# BAB II

# KAJIAN PUSTAKA

## A. Dokumen Berita untuk Pengklasifikasian Data

Berita sendiri memiliki arti yakni informasi terbaru mengenai sesuatu yang sedang atau telah terjadi yang disajikan baik melalui media cetak maupun lisan (Ardianto & Erdiyana et al, 2004). Pada penelitian ini terdapat kumpulan dokumen berita tervalidasi berita *hoax* yang nantinya digunakan untuk pembuatan *library* kata. Berita *hoax* ini didapat dari website *turnbackhoax.id* yang memilki beberapa berita yang terbukti bukan berita benar atau tervalidasi berita *hoax.* Penggunaan klasifikasi dinilai tepat untuk identifikasi berita hoax dikarenan dalam metode klasifikasi terdapat teknik untuk mengidentifikasi pola dari kata pada berita yang terindikasi *hoax.* Data dokumen yang sudah dikumpulkan dibagi menjadi 2 buah bagian yakni data uji dan data latih. Data latih yakni berita tervalidasi hoax yang nantinya akan digunakan untuk pembuatan *library* kata. Data uji yakni data baik berita *hoax* maupun non *hoax* yang digunakan untuk mengukur performa sistem yang digunkan untuk klasifikasi berita *hoax*.

## B. Metode *Waterfall* untuk Pembuatan Aplikasi Klasifikasi Berita *Hoax*

Metode *waterfall* yakni metode pengembangan yang digunakan untuk penelitian dalam bidang sistem informasi. Model ini melakukan pendekatan secara urut dan sistematis yang dimulai dari tahap awal kebutuhan sistem kemudian tahap analisis, selanjutnya tahap desain, tahap *coding*, tahap *testing*, tahap *verifikasi*, dan tahap *maintenance* (Pascapraharastyan et al, 2014). Model metode ini dijuluki dengan *waterfall* dikarenakan tahapan tahapn yang dilakukan berjalan secara runtut dari tahap awal dilanjut dengan tahap selanjutnya yang dapat dikerjakan jika tahap sebelumnya telah diselesaikan.

## *C. Natural Language Processing*

*Natural Languge Processing* adalah sebuah pemrosesan bahasa alami yang digunakan untuk mengkaji bahasa manusia ke dalam bahasa computer. NLP ini berguna untuk memproses dan memahami bahasa manusia ke dalam komputer sehingga maksud dari target dapat tersampaikan dengan benar ke dalam komputer. Dalam hal ini salah satu kesulitan yang dihadapi yakni soal ambiguitas dari kata yang diberikan oleh manusia sehingga komputer salah dalam memproses maksud yang diinginkan. (Mariana Neves et al, 2016)

Penerapan teknik ini dalam dunia teknologi yakni steaming / pemotongan kata menjadi bentuk dasar, pembuatan ringkasan sebuah cerpen,pembuatan botchat untuk took-toko online dan lain sebagainya. Dalam hal ini *Natural Language processing* sangat berperan penting guna menghubungkan antara Bahasa Manusia ke dalam komputer sehingga dapat diolah sedemikian rupa sesuai kebutuhan.

Berikut adalah tahapan-tahapan pemrosesan dari NLP :

1. *Case Folding*

   Yakni perubahan huruf besar menjadi huruf kecil pada berita

2. *Tokenizing*

   Yakni sebuah proses pemisahan text menjadi per kata sehingga tiap kata dapat diolah dengan mudah.

3. *Stopwords Removal*

   Stopword merupakan kata umum yang digunakan / kata yang tidak penting untuk dianalisi / diproses. Seperti contoh kata : dan, atau, akan tetapi, jika. Tujuan dari penghapusan kata ini yakni mengurangi pembengkakan jumlah index yang digunakan.

