BAB 2 LANDASAN KEPUSTAKAAN

Di dalam bab ini akan membahas tentang kepustakaan yang digunakan pada penilitian ini. Pustaka ini berisikan penelitian-penelitian yang telah dilakukan terdahulu dan metode terkait yang digunakan untuk penelitian ini.

2.1 Tinjauan penelitian

Tinjauan penelitian terdahulu yang sudah dilakukan, digunakan untuk pembahasan penelitian terkait dengan penelitian ini. Salah satu penelitian oleh Nurjanah et al. pada tahun 2017 dengan judul "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode *K-Nearest Neighbor* dan Pembobotan Jumlah Retweet", menjelaskan bahwa nilai *k* metode *K-Nearest Neighbor* memiliki pengaruh pada penelitian analisis sentimen opini masyarakat di Twitter mengenai tayangan televisi ini dengan hasil akurasi yang optimal mencapai 80.83 saat nilai k = 3.

Penelitian selanjutnya Parlar et al. (2016) membandingkan performa dari *feature selection Chi-Square*, *Document Frequency Difference* (DFD), dan metode usulannya *Query Expansion Ranking* (QER). *Feature selection* yang dibandingkan performanya ini digunakan untuk penelitan analisis sentimen, dan hasilnya menunjukkan bahwa *feature selection Query Expansion Ranking* (QER) yang memiliki nilai akurasi tertinggi dibandingkan *feature selection* lainnya.

2.2 Twitter

Twitter adalah salah satu jejaring media sosial yang tengah populer saat ini. Twitter menyediakan layanan *microblogging*, sehingga penggunanya dapat mengirimkan pesan, membagikan pendapat tentang sesuatu atau membagikan cerita yang dialaminya dalam bentuk tulisan yang biasa disebut dengan *tweets*. Penulisan *tweets* ini hanya dibatasi dengan 140 karakter dalam sekali penulisan. Penjelasan dari fitur-fitur yang ada di dalam Twitter, yaitu sebagai berikut (O'Reilly dan Milstein, 2012):

1. Halaman Utama (*Home*)

   Halaman utama (*home*) ini memuat tentang *tweets* yang diposting oleh pengguna lain yang diikuti oleh kita.

2. Profil (*Profile*)

   Halaman profil (*profile*) ini memuat profil atau data diri tentang kita yang telah kita tulis dan simpan saat pertama kali mendaftar Twitter.

3. *Followers*

   Akun pengguna lain yang mengikuti kita atau menjadikan kita temannya sehingga semua yang kita posting dapat dibaca dan dimuat dalam halaman utama Twitternya.

4. *Following*

   Akun pengguna lain yang kita ikuti sehingga semua yang diposting oleh akun tersebut dapat kita baca dan dimuat dalam halaman utama (*home*) akun Twitter kita.

5. *Mentions*

   Konten ini berisi pesan percakapan yang bisa langsung menandai nama akun Twitter pengguna lain untuk dapat saling membalas pesan.

6. *Favorite*

   Memberikan tanda pada *tweets* sehingga *tweets* tersebut tidak hilang oleh halaman sebelumnya.

7. Pesan Langsung (*Direct Message*)

   Sebuah pesan yang kita kirim untuk pengguna lain yang sifatnya pribadi.

8. *Hashtag* (#)

   Tanda yang ditulis di depan sebuah kata tentang topik tertentu sehingga akan memudahkan pengguna lain untuk mencari topik yang serupa.

9. *List*

   Berisi seluruh pengguna Twitter yang diikuti agar memudahkan untuk melihat nama pengguna (*username*) secara keseluruhan.

10. Topik Terkini (*Trending Topic*)

    Topik yang sedang ramai diperbincangkan atau topik yang banyak ditulis oleh pengguna Twitter dalam waktu yang bersamaan.

2.3 Kurikulum 2013 (K-13)

Kurikulum 2013 atau biasa disebut dengan K-13 merupakan kurikulum pendidikan terbaru yang ditetapkan oleh pemerintah pada pertengahan tahun 2013 untuk menggantikan kurikulum lama yaitu Kurikulum Tingkat Satuan Pendidikan (KTSP). Pemerintah mengharapkan kurikulum baru ini dapat menjadikan siswa menguasai beberapa kompetensi, seperti:

- Kemampuan berkomunikasi
- Kemampuan hidup dalam masyarakat yang mengglobal
- Kemampuan untuk berpikir jernih dan kritis
- Memiliki kecerdasan sesuai bakat dan minat

Banyak terdapat perubahan mendasar dari pergantian kurikulum yang baru seperti konsep dari kurikulum, buku pelajaran, proses pembelajaran, serta penilaiannya. Dari konsepnya Kurikulum 2013 ini dapat menyeimbangkan antara *hardskill* dan *softskill*, dimulai dari Standar Kompetensi Lulusan, Standar Isi, Proses serta Penilaian.

