Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 5, No. 11, November 2021, hlm. 5075-5079 http://j-ptiik.ub.ac.id

Perbandingan Pembobotan *Term Frequency-Inverse Document Frequency* dan *Term Frequency-Relevance Frequency* terhadap Fitur N-Gram pada Analisis Sentimen

Randy Ramadhan[1], Yuita Arum Sari[2], Putra Pandu Adikara[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]randyyramadhan14@gmail.com, [2]yuita@ub.ac.id, [3]adikara.putra@ub.ac.id

Abstrak

Analisis sentimen merupakan salah satu metode yang digunakan untuk mengekstrak sentimen dalam kalimat berdasarkan isinya. Analisis sentimen merupakan salah satu metode dalam text mining yang menggunakan proses text preprocessing yang setelahnya terdapat suatu proses yaitu pembobotan kata. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode pembobotan kata yang paling populer dari kategori *unsupervised term weighting* yang dilansir tidak cocok untuk melakukan pengelompokan teks. *Term Frequency-Relevance Frequency* (TF-RF) merupakan metode penggabungan antara TF dan RF dengan tujuan untuk mendapatkan performansi yang lebih baik, metode ini berfokus pada seluruh dokumen yang mengandung term atau tidak mengandung term. *Twitter* merupakan tempat untuk masyarakat mencurahkan isi pikiran tentang pandemi yang dialami. Ulasan tentang hal karyawan dirumahkan pada *Twitter* perlu diklasifikasikan menjadi ulasan positif, negatif, dan netral, yang berguna untuk menjadi pertimbangan perusahaan dan pemerintah untuk melakukan keputusan dalam kebijakan PSBB. Terdapat beberapa tahap penelitian ini yaitu *preprocessing* untuk pemrosesan dokumen, dan menggunakan fitur *unigram* dan *bigram* serta pembobotan kata menggunakan metode TF-IDF dan TF-RF lalu dalam pengklasifikasian menggunakan metode klasifikasi *K-Nearest Neighbor*. Data yang digunakan sebanyak 246 data latih dan 90 data uji. Hasil terbaik dari perbandingan evaluasi yang didapatkan adalah dengan menggunakan pembobotan kata TF.RF dengan fitur *unigram* pada klasifikasi KNN dengan nilai K = 3 yaitu *accuracy* dengan 0,677, *precision* sebesar 0,526, *recall* dengan 0,654, serta *f-measure* dengan nilai 0,583. Nilai bigram tidak berpengaruh besar dalam penelitian ini dikarenakan nilai *f-measure* terbaik didapatkan bigram dengan nilai 0,591, serta nilai unigram terbaik dengan nilai 0,583.

Kata kunci: *analisis sentimen, karyawan dirumahkan, TF.IDF, TF.RF, unigram, bigram, KNN*

Abstract

Sentiment analysis is a method used to extract sentiments in sentences based on their content. Sentiment analysis is a method in text mining that uses a text preprocessing process after which there is a process, namely word weighting. Term Frequency-Inverse Document Frequency (TF-IDF) is the most popular word-weighting method from the unsupervised term weighting category reported which is not suitable for grouping texts. Term Frequency-Relevance Frequency (TF-RF) is a method of combining TF and RF with the aim of getting better performance, this method focuses on all documents that contain terms or do not contain terms. Twitter is a place for people to express their thoughts about the pandemic they are experiencing. Reviews about employees being sent home on Twitter need to be classified into positive, negative, and neutral reviews, which are useful for companies and government consideration to make decisions in PSBB policies. There are several stages of this research, namely preprocessing for document processing, and using unigram and bigram features as well as word weighting using the TF-IDF and TF-RF methods in classification using the K-Nearest Neighbor classification method. The data used were 246 training data and 90 test data. The best results from the evaluation comparisons obtained are using TF.RF word weighting with the unigram feature in the KNN classification with a value of K = 3, namely accuracy of 0.677, precision of 0.526, recall of 0.654, and f-measure of 0.583. Bigram value does not have a big effect in this study because the best f-measure value is obtained Bigram with a value of 0.591, and the best unigram value is with a value of 0.583.