4. *Stemming*

   Proses *stemming* yakni sebuah penghapusan kata yang memilki awalan / akhiran sehingga didapat kata dasar. Guna dari *stemming* ini yakni

meminimalisir persamaan kata yang memiliki perbedaan awalan / akhiran sehingga dapat memperkecil jumlah indeks yang akan diproses.

## D. Teknik *TF-IDF*

Metode *TF-IDF* (*Term Frequency and Inverse Document Frequency)* yakni sebuah metode pembobotan sebuah kata yang digunakan untuk memberikan nilai seberapa penting kata tersebut dalam pembentukan text yang digunakan. Dalam metode ini memiliki algoritma yang digunakan untuk mengukur bobot tiap-tiap kata dalam sebuah dokumen. Semakin besar bobot sebuah kata, maka semakin penting pula kata tersebut dalam sebuah dokumen. (Frista Gifti et al, 2018)

Proses dari TF-IDF yakni *Term Frequent* akan menghitung frekuensi kata yang muncul dan dibandingkan jumlah kata yang terdapat dalam sebuah dokumen. Berikut rumus persamaan dari metode Tf:

$$\text{Tf(i)} = \frac{freq(ti)}{\sum freq(t)}$$

Keterangan :

Tf (i) = Nilai *Term Frequent* sebuah kata dalam sebuah dokumen

Freq (ti) = Frekuensi kemunculan sebuah kata dalam sebuah dokumen

Freq (t) = Jumlah keselurahan kata dalam sebuah dokumen

Untuk proses IDF (*Inverse Document Frequent*) yakni menghitung jumlah seluruh dokumen yang dibandingkan dengan dokumen yang memiliki kata t muncul. Berikut rumus persamaan yang digunakan :

$$\text{Idf(i)} = \log \frac{|D|}{|\{d : ti \in d\}|}$$

Keterangan :

Idf = Nilai *Inverse Documen Frequent* yang terdapat dalam dokumen

|D| = Jumlah seluruh dokumen

$|\{d : ti \in d\}|$ = Jumlah dokumen yang mengandung kata (t)

Setelah diketahui Tf dan Idf dari 2 buah rumus persamaan yang telah dipaparkan maka langkah selanjutnya yakni mencari weight / bobot dari sebuah kata dengan mengalikan hasil dari TF dan IDF sebuah kata.

$$\text{W} = \text{tf(i)} \times \text{idf(i)}$$

Keterangan :

W = *Weight* / Bobot dari sebuah kata

Tf(i) = *TermFrequent* dari sebuah kata

Idf(i) = *Inverse Document Frequent* dari sebuah kata

## E. *Apriori Methode*

Metode *apriori* yakni sebuah metode yang digunakan untuk mencari pola frequensi penjualan sehingga dapat memaksimalkan laba penjualan dari sebuah toko. Akan tetapi metode ini dapat di aplikasikan kedalam stemming kata dengan memanfaatkan hasil *support* dari metode *apriori.* Cara kerja algoritma ini yakni :

1. *Itemset Frequent* yakni pencarian item yang sering muncul bersamaan dalam sebuah data. Contoh : kata berita dan hoax sering muncul dalam sebuah paragraph.
2. Knowledge Pencarian data / informasi yang penting dalam sebuah paragraph
3. Support (nilai penunjang) yakni presentasi dari record yang mengandung kombinasi item dibanding dengan jumlah record contoh : jika ada kata a dan b maka *support* dari {a,b} yakni peluang sebuah kata a dan b yang muncul dalam sebuah dokumen. (Gilang Abi Saputro et al, 2017)

   Rumus untuk menghitung nilai *support* pada suatu itemset yakni:

$$\text{Support (A)} = \frac{jumlah\ berita\ yang\ mengandung\ kata\ A}{Total\ Berita}$$

   Keterangan :

   *Support(A)* = Nilai Penunjang dari sebuah kata

## F. Bahasa Pemorgraman *Python*

Bahasa pemrograman *python* merupakan bahasa pemrograman yang termasuk kedalam *highlevel language. Hightlevel language* merupakan kategori bahasa pemrograman yang mendekati bahasa manusia. Penulisan dari bahasa python ini sangat *simple* sehingga tidak memerlukan banyak *space* untuk digunakan. Selain itu, keuntungan dari bahasa pemrograman *python* yakni tidak memakan banyak waktu dalam pembuatanya dikarenakan bahasa yang tidak terlalu rumit.
(Richard Halterman et al, 2011)

Bahasa pemrograman python termasuk kedalam bahasa pemrograman *open source* yakni dapat digunakan secara bebas sehingga banyak perusahaan yang memanfaatkan bahasa pemrograman *python* untuk memberikan pelayanan. Salah satu aplikasi yang dapat digunakan untuk mengembangkan sebuah software dengan mendukungbahasa pemrograman python yakni pycharm. Python juga mendukung pengembangan pada berbagai sistem operasi seperti : Linux, Android, Mac Os, Windows, Palm.

## G. Database MySQL

Yakni sebuah manajemen basis data SQL yang digunakan untuk penyimpanan, pengaturan dan pengolahan data yang nantinya akan dapat diakses dengan mudah.(Haris Saputro et al, 2012) MySQL termasuk kedalam *system database* gratis yang didistribusikan dibawah lisensi GPL (*General Public License*). Pada awalnya, MySQL ditemukan oleh Michael Monty Widenius pada tahun 1979 dari Swedia. Kelebihan dari database MySQL ini yakni :

1. Memiliki keamanan yang cukup baik sehingga pengguna dapat mengguanakan database ini untuk keperluan pribadi maupun komersial.
2. Gratis sehingga dapat digunakan oleh siapapun
3. Stabil dalam pengoperasianya sehingga dapat digunakan dengan optimal
4. Fleksibel dalam berbagai macam program sehingga dapat di akses dalam berbagai platform dan berbagai bahasa program

## H. Pengukuran Performa

Setelah sebuah *system* telah selesai dibuat dan menampilkan hasil yang telah diinginkan, maka langkah selanjutnya yakni akan dilakukan pengukuran performa dari *system* yang dibuat. Pengukuran performa dilakukan bertujuan menguji kinerja dan akurasi dari *system* yang sudah dibuat.Pengukuran performa ini dilakukan dengan menggunakan metode *Precision , Recall* dan *Accuracy.* (Frista Gifti et al, 2018)

*Precision dan Recall* yakni berguna untuk mengukur keefektifan pengambilan informasi. Rumus formula dari *Precision* dan *Recall* yakni :

$$Precision = \frac{\#(Dokumen\ hoax\ terklasifikasi\ hoax)}{\#(Jumlah\ h\ dokumen\ terklasifikasi\ hoax)}$$

$$Recall = \frac{\#(Dokumen\ hoax\ terklasifikasi\ hoax)}{\#(Jumlah\ h\ dokumen\ hoax\ yang\ diuji)}$$

| | Relevan | Tidak Relevan |
|---|---|---|
| Diambil | *True positve (tp)* | *Flase postive (fp)* |
| Tidak Diambil | *False negative (fn)* | *True negative (tn)* |

**Gambar 2.1 *Precision and Recall***

Berdasarkan gambar 2.1 dapat dirumuskan formula perhitungan akurasi sebuah *system* yakni

$$P = tp / (tp+fp)$$

$$R = tp / (tp+fn)$$

Keterangan :

P = *Precision*
R = *Recall*
Tp = *True Positive*
*Fp* = *False Positive*
*Tn* = *True Negative*
*Fn* = *False Negative*

Contoh terdapat 10 buah dokumen berita yang terdiri dari 5 buah konten berisi *fake news* dan 5 buah konten berisi berita *real.* Jika dalam *system* ini mendeteksi terdapat 6 buah berita yang terindikasi *fake* yakni 4 buah berisi konten *fake news* dan 2 buah termasuk berita *real* maka 4 dari berita yang telah dipilih termasuk tn (*true positive*) 2 yang dipilih termasuk tn (*false psotive*). 1 berita *fake* yang tidak terdeteksi termasuk *fn* (*false negative*). Dan 4 berita sisa yang tidak terdeteksi termasuk tn (*True Negative*).

Selain *Precision and Recall,* Perhitungan akurasi *system* juga diperlukan untuk memastikan seberapa akurat *system* dapat digunakan. maka langkah selanjutnya yakni perhitungan akurasi *system* dengan menggunakan rumus persamaan :

$$\text{ac} = \frac{\sum match\ h}{\sum tp} \text{ x } 100\%$$

Keterangan :

ac = Tingkat akurasi

$\sum match\ h$ = Jumlah hasil deteksi yang benar

$\sum tp$ = Jumlah data yang diuji

Jumlah deteksi benar yakni hasil dari penjumlahan *true positive* dan *True negative.*Lalu hasil tersebut akan dibagi dengan total dari seluruh data yang digunakan untuk pengujian. Setelah itu dikalikan 100% agar dapat mencari persentase tingkat akurasi dari peneltian ini.

## I. Penelitian yang Relevan

Dalam penelitian ini dilakukan penelusuran studi literatur pada beberapa penelitian terkait dengan klasifikasi berita hoax . Beberapa penelitian tersebut antara lain :

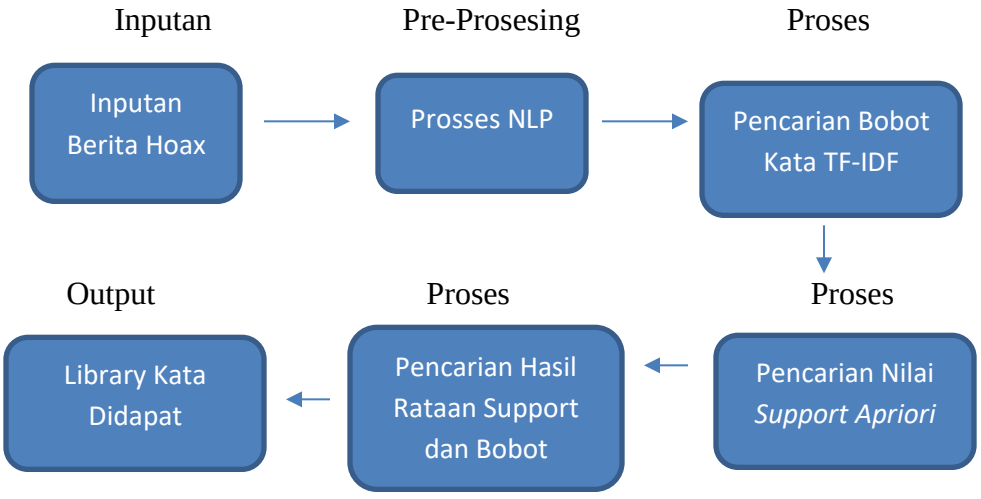
1. Penilitian yang dilakukan oleh Frista Gifti (2018) menjelaskan pendeteksian konten hoax berbahasa Indonesia menggunakan metode *Levenshtein Distance.* Dalam penilitan tersebut penulis menggunakan metode *Tf-Idf* untuk memberikan bobot kata yang nantinya akan dicari jarak kata asal dengan kata sumber menggunakan metode *Levenshtein Distance* pada sebuah berita. Hasil batas yang didapat yakni 0,0014 pada data 100 berita yang terindikasi berita *hoax.* Metode yang diambil dari penelitian ini yakni proses pengklasifikasian berita *hoax.*
2. Penelitian yang dilakukan oleh Marin Vukovic (2009) yang berjudul "*An Intelegent Automatic Hoax Detection Sistem*" menjelaskan tentang pengklasifikasian email hoax dengan membandingkan pola tersimpan yang sama. Kelemahan dari peneleitian ini yakni jika ada email yang memiliki pola baru, maka *system* belum dapat mengidentifikasi email tersebut. Metode yang diambil dari penelitian ini yakni proses pengklasifikasian berita *hoax*.
3. Penelitian yang dilakukan oleh Munjiah Nur Saadah, Widar Atmagi, Dyah S, Agus Zainal (2013) yang berjudul "Sistem Temu Kembali Dokumen Texs dengan pembobotan TF-Idf dan LCS" menjelaskan tentang bagaimana membangun system pengembalian sejumlah dokumen dengan metode tertentu yang memiliki relevansi tinggi sesuai dengan permintaan pengguna. Metode yang digunakan penulis yakni metode pembobotan dengan menggunakan Tf-Idf yang disesuaikan dengan menggunakan *LCS* guna mempertimbangkan kemunculan urutan kata yang sama antara query dengan text dalam document. Metode yang diambil dari penelitian ini yakni metode pembobotan kata.
4. Penelitian yang dilakukan oleh Gilang Abi Saputro (2017) yang berjudul "Penerapan Algoritma Apriori untk Mencari Pola Penjualan di Cave" yang

bertujuan untuk mencari pola penjualan dalam cave menggunakan metode *Apriori.* Dalam hal ini penulis melakukan pengujian system sebanyak 7 kali menggunakan data transaksi jurney coffe dengan merubah parameter *minimum support* dan *minimum confidence.* Konsep yang diambil dari penelitian ini yakni metode apriori yang digunakan.

## J. Kerangka Pemikiran

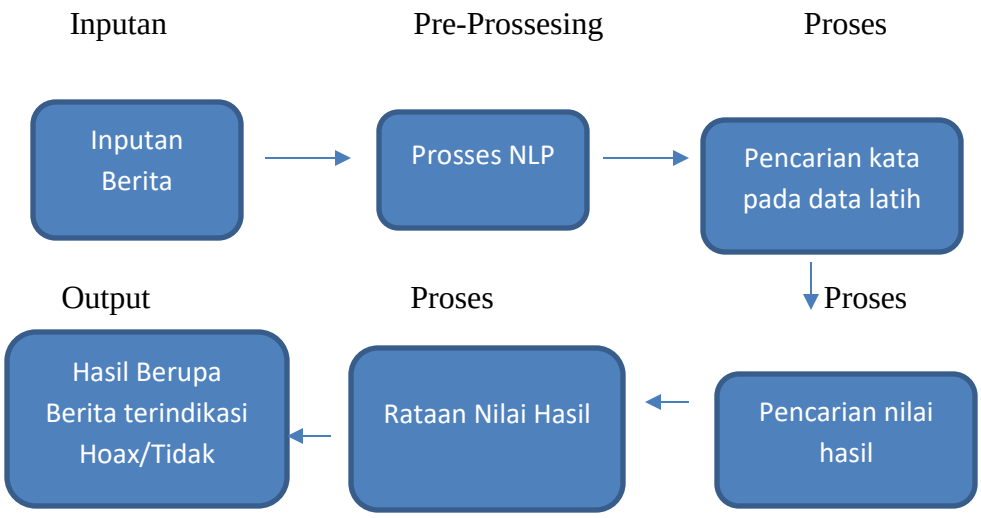
Penelitian ini menggunakan parameter berupa beriita hoax yang nantinya akan di proses dengan menggunakan metode *NLP ,Apriori dan TF-Idf untuk dijadikan dataset.* Hasil dari inputan tersebut yakni berupa kata yang memiliki nilai bobot tersendiri yang nantinya digunakan untuk menilai berita baru yang di inputkan sebagai data uji. Setelah di proses, hasil dari meotde akan diuji akurasi dengan menggunakan perhitungan akurasi *system.*

Proses inputan data latih :




**Gambar 2.2 Proses Input Data Latih**

Proses inputan data uji :




**Gambar 2.3 Proses Input Data Uji**