Kemudian perubahan terjadi saat proses mengajar buku yang dipakai adalah buku siswa yang ditekankan pada *activity base* dan buku guru yang memuat panduan bagi guru untuk mengajarkan materi kepada siswanya. Proses pembelajaran di Kurikulum 2013 ini mendukung kreativitas siswa, di sini mengacu dalam buku Innovators DNA, Dyers, J.H. et al (2011) menjelaskan bahwa dari 3/3 bagian kemampuan kreativitas seseorang, 2/3 diperoleh melalui pendidikan dan 1/3 sisanya berasal dari genetik namun untuk kemampuan kecerdasan 1/3 dari pendidikan, dan 2/3 sisanya dari genetik. Kemampuan kreativitas dapat diperoleh melalui proses mengamati, bertanya, menalar, dan mencoba banyak hal sehingga proses pembelajaran kurikulum 2013 akan mengedepankan pengalaman personal.

Dengan perubahan dari kurikulum ini dapat menjadikan kurikulum pendidikan yang digunakan menyesuaikan perkembangan masyarakat, ilmu pengetahuan dan teknologi. Namun, setiap kali perubahan kurikulum dilakukan, selalu saja disambut pro dan kontra. Kurikulum 2013 menuai banyak kritik dan protes. Kritik dan protes datang dari berbagai kalangan menyangkut isi dan kemasan kurikulum, kesiapan guru dan lain-lain. Seperti pada akhir tahun 2014 saat Menteri Pendidikan dan Kebudayaan, Anies Baswedan menerbitkan peraturan bahwa Kurikulum 2013 (K-13) di sekolah rintisan diberhentikan sementara dan untuk sekolah rintisan dapat melaporkan kepada kepala dinas pendidikan untuk menggunakan kembali Kurikulum Tingkat Satuan Pendidikan (KTSP). *Website* kemenkopmk.go.id menyebutkan kebijakan tersebut tertera pada Permendikbud nomor 160 tahun 2014 yang efektif diberlakukan pada tanggal 12 Desember 2014. Permendikbud ini paling lama diberlakukan sampai tahun ajaran 2019/2020. Oleh sebab itu hingga sampai saat ini bahasan tentang Kurikulum 2013 masih banyak diperbincangkan oleh masyarakat pada media sosial khususnya Twitter dan salah satu *website* berita *online* yaitu kompas, menuliskan bahwa Kurikulum 2013 pada akhir tahun 2014 sempat menjadi *trending topic*.

2.4 *Text mining*

Text Mining bertujuan menemukan pola yang penting atau berguna dari sekumpulan dokumen. *Text mining* juga memiliki fungsi utama yaitu *information extraction* (IE), kategorisasi teks, *summarization*, *clustering* teks, *information monitor*, kemudian *question and answer* (Mustafa, 2009).

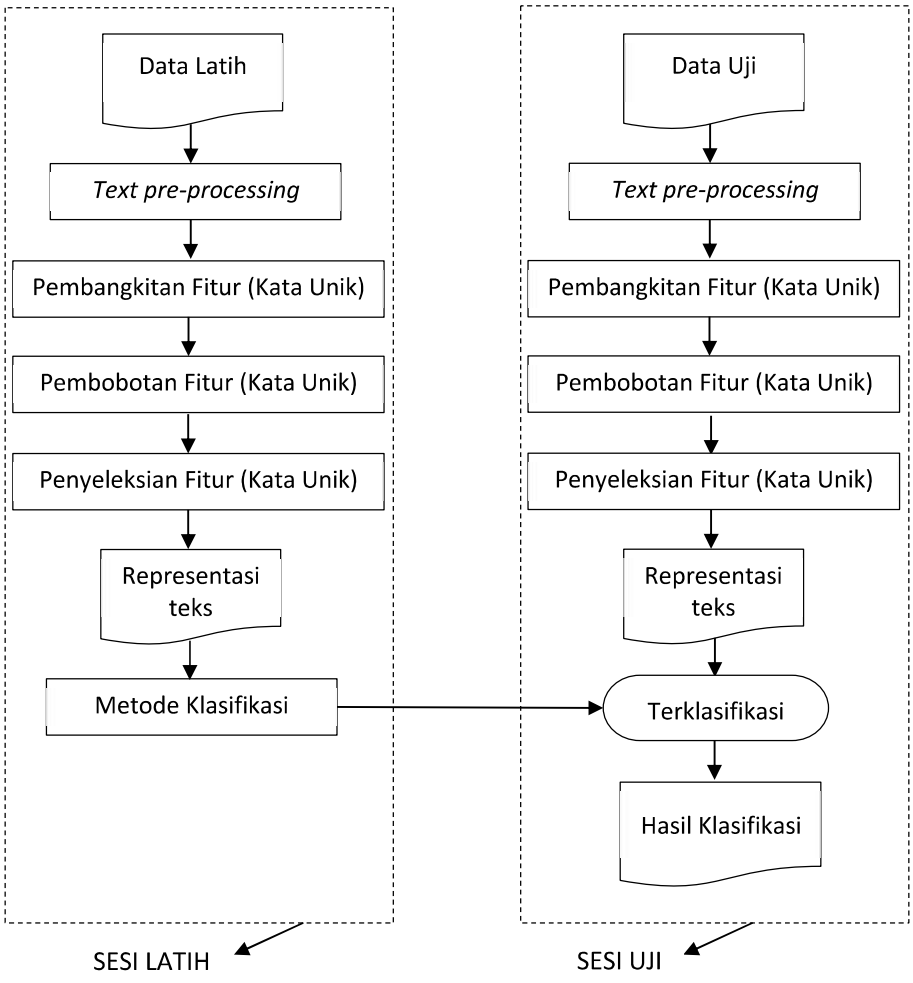
Terdapat manfaat juga dalam *text mining* ini yaitu dalam proses pencariannya dan pembuatan inovasi untuk dapat membantu manusia mengerti informasi serta menggunakan informasi dari sebuah *repository document* (Hand, 2010).

2.5 Analisis sentimen

Analisis sentimen bisa disebut juga *opinion mining* ini ditujukan untuk menganilisis opini, pendapat, penilaian, sikap, emosi dan sentimen terhadap produk, jasa, peristiwa, topik, organisasi dan lain sebagianya. Analisis sentimen merupakan salah satu bidang studi yang menjadi penelitian aktif dalam *natural language processing* sejak awal tahun 2000. Amerika Serikat merupakan salah

satu negara maju yang terdapat setidaknya 20-30 perusahaan yang menjual jasa untuk melayani dalam bidang analisis sentimen (Liu, 2010).

Menurut penelitian yang berjudul "*Affective-feature-based Sentiment Analysis using SVM Classifier*" dilakukan oleh Luo et al., sentimen analisis bertujuan menemukan orientasi emosi dari ulasan pengguna secara otomatis. Analisis sentimen jauh lebih kompleks daripada *topic mining*, analisis sentimen merupakan klasifikasi biner dengan pembagian klasifikasiannya menjadi dua jenis kelas yang pertama adalah positif dan yang kedua adalah negatif (Luo et al., 2016). Gambar 2.1 merupakan diagram alir proses klasifikasi sentimen analisis yang diambil dari penelitian Luo et al.



Gambar 2.1 Diagram alir proses klasifikasi sentimen analisis

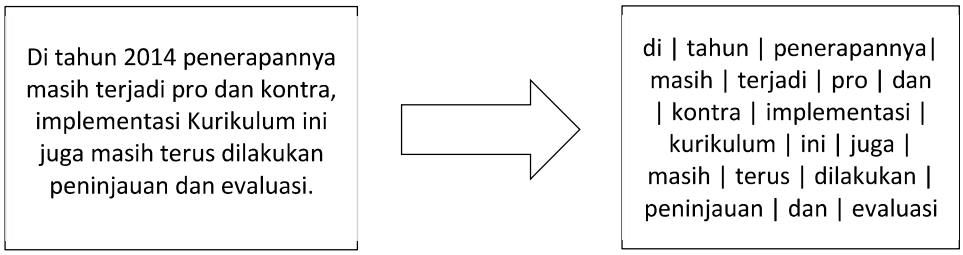
Sumber: [Luo et al., 2016]

2.6 *Text pre-processing*

Tahapan yang ada dalam *text pre-processing* ini meliputi tokenisasi, *filtering*, dan *stemming*.

2.6.1 Tokenisasi

Proses pemisahan suatu rangkaian kalimat yang menyusun sebuah dokumen dengan dijadikan kata per kata dan dilakukan penghilangan tanda baca, karakter dan angka selain huruf *alphabet*. Token disebut juga istilah (*term*) atau kata, misalnya sebuah token merupakan sebuah urutan karakter yang terdapat dalam dokumen yang dikelompokkan dan nilainya berguna untuk diproses (Amin, 2013). Pada Gambar 2.2 merupakan contoh ilustrasi proses tokenisasi.

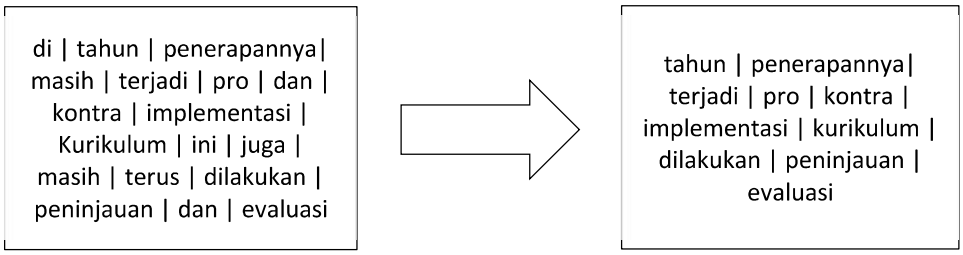


Gambar 2.2 Ilustrasi tokenisasi

2.6.2 *Filtering*

Filtering atau yang biasa disebut dengan *stopword removal*, proses ini memuat tahap pengambilan *term* yang dianggap penting atau menghapus *term* yang dianggap tidak penting atau tidak relevan. Terdapat 2 algoritme yang ada dalam *filtering*, yaitu pertama algoritme *stoplist* (membuang *term* yang kurang penting) ini merupakan kata atau *term* yang dapat dibuang dalam pendekatan *bag-of-words*. Dan *wordlist* (menyimpan *term* yang dianggap penting) yang mungkin digunakkan sebagai kata kunci dalam pencarian dokumen.

Digunakan daftar *stopword* bahasa Indonesia seperti; dan, ke, yang, di, kepada, dan yang lainnya. Keuntungan dalam proses eliminasi *stopword* adalah mengurangi space pada tabel *term index* hingga 40% atau lebih (Amin, 2013). Pada Gambar 2.3 merupakan contoh ilustrasi *filtering*.

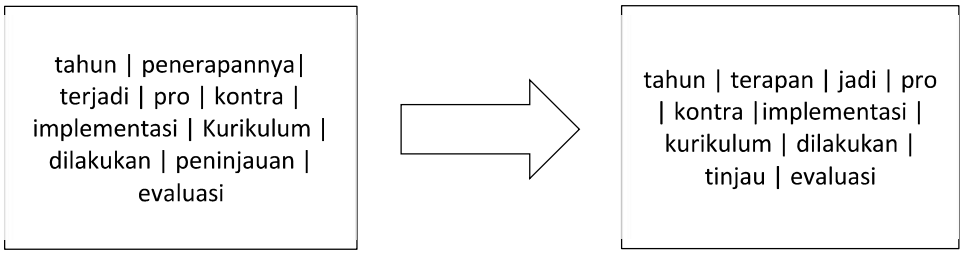


Gambar 2.3 Ilustrasi *filtering*

2.6.3 *Stemming*

Dalam proses ini dilakukan mengubah *term* yang masih melekat seperti awalan, sisipan dan akhiran. Pergantian *term* menjadi kata dasar dan harus sesuai struktur morfologi bahasa Indonesia yang benar (Amin, 2013). Stemming Bahasa Indonesia telah banyak dikembangkan oleh para peniliti, salah satunya adalah

library Sastrawi *stemming* yang ada pada Github. Penelitian yang dilakukan oleh Agusta (2009), melakukan perbandingan kinerja algoritme Nazief dan Adriani dengan algoritme Porter. Hasilnya menyatakan algoritme Nazief dan Adriani memiliki keakuratan yang lebih tinggi. Berikut pada Gambar 2.3 merupakan salah satu ilustrasi dari tahap *stemming*.



Gambar 2.3 Ilustrasi *stemming*

2.7 *Term weighting* (TF-IDF)

Term Frequency (TF) merupakan jumlah kemunculan sebuah *term* atau kata pada sebuah dokumen. *Local weight* yag biasa disebut dengan TF (*term frequency*) memiliki fungsi untuk menentukan bobot dari *term* t pada dokumen tertentu, yang nantinya akan menghasilkan estimasi berdasarkan frekuensi atau *relative frequency* dari *term* t pada dokument tertentu (Kao et al., 2007). Ada empat cara yang dapat digunakan untuk menghitung serta mendapatkan nilai TF tersebut, yaitu Raw TF, Logarithmic TF, Binary TF, dan Augment TF (Fitri, 2012). Namun, pada penelitian ini digunakan Logarithmic TF dengan Persamaan pada 2.1 berikut ini.

$$TF = 1 + \log(D) \quad (2.1)$$

Keterangan:
D adalah kemunculan *term* pada dokumen

Inverse Document Frequency (IDF) memiliki fungsi jika suatu *term* kemunculannya banyak tersebar di seluruh dokumen sehingga untuk mengurangi bobot *term* tersebut digunakan IDF. TF dan IDF merupakan satu kesatuan yang baik digunakan untuk jumlah dokumen yang besar. Berikut perhitungan IDF pada Persamaan 2.2 dan perhitungan dari TF-IDF pada Persamaan 2.3.

$$\boldsymbol{IDF = \log_{10}(\frac{N}{DF})} \quad (2.2)$$

Keterangan:
N = Jumlah banyaknya koleksi dokumen
DF = Jumlah banyaknya dokumen yang mengandung *term*

$$\boldsymbol{Term\ Weighting = TF * IDF} \quad (2.3)$$

Keterangan:
TF = Nilai hasil *Term Frequency*
IDF = Nilai hasil *Inverse Document Frequency*

2.8 *Feature selection Query Expansion Ranking*

Feature selection atau penyeleksian fitur ini berguna dalam pengurangan dari dimensi data, bisa juga menghilangkan data yang tidak relevan. Menurut Jain dan Zongker (1997) masalah *feature selection* didefinisikan saat diberikan sekumpulan fitur lalu dipilih beberapa fitur yang mampu memberikan hasil yang terbaik pada proses klasifikasi. Beberapa tahun terakhir, banyak selection methods fitur telah diusulkan seperti *Mutual Information* (MI), *Chi-Squared* (CHI), *Document Frequency* (DF), *Information Gain* (IG), *Cross Entropy* (CE) dan sebagainya (Liu et al., 2010).

Feature selection yang digunakan penulis adalah *Query Expansion Ranking*. *Query Expansion* proses yang mana *query* awal diformulasi dengan ditambahkannya beberapa *term* untuk meningkatkan perfoma dalam proses *information retrieval* (Sukarno, 2016).

Mengacu pada penelitian yang dilakukan oleh Parlar et al. (2016) menjelaskan fakta bahwa para ahli riset telah mengembangkan teknik *Query Expansion* untuk memperbaiki proses menemukan dokumen yang cocok untuk sebuah *query*. Dokumen yang telah ditemukan oleh sistem *retrieval* dan dianggap cocok dengan *query*-nya akan dikembalikan ke *user*, kemudian *user* akan menandai dokumen tersebut dan *term* yang ada pada dokumen tersebut akan diproses serta diberi skor. *Query* awal akan diperluas dengan *term* yang memiliki skor terbaik selanjutnya *query* tersebut diberikan kembali ke sistem untuk dilakukan pencarian dokumen yang paling cocok. Penelitian tersebut juga terinspirasi dari *probabilistic weighting model* yang digunakan untuk menetapkan skor sebuah kata. Berikut Persamaan 2.4, Persamaan 2.5 dan Persamaan 2.6 menunjukkan proses perhitungan yang digunakan untuk *feature selection*.

$$pf = \frac{df_{+}^{f} + 0.5}{n^{+} + 1.0} \quad (2.4)$$

Keterangan:

pf = Nilai probabilitas *term f* pada dokumen data latih kategori positif.

df_{+}^{f} = Jumlah dokumen yang mengandung *term f* yang ada pada data latih kategori positif.

