* dengan Algoritma *winnowing* untukpenentuan artikel.

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.*

# Tabel 1 Jumlah Data Artikel

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

]

Biaya dan jadwal penelitian disusun dengan mengisi langsung tabel berikut dengan memperbolehkan penambahan baris sesuai banyaknya jenis pengeluaran dan kegiatan.

BIAYA DAN JADWAL

|  |  |  |
| --- | --- | --- |
| No. | Jenis Pengeluaran | Biaya yang diusulkan (Rp) |
| 1 | Honor Peneliti | 4.000.000,- |
| 2. | Peralatan penunjang | 1.500.000,- |
| 3 | Bahan pakai dan peralatan | 2.000.000,- |
| 4 | Biaya Perjalanan | 2.500.000,- |
|  | **T O T A L** | **10.000.000,-** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Keterangan | Tahapan | Mei | Juni | Juli | Agustus |
| No | Aktivitas | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1. | Persiapan .Persiapan dilakukan dengan melakukan :* Identifikasi masalah yang terjadi. Identifikasi keunggulan sistem dalam mengatasi si masalah.
* Identifikasi kebutuhan– kebutuhan dalam penerapan system
* Penyusunan Proposal Penentuan tim yang akan terlibat dalam proyek
 | Planing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2. | Penelusuran requirement melalui wawancara dan observasi terhadap kebutuhan proses untuk pengumpulan data requirement tersebut | Analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3. | Memutuskan bagaimana suatu sistem beroperasi, dalam suatu perangkat keras/perangkat lunak, user interface, forms, reports,program khusus seperti database dan file-file yang digunakan. | Design |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4. | Uji coba | Implemntasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Publikasi dan HKI | Submitted |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 5. | Laporan Akhir | Finishing |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Daftar pustaka disusun dan ditulis berdasarkan sistem nama dan tahun dengan urutan abjad nama pengarang, tahun penerbitan, judul tulisan, nama jurnal, volume dan nomor penerbitan, serta halaman dimana artikel tersebut dimuat. Hanya pustaka yang disitasi pada usulan penelitian yang dicantumkan dalam Daftar Pustaka.

DAFTAR PUSTAKA

[

Ahmad Fauzi, G. (2018). INFORMATION RETRIEVAL SYSTEM PADA FILE PENCARIAN DOKUMEN TESIS BERBASIS TEXT MENGGUNAKAN METODE VECTOR SPACE. *Jurnal PILAR Nusa Mandiri* , 41-46.

Annisa, R. (2019). ANALISIS KOMPARASI ALGORITMA KLASIFIKASI DATA MINING UNTUK PREDIKSI PENDERITA PENYAKIT JANTUNG. *Jurnal*

*Teknik Informatika Kaputama (JTIK)* , 22-28.

Anwar, F. (2017). Perubahan dan Permasalahan Media Sosial. *Jurnal Muara Ilmu Sosial, Humaniora, dan Seni*, 137-144.

Eva Y Puspaningrum, B. N. (2018). A Comparison of the Vector Space Model Method and Winnowing Algorithm to Measure the Similarity of Documents. *The 5th International Conference on Information Technology and Bussiness* , 38-45.

Ginabila. (2020). Information Rerieval dan Perhitungan Kemiripan Dokumen Pada Indonesian Heritage Library Menggunakan Vector Space Model. *Jurnal Teknik Informatika Unika St. Thomas*, 236-245.

Glen Hizkia Oge Mangundap, H. S. (2022). Implementasi Algoritma Winnowing pada Aplikasi Pendeteksi Kemiripan Dokumen. *Jurnal Edukasi dan Penelitian Informatika*, 147-153.

HENDRIAN, S. (2018). ALGORITMA KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI SISWA DALAM MEMPEROLEH BANTUAN DANA PENDIDIKAN. *Faktor Exacta 11*, 266-273.

I Gede Iwan Supida, K. S. (2019). Pendekatan Penentuan Bobot dengan Surrogate Weighting Procedures untuk Metode Simple Additive Weighting dalam Pengambilan Keputusan Multikriteria. *International Journal of Natural Sciences and Engineering*, 113-121.

Indri Tri Hapsari, B. S. (2017). APLIKASI INFORMATION RETRIEVAL UNTUK PENCARIAN DOKUMEN LAPORAN PENELITIAN. *Jurnal Informatika*

*Polinema*, 22-28.

Mardi Siswo utomo, J. S. (2022). TEXT SUMMARIZATION PADA ARTIKEL BERITA MENGGUNAKAN VECTOR SPACE MODEL DAN COSINE SIMILARITY.

*Dinamika Informatika*, 11-24.

Ni Komang Widyasanti, I. K. (2018). Seleksi Fitur Bobot Kata dengan Metode TFIDF untuk Ringkasan Bahasa Indonesia. *Merpati*, 119-126.

Pratama, E. E. (2018). Information Retrieval pada Proses Penyimpanan dan Pencarian Dokumen Digital Menggunakan Metode Text Mining. *Seminar Nasional Teknologi Informasi Dan Komunikasi*, 736 - 742.

Rina Anggeriani, F. M. (2017). Pemanfaatan Vector Space Model Untuk Peringkasan Berita. *Journal of Big Data Analytic and Artificial Intelligence*, 16-22.

Tongkotow Liedfray, F. J. (2022). Peran Media Sosial Dalam Mempererat Interaksi Antar Keluarga Di Desa Esandom Kecamatan Tombatu Timur Kabupaten Minahasa Tenggara. *Jurnal Ilmiah Sociaty*, 1-13.

]