ISSN: 2477-3042 SURYA INFORMATIKA, VOL. 11 No.1, November 202

SENTIMENT EMBEDDINGS WORD2VEC PADA KLASIFIKASI KEPUASAN KARYAWAN PADA MANAJEMEN RTO GROUP

Edy Subowo[1] , Tohir Jaya[2], Nuridin [3]

Informatika

Universitas Muhammadiyah Pekajangan Pekalongan

Jl. Raya Pahlawan No. Gejlig – Kajen Kab. Pekalongan

Telp.: (0285) 385313, e-mail: fastikom.umpp@gmail.com

ABSTRAKSI

Penyebaran informasi yang semakin meningkat di media sosial memudahkan pengguna untuk mengungkapkan pandangan dan pendapatnya. Opini dan reaksi dapat berupa opini positif atau negatif atau dapat diartikan sebagai sentimen. Pada riset ini dibuat sistem analisis sentimen berdasarkan data kebijakan pemerintah pada media sosial Twitter. Dalam membangun sistem analisis sentimen ini, data yang digunakan adalah data yang berisi tweet dengan keyword yang telah ditentukan dan menggunakan Word2Vec. feature expansion dapat mengoreksi perbedaan kosakata dalam data tweet yang random dan terbatas untuk mendapatkan hasil pemrosesan kata yang maksimal sehingga didapatkan akurasi sebesar 79,52%.

Kata Kunci : analisis sentimen, Word2Vec

Abstract

The increasing dissemination of information on social media makes it easier for users to express their views and opinions. Opinions and reactions can be positive or negative opinions or can be interpreted as sentiments. In this research, a sentiment analysis system was created based on government policy data on Twitter social media. In building this sentiment analysis system, the data used is data that contains tweets with predetermined keywords and uses Doc2Vec. feature expansion can correct for vocabulary differences in random and limited tweet data to get maximum word processing results so that an accuracy of 79.52% is obtained..

***Keywords** : sentiment analysis, Doc2Vec*

1. Pendahuluan

1.1 *Latar Belakang*

Teknik yang digunakan dalam pengekstrakan informasi yang berupa pandangan (sentimen) dari individu/kelompok terhadap isu atau kejadian yang ada. Analisis Sentimen dapat dimanfaatkan untuk mengungkap beberapa hal seperti opini publik terhadap suatu isu yang berkembang, kepuasan terhadap pelayanan, kebijakan yang dibuat, cyber bullying, memprediksi harga saham, serta analisis kompetitor berdasarkan data yang dikumpulkan.

Analisis sentimen adalah sebuah cara yang digunakan untuk mengolah komentar yang diberikan oleh pemesan atau pelanggan melalui berbagai media, mengenai sebuah produk, jasa ataupun sebuah instansi. Permasalahan yang dihadapi RTO Group adalah pengolahan komentar pelanggan terhadap produk dan layanan yang banyak masuk untuk ritel, sehingga pihak Ritel RTO mengalami kesulitan dalam menangani komentar tersebut. Oleh karena itu dirancang sebuah sistem yang dapat membantu pihak R dalam mengetahui dan mengelompokkan komentar pelanggan, berdasarkan kelompok kategori dan sentimen. Hal tersebut dapat digunakan pihak RTO untuk evaluasi produk dan layanan RTO [1].

Klasifikasi merupakan proses memprediksi kelas dari data yang diberikan. Kelas diartikan juga sebagai kategori atau label/target. Klasifikasi masuk ke dalam kategori metode supervised learning yang merupakan pengembangan dari ilmu machine learning. Implementasi klasifikasi banyak dimanfaatkan di beragam disiplin ilmu, salah satunya di bidang ilmu komputer. Salah satu topik pada bidang ilmu komputer yang memanfaatkan metode klasifikasi dalam penyelesaiannya adalah Analisis sentimen. Beberapa kasus analisis sentimen yang cukup sering dibahas adalah seperti analisis sentimen mengenai suatu produk, tokoh publik, atau review terhadap suatu tempat. Dalam melakukan proses klasifikasi, ada beberapa metode yang bisa digunakan untuk diimplementasikan, seperti Support vector machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Tree, Neural Network.

Deep learning adalah bagian dari kecerdasan buatan serta pembelajaran mesin dimana metode deep learning adalah pengembangan dari neural network multiple layer yang berfungsi untuk memberikan ketepatan sebagaimana dalam object detection, speech recognition, language translation, dan lain sebagainya. Deep learning saat ini banyak dikembangkan dalam berbagai sektor karena fungsi deep learning yang dapat mempermudah pekerjaan

ISSN: 2477-3042 SURYA INFORMATIKA, VOL. 11 No.1, November 202

manusia terutama dalam mengolah data sehingga dapat menghasilkan informasi yang dapat dimanfaatkan [2].

Word Vector Representation (Representasi Vektor Kata) merupakan sebuah hasil pembelajaran dari

algoritma deep learning yang bertujuan untuk mengekstraksi kata. Terdapat beberapa metode yang dapat digunakan dalam mengimplementasikan Word Vector Representation diantaranya terdapat metode Word2vec yang telah dikembangkan oleh Google dan Word2Vec yang dikembangkan oleh Stanford University [3].

Berdasarkan berbagai riset dan permasalahan yang telah dijelaskan, maka pada riset ini, dilakukan analisis sentimen menggunakan Word2Vec karena memiliki kriteria sebagai pemroses arti pandaan kata. Alasan penggunaan teknik tersebut adalah untuk dapat mengatasi ketidakcocokan kosakata yang berasal dari data tweet yang kalimatnya terbatas sehingga muncul berbagai macam variasi kata, maka digunakan teknik feature expansion dengan metode word embedding Word2Vec.

1.2 *Batasan Masalah*

Batasan yang menjadi ruang lingkup di riset ini, yaitu data yang digunakan ialah data sentimen Bahasa Indonesia yang bersumber dari twitter sebanyak 500 tweet yang berkaitan dengan RTO, proses pelabelan sentimen dilakukan secara manual menjadi dua kategori, yaitu positif serta negