BAB II
LANDASAN TEORI

2.1 Text Mining

Text mining adalah salah satu penambangan informasi yang berguna dari data – data yang berupa tulisan dokumen atau text dalam bentuk klasifikasi maupun clustering (H, 2015). *Text mining* masih merupakan bagian dari data mining dimana akan memproses data – data atau *text* – *text* serta dokumen – dokumen yang bisa jadi dalam jumlah sangat besar. Untuk memproses data yang sangat besar tentulah akan memakan sumber daya yang tidak sedikit kaitanya dengan pengolahan data tersebut. Disinilah diperukanya sebuah pemrosesan awal atau *preprocessing data text* tersebut sebelum data tersebut di lakukan proses *text mining* sesuai algoritma yang akan diterapkan.

2.2 Preprocessing

Preprocessing merupakan tahap untuk memproses data teks untuk didapat dianalisa. Pada tahap *pre-processing* data, data *tweet* mentah terlebih dahulu dilakukan proses *case folding*, *tokenizing*, *stemming*, serta *filtering*. Hasil dari tahapan ini menghasilkan fitur yang digunakan sebagai data pembelajaran mesin oleh NBC (Wahana & Chrisnanto, 2016). Dalam tahapan *preprocessing*, terdiri dari proses *case folding*, *tokenizing*, *stemming* dan *filtering*. Berikut ini merupakan penjelasan dari tahapan *preprocessing*:

1.) Pada tahapan *case folding*, teks dilakukan proses perubahan dari huruf besar menjadi huruf kecil dan menghilangkan seluruh tanda baca pada kalimat.

2.) Pada tahapan *tokenizing*, setiap kata akan dipisahkan berdasarkan spasi yang ditemukan.

3.) Pada tahapan *stemming*, yaitu pengubahan kata berimbuhan menjadi kata dasar.

4.) Pada tahapan *filtering*, yaitu pembuangan kata-kata tidak penting dari hasil token

2.3 Naïve Bayes Classifier

Naïve Bayes Classifier adalah suatu model independen yang membahas mengenai klasifikasi sederhana berdasarkan teorema Bayes. Naïve Bayes merupakan suatu algoritma yang dapat mengklasifikasikan suatu variable tertentu dengan menggunakan metode probabilitas dan statistik. Secara garis besar algoritma Naïve Bayes dapat dijelaskan seperti persamaan (Kurniawan, 2018):

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} \quad (1)$$

Keterangan:

R : Data yang belum diketahui kelasnya

S : Hipotesis pada data R yang merupakan class khusus

P(R|S) : Nilai probabilitas pada hipotesis

R yang berdasarkan kondisi S

P(R) : Nilai probabilitas pada hipotesis R

P(S|R) : Nilai probabilitas S yang berdasarkan dengan kondisi hipotesis R

P(S) : Nilai probabilitas S

2.4 K Nearest Neighbor

K-*Nearest Neighbor* merupakan metode yang memakai supervised algorithm dengan hasil dari query instance terbaru, diklasifikasikan berlandaskan yang paling banyak dari suatu kategori pada K-*Nearest Neighbor*. Algoritma ini merupakan mengelompokkan objek baru bedasarkan atribut dan *sample latih*. Selain itu, *classifier* berdasarkan pada memori. Jika diberikan titik *query*, maka akan ditentukan sejumlah k objek atau (titik *training*) yang paling dekat dengan titik query. Adapun klasifikasi menggunakan voting merupakan yang terbanyak diantara klasifikasi dari k obyek. Disisi lain, algoritma KNN menggunakan klasifikasi tetangga sebagai nilai prakiraan dari *query instance* yang terbaru(Setiawan, T. A., Wahono, R. S., & Syukur, 2015).

Algoritma KNN cukup mudah karena bekerja menurut jarak terdekat dari *query instance* ke *sample latih* untuk menentukan KNN-nya. Selain itu, *sample latih* digambarkan ke ruang berdimensi banyak, didalam masing-masing dimensi tersebut bisa merepresentasikan fitur data. Lalu, ruang dibagi menjadi beberapa mengikuti klasifikasi sample latih. Kemudian, satu titik diruang ini akan di beri tandai kelas c, jika kelas c adalah klasifikasi terbanyak ditemui pada k buah tetangga terpendek dari titik itu. Dekat atau jauhnya tetangga dihitung menggunakan Euclidean Distance.

Euclidian Distance dirusmuskan pada persamaan (2).

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

2.5 Confusion Matrix

Untuk menguji hasil klasifikasi pada sistem yang telah dibangun, maka dibutuhkan suatu metode perhitungan evaluasi performasi yaitu dengan menghitung nilai precission, recall dan f1-measure. Dalam evaluasi performasi ini akan dihitung nilai akurasi dan F1-*Measure*. Akurasi adalah bagaimana suatu sistem dapat melakukan klasifikasi *true* pada data true dan false, sedangkan f1-measure untuk menilai performasi dari keseluruhan sistem dengan menghitung nilai *precision* dan *recall*. Adapun perhitungan dapat dilihat pada *confussion matrix*: (Arini, Wardhani, & Octaviano, 2020)

Tabel 4. 2.1 Confusion Matrix

		Nilai Aktual	
		Positive	Negative
Nilai Prediksi	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TP (*True Positive)* merupakan prediksi positif dan nilai sebenarnya positif, TN (*True Negative*) merupakan prediksi negatif dan nilai sebenarnya negatif, FP (*False Positive*) merupakan prediksi positif dan nilai sebenarnya negatif dan FN (*False Negative*) merupakan prediksi negatif dan nilai sebenarnya positif.

Rumus dari akurasi dapat dilihat pada persamaan berikut :

$$Precission = \frac{TP}{TP+FP} \quad (1) \qquad Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+TN} x100\% \quad (3) \qquad F1 - Measure = 2x \frac{precission*recall}{precission+recall} \quad (4)$$

2.6 Jurnal Penelitian Yang Berkaitan

Penelitian yang dilakukan Nur Shafiya Nabilah Salam, Ahmad Afif Supianto dan Andi Reza Perdanakusuma yang berjudul "Analisis Sentimen Opini Mahasiswa Terhadap Saran Kuesioner Penilaian Kinerja Dosen dengan Menggunakan TF-IDF dan K-Nearest Neighbor" . Data pada kolom saran tentang perspektif mahasiswa terhadap kinerja dosen, hingga saat ini belum diolah secara mendalam untuk dijadikan bahan evaluasi. Untuk itu penelitian ini dilakukan agar dapat membantu Tim UJM mengolah data dari kolom saran yaitu menggunakan Analisis Sentimen pada tingkat kalimat, dengan metode klasifikasi K-Nearest Neighbor. Dari 2210 data opini mahasiswa Program Studi Teknologi Informasi yang dianalisa dalam tiga semester yaitu Semester Genap 2016/2017, Ganjil 2017/2018, dan Genap 2017/2018, diperoleh hasil klasifikasi dengan rata-rata Accuracy 81%. Hasil penelitian ini berupa daftar data opini dan grafik frekuensi data opini yang terklasifikasi dan divisualisasikan menjadi tampilan dashboard, serta dapat ditampilkan dengan fungsi filter berdasarkan nama dosen dan mata kuliah. Hasil dari klasifikasi yang telah dilakukan dengan metode K-Nearest Neighbor kemudian dilakukan pengujian menggunakan confusion matrix, dan diperoleh nilai dari Accuracy, Precision, Recall dan F1-Score dari ketiga semester. Setelah di rata-rata diperoleh nilai Accuracy, Precision, Recall, dan F1-Score sebesar 0.81 atau 81%. (Salam, 2019)

Sedangkan Penelitian serupa juga dilakukan oleh Mulyono dan teman – teman nya yang berjudul "Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes". Pada penelitian ini dilakukan analisis sentimen terhadap postingan opini pelanggan online marketplace di Indonesia pada twitter yang nantinya bisa digunakan untuk menentukan rating online marketplace agar masyarakat tidak salah pilih situs marketplace untuk berbelanja di masa depan. Langkah pertama melakukan koleksi data opini masyarakat di twitter dari situs belanja online. Kemudian dilakukan pre-processing pada data yang meliputi cleansing data, case folding, tokenizing, case normalization, stop word, convert negation dan stemming. Selanjutnya dilakukan proses pemberian label (kelas) pada data tersebut yang dilakukan oleh ahli bahasa. Hasil dari proses clustering tercipta dua kelas data yaitu kelas data positif dan kelas data negatif dengan jumlah total 1200 data. Data yang sudah memiliki kelas data ini, digunakan sebagai data training untuk mesin pengklasifikasi, dalam riset ini menggunakan algoritma mesin pengklasifikasi Naïve Bayes. Terakhir, kami mengukur kinerja dari mesin pengklasifikasi menggunakan 10-fold cross validation. Hasil evaluasi menunjukkan rata-rata akurasi sebesar 93.33% (Muljono, M., Artanti, D. P., Syukur, A., & Prihandono, A, 2018).