Keywords: *sentiment analysis, dismissed employees, TF.IDF, TF.RF, unigram, bigram, KNN*

1. PENDAHULUAN

Analisis sentimen merupakan salah satu metode yang digunakan untuk mengekstrak sentimen dalam kalimat berdasarkan isinya. Analisis sentimen menganalisis dokumen online seperti blog, review, komentar dan mengategorikan sebagai positif negatif dan netral (Saberi & Saad, 2017). Analisis sentimen merupakan salah satu metode dalam *text mining* yang menggunakan proses *text preprocessing* yang setelahnya terdapat suatu proses yaitu pembobotan kata.

Pembobotan kata merupakan tahap yang sangat penting setelah proses *preprocessing*, tujuan pembobotan kata merupakan mengubah data yang belum terstruktur menjadi lebih terstruktur, setelah itu data yang telah terstruktur dikategorikan menggunakan metode *classifier*. Nilai pembobotan kata dapat berbeda-beda berdasarkan metode pembobotan kata itu sendiri. *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan metode pembobotan kata yang paling populer dari kategori *unsupervised term weighting* yang dilansir tidak cocok untuk melakukan pengelompokan teks (Carvalho & Guedesa, 2020) terbukti pada penelitian yang dilakukan oleh Diki Susandi menggunakan TF.IDF, akurasi dan kelengkapan sistem dalam mengklasifikasi dokumen adalah 70,6% (Susandi, 2016), *Term Frequency-Relevance Frequency* (TF-RF) merupakan metode penggabungan antara TF dan RF dengan tujuan untuk mendapatkan performansi yang lebih baik, metode ini berfokus pada seluruh dokumen yang mengandung *term* atau tidak mengandung *term* (Lan et al., 2009), pada penelitian yang dilakukan oleh Thopo Martha Akbar dengan membandingkan beberapa metode pembobotan kata terhadap performansi kategorisasi teks, berdasarkan hasil evaluasi TF.RF lebih baik dari metode yang lain untuk sebagian besar pengujian yang dilakukan (Akbar, 2012) pada dasar ini akan dilakukan perbandingan dari kedua metode untuk melihat hasil dari evaluasi dari kedua metode dengan menggunakan analisis sentimen pada media sosial Twitter.

Twitter merupakan sosial media yang berguna untuk penggunanya untuk *posting* dan membaca pesan teks atau cuatan hingga 140 karakter (panjang pesan teks) yang dapat digunakan dari ponsel serta perangkat internet berupa aplikasi serta situs web (Maclean et al., 2013). Hal ini yang menyebabkan opini dari pengguna Twitter yang menjadi bahan analisis untuk menjadikan landasan untuk penelitian ini, mengingat maraknya pandemi yang terjadi di Indonesia yang terjadi awal Maret 2020.

Pada tahun 2019 terdapat virus baru yang dapat ditularkan dari manusia ke manusia di China dan menyebar sangat luas dan cepat di 190 negara dan teritori lainnya, dari sampel yang diteliti menunjukkan bahwa ada coronavirus baru, penyebaran virus ini sangat cepat, dan kemudian WHO menamakan nama baru pada virus ini yaitu *Severe Acute Respiratory Syndrome Coronavirus-2* (SARS-CoV-2), dan wabah ini dinamakan *Coronavirus Disease* atau biasa disebut COVID-19. Hingga tanggal 2 Maret 2020 virus tersebut pertama kali di Indonesia dengan kasus sejumlah 2 kasus dan sampai tanggal 15 September 2020 terdapat kurang lebih 222 ribu kasus di Indonesia (Susilo et al., 2020).

Kebijakan tanggap darurat dan Pembatasan Sosial Berskala Besar (PSBB) menjadi kebijakan pemerintah untuk mengatasi penyebaran Covid-19. Pandemi ini menimbulkan dampak terhadap masyarakat diberbagai bidang terutama bidang perekonomian, dan dampak terbesar yang dapat dirasakan adalah pada karyawan yang bekerja di beberapa perusahaan, beberapa perusahaan memilih untuk memulangkan atau melaksanakan *Work From Home* (WFH). Dan tidak sedikit masyarakat atau karyawan yang mengeluh terhadap pandemi yang melanda Indonesia. Sosial media adalah tempat yang banyak dipilih untuk para penggunanya mengutarakan pendapat serta keluhan terhadap pandemi ini, salah satunya adalah Twitter.

Pengelompokan *tweet* opini masyarakat mengenai Karyawan Dirumahkan dapat dilakukan dengan suatu metode klasifikasi. Metode yang digunakan pada penelitian ini adalah metode *K-Nearest Neighbor* (KNN) dengan perbandingan pembobotan kata yang telah dibahas sebelumnya, serta fitur *N-gram* untuk mendapatkan nilai dari term nya. *N-gram* dilakukan pada pemisahan teks atau kalimat dengan panjang n dari posisi tertentu dalam suatu teks, beberapa jenis *N-gram* yang sering digunakan adalah *Unigram*, *Bigram*, dan *Trigram* (Ahmed, 2017)

2. DASAR TEORI

2.1. Twitter

Twitter adalah contoh sistem *weibo* yang memungkinkan pengguna mengirim dan menerima posting singkat yang disebut *tweet*. *Tweet* dapat berisi hingga 140 karakter dan dapat berisi tautan ke situs web dan sumber terkait. Tepatnya, dalam beberapa tahun terakhir, *Twitter* resmi menambah jumlah karakter yang bisa digunakan hingga 280 karakter pada 7 November 2017.

2.2. Karyawan Dirumahkan

Karyawan dirumahkan disebabkan oleh dampak pandemi Covid-19. Beberapa perusahaan mengalami kerugian yang sangat signifikan sehingga perusahaan mengambil langkah-langkah untuk mengatasi kerugian tersebut, diantaranya *Work From Home* (WFH), PHK paksa dan PHK sementara. Bisnis yang ditutup telah selesai, yang menyebabkan banyak karyawan mengeluh tentang pandemi saat ini di Indonesia.

2.3. Analisis Sentimen

Analisis sentimen adalah bidang penelitian yang menganalisis pandangan, emosi, evaluasi, penilaian, sikap, dan emosi orang tentang produk, organisasi, individu, masalah, peristiwa, atau topik. (Liu, 2012). Masalah penting dalam analisis sentimen adalah untuk menentukan ekspresi emosional dalam teks dan apakah ekspresi subjek positif (menguntungkan) atau negatif (tidak menguntungkan) (Nasukawa & Yi, 2003).

2.4. Preprocessing Text

Teks preprocessing merupakan item yang dipesan untuk mempermudah dalam proses pengambilan data karena data yang diolah akan lebih terstruktur. Untuk preprocessing teks pada sistem pencarian akan dilakukan *data cleaning, case folding*, tokenisasi, *filtering* dan *stemming.*

2.5. N-Gram

N-gram adalah proses yang digunakan untuk penambangan teks dan pemrosesan bahasa, yang memperoleh hingga n karakter dari sebuah *string*. Dalam segmentasi kata, metode ini dilakukan dengan memisahkan hingga n kata dari rangkaian kata (paragraf, kalimat, bacaan) dibaca terus menerus dari awal sampai akhir dokumen. *N-gram* ukuran satu disebut *unigram*, *N-gram* ukuran dua disebut *bigram*, *N-gram* ukuran tiga disebut *trigram*, dan seterusnya (Sugianto et al., 2013).

2.6. Pembobotan Kata

Pembobotan kata merupakan salah satu metode yang memberikan manfaat. Diharapkan dapat ditemukan pentingnya kata dalam kalimat yang telah ditentukan sebelumnya. Mencari teknik pembobotan yang terbaik adalah dengan menggabungkannya dengan teknik pembobotan yang benar (Tantyoko, 2019). Metode pembobotan kata yang digunakan pada penelitian TF.IDF dan TF.RF yang dapat dilihat pada Persamaan1 dan Persamaan2.

$$Wt,d = tft,d\ x \log\left(\frac{D}{df}\right) \quad (1)$$

$$TFt,d\ x\ RF = TFd,t\ x\ RF\ c,t \quad (2)$$

Keterangan:
Tft,d = Jumlah kemunculan term (td) dalam dokumen (Dt).
RF = Relevant Frequency.

2.7. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) merupakan algoritma berbasis pembelajaran dengan dataset training disimpan. Sehingga klasifikasi untuk record baru yang tidak diklasifikasikan diperoleh dengan membandingkan record yang paling banyak mirip dengan data latih. *Cosine Distance* dan *Euclidean Distance* merupakan cara menghitung jarak antara dua data. Terdapat Persamaan 3 dan Persamaan 4 yang menjelaskan perhitungan kedua metode tersebut.

$$CosSimDistance(d_i, q_i) = 1 - \left(\frac{\sum_{i=1}^{n} d_i x q_i}{\sqrt{\sum_{i=1}^{n}(d_i)^2 x}\sqrt{\sum_{i=1}^{n}(q_i)^2}}\right) \quad (3)$$

$$d_{ij} = \sqrt{\sum_{i=1}^{n}(d_{ik} - d_{jk})^2} \quad (2)$$

Keterangan:
dij = Euclidean Distance dokumen i terhadap j

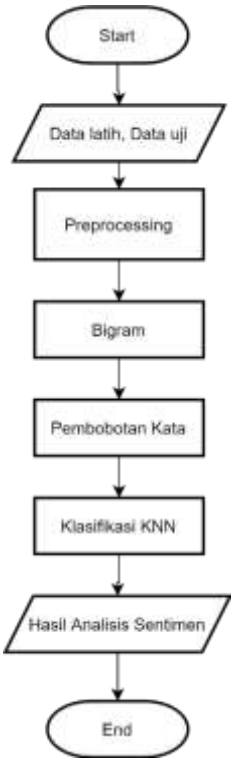
3. METODOLOGI PENELITIAN

3.1. Data Penelitian

Data yang digunakan sebanyak 243 data latih dan 90 data uji, Terdapat 135 ulasan negatif, 59 ulasan positif, 49 ulasan netral pada data latih. Data dikumpulkan secara manual dari halaman web *www.twtter.com* dengan *keyword* karyawan dirumahkan.

3.2. Implementasi Sistem

Implementasi sistem pada penelitian ini terdiri dari beberapa tahapan seperti yang ditujunkan pada Gambar 1.

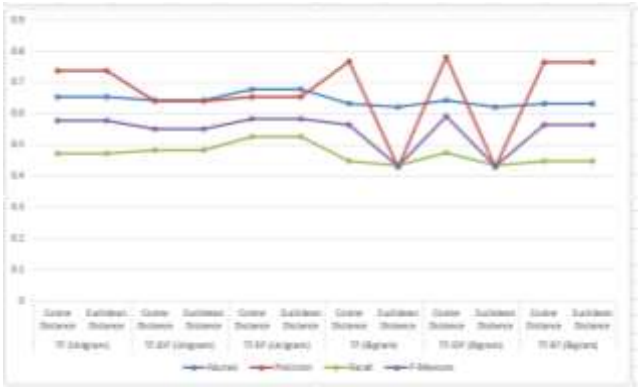


Gambar 1 Diagram Alir Implementasi Sistem

Dalam penelitian ini input yang digunakan pada tahap ini berupa *tweet* tentang pegawai yang dipulangkan, dan *tweet* tersebut menjadi data pelatihan dan data uji. Ada tahap *preprocessing* yang berguna untuk mengolah suatu dokumen agar dapat digunakan untuk tahapan klasifikasi, fitur *bigram* yang dilakukan untuk menghasilkan *term* dengan dua kata, dengan menggunakan TF.IDF dan TF.RF untuk pembobotan kata, serta menggunakan metode *K-Nearest Neighbor* (KNN) untuk klasifikasi. Setelah menyelesaikan tahapan tersebut, keluarannya akan berada pada kategori positif, negatif dan netral.

4. HASIL ANALISIS

Pada bagian ini membahas tentang pengujian yang dilakukan dan analisis terhadap hasil pengujian tersebut. Pengujian yang dilakukan menggunakan 90 *tweet* yang dijadikan sebagai data uji, dengan 52 ulasan negatif, 25 ulasan netral, dan 13 ulasan positif. Pengujian dilakukan berdasarkan nilai K, variasi *distance* serta menggunakan n-gram.



Gambar 2 Hasil Analisis Pengujian

Pada Gambar 6.1 merupakan hasil dari analisa perhitungan evaluasi pengujian dengan nilai K = 3 karena memperoleh hasil *accuracy* terbaik pada metode TF.RF menggunakan unigram term dengan nilai evaluasi pada varian jarak *Cosine Distance* dan *Euclidean Distance* yang sama yaitu 0,677. Namun, untuk nilai *precision* yang tertinggi adalah metode TF.IDF menggunakan unigram term dengan 0,565 dengan varian jarak *Cosine Distance* dan *Euclidean Distance*.

Nilai *precision* tertinggi diraih pada metode TF.IDF dengan metode *Cosine Distance* dengan nilai 0,782 hal ini disebabkan oleh banyaknya hasil prediksi Negatif yang dihasilkan pada hasil ini, karena bigram pada TF.IDF menerima informasi term yang didapatkan dalam dokumen kecil, dikarenakan prediksi data uji yang telah sebelumnya lebih banyak dari kategori lainnya dan hasil prediksi negatif lebih banyak dari kategori lainnya, contoh term yang pasti dinilai negatif adalah "karyawan, dirumahkan", dikarenakan kalimat "karyawan dirumahkan" merupakan kata kunci dari pencarian data kali ini, dan kategori yang dominan pada data latih adalah negatif, maka sistem akan menyimpan kata "karyawan dirumahkan" menjadi masuk ke kategori negatif.

Pada nilai *accuracy* terjadi penurunan yang rendah ke arah bigram term, yang disebabkan kebanyakan satu term bigram hanya ada pada satu dokumen, yaitu dokumen ketika term itu muncul salah satunya adalah term "karyawan, dirumahkan". Varian jarak pada penelitian ini cenderung memiliki perhitungan jarak yang hampir sama, pada *accuracy* tertinggi perhitungan *Cosine Distance* dan *Euclidean Distance* sama. Namun dari segi *recall*, *Euclidean Distance* lebih memiliki nilai yang rendah disebabkannya hasil prediksi yang tidak relevant terhadap hasil aktual dari data uji.

5. PENUTUP

Hasil *accuracy* terbaik dihasilkan dengan menggunakan metode TF.RF serta menggunakan unigram *term*, menghasilkan nilai *accuracy* sebesar 0,677 dengan menggunakan variant jarak *Cosine Distance* dan *Euclidean Distance*. Penggunaan term bigram tidak berpengaruh besar terhadap penelitian ini, terlihat pada hasil f-measure bigram memiliki nilai f-measure tertinggi dengan nilai 0,591 serta unigram memiliki nilai tertinggi 0,583, hanya terpaut 0,01 di atas hasil f-measure dari unigram, namun unigram mendapat nilai *accuracy* tertinggi sebesar 0.677 dibandingkan bigram.

Penambahan data latih sangat disarankan untuk penelitian selanjutnya, mengingat saat proses klasifikasi data latih lebih beragam serta *accuracy* juga bertambah. Serta pelabelan kelas pada data latih dan data uji sangat dibutuhkan ketelitian pelabelan kelas sangat berpengaruh pada hasil evaluasi. Perlu ditambahnya metode klasifikasi lain untuk melakukan perbandingan yang dilakukan agar didapat metode terbaik untuk melakukan analisis sentimen.

6. DAFTAR PUSTAKA

Ahmed, H. (2017). Detecting Opinion Spam and Fake News Using *N-gram* Analysis and Semantic Similarity. *University of Ahram Canadian*.

Asriningtias, Y., Mardhiyah, R., Studi, P., Informatika, T., Bisnis, F., Informasi, T., & Yogyakarta, U. T. (2014). *APLIKASI DATA MINING UNTUK MENAMPILKAN INFORMASI*. *8*(1), 837–848.

Deolika, A., Kusrini, K., & Luthfi, E. T. (2019). Analisis Pembobotan Kata Pada Klasifikasi Text Mining. *Jurnal Teknologi Informasi*, *3*(2), 179.

Ernawati, S., & Wati, R. (2018). Penerapan Algoritma K-Nearest Neighbors Pada Analisis Sentimen.

Indhiarta, W. C. (2017). *Penggunaan N-gram Pada Analisa Sentimen*. 1–18. Introduction, A., & Retrieval, I. (2009). *Online edition (c) 2009 Cambridge UP*. *c*.

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(4), 721–735.

Liu, B. (2012). Sentiment Analysis and Opinion Mining Morgan & Claypool Publishers. *Language Arts & Disciplines*, *May*, 167.

Maclean, F., Jones, D., Carin-Levy, G., & Hunter, H. (2013). Understanding twitter. *British Journal of Occupational Therapy*, *76*(6), 295–298.

Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, *March*, 70–77.

Saberi, B., & Saad, S. (2017). Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(5), 1660–1666.

Sugianto, S. A., Liliana, & Rostianingsih, S. (2013). Pembuatan Aplikasi Predictive Text Menggunakan Metode *N-gram*-based. *Jurnal Infra*, *1*(2). https://www.neliti.com/id/publications/105718/pembuatan-aplikasi-predictive-text-menggunakan-metode-*n-gram*-based

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, H., Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., & Yunihastuti, E. (2020). Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*, *7*(1), 45.