n^{+} = Jumlah seluruh dokumen data latih kategori positif.

$$qf = \frac{df_{-}^{f} + 0.5}{n^{-} + 0.5} \quad (2.5)$$

Keterangan:

qf = Nilai probabilitas *term f* pada dokumen data latih kategori positif.

df_{-}^{f} = Jumlah dokumen yang mengandung *term f* yang ada pada data latih kategori negatif.

n^{-} = Jumlah seluruh dokumen data latih kategori negatif.

$$score_f = \frac{|pf + qf|}{|pf - qf|} \quad (2.6)$$

Keterangan:
$score_f$ = Hasil perhitungan *Query Expansion Ranking* untuk *term f*.
pf = Nilai probabilitas *term f* pada dokumen data latih kategori positif.
qf = Nilai probabilitas *term f* pada dokumen data latih kategori positif.

2.9 *K-Nearest Neighbor*

K-Nearest Neighbor merupakan salah satu metode yang banyak digunakan untuk klasifikasi pada teknik Data Mining. Metode ini secara umum bekerja dengan cara melakukan perhitungan jarak antara data latih dan data uji, untuk data latih dimasukkan terlebih dahulu dan perlu adanya pelabelan setiap data. Penentuan kelas pada data uji akan ditentukan berdasarkan nilai *k* tetangga terdekat dari nilai jarak data latih teratas yang telah diurutkan (Sani, et al., 2016), kemudian mayoritas dari nilai *k* tetangga terdekat akan menjadi dasar untuk keputusan dari kategori data uji (Yu, 2010). Namun pada penelitian ini perhitungan jarak dengan tetangga terdekat menggunakan metode *cosine similarity*.

Dalam pengklasifikasian teks, semakin besar nilai *cosine similarity* akan semakin dekat tingkat kemiripan antara data uji dan data latihnya dan sebaliknya jika nilai *cosine similarity*-nya semakin kecil maka akan semakin jauh tingkat kemiripan antara data uji dan data latihnya (Luhulima, 2015). Perhitungan *cosine similarity* dapat dilihat pada Persamaan 2.7.

$$\cos Sim(A,B) = \frac{\Sigma_{i=1}^{n} A_i B_i}{\sqrt{\left(\Sigma_{i=1}^{n} A_i\right)^2}\sqrt{\left(\Sigma_{i=1}^{n} B_i\right)^2}} \quad (2.7)$$

A merupakan data uji, dan *B* merupakan data latih. A_i dan B_i merupakan bobot nilai yang diberikan untuk setiap *term* yang ada.

Setelah semua proses perhitungan selesai, nilai dari proses perhitungan *cosine similarity* diurutkan mulai dari nilai yang terbesar hingga terkecil. Kemudian setelah selesai diurutkan data akan diambil sebanyak *k* dengan tujuan penentuan kelas dari data uji.

2.10 Evaluasi

Dalam penelitian ini, perhitungan akurasi dilakukan dengan menggunakan *confussion matrix*. *Confussion matrix* adalah salah satu metode untuk evaluasi yang menggunakan matrix seperti pada Tabel 2.2.

Tabel 2.2 *Confussion Matrix*

		Nilai sebenarnya	
		True	*False*
Nilai prediksi	*True*	TP (*True positive*)	FP (*False Positive*)
	False	FN	TN

		(*False Negative*)	(*True Negative*)

Sumber: [Han & Kamber, 2006]

Keterangan :

- TP (*True Positive*) menunjukkan dokumen yang hasil pengklasifikasian manual (positif) dan pengklasifikasian oleh sistem (positif).
- FN (*False Negative*) menunjukan dokumen yang hasil pengklasifikasian manual (negatif) dan pengklasifikasian oleh sistem (positif).
- FP (*False Positive*) menunjukkan dokumen yang hasil pengklasifikasian manual (positif) pengklasifikasian oleh sistem (negatif).
- TN (*True Negative*) meenunjukkan dokumen yang hasil pengklasifian manual (negatif) dan pengklasifikasian oleh sistem (negatif).

Sehingga diperoleh formulasi evaluasi untuk melakukan pengujian pada hasil klasifikasi dengan rumus pada Persamaan 2.8 berikut.

$$\text{Akurasi} = \frac{TN+TP}{(FP+FN+TP+TN)} \text{x} 100\% \qquad (2.8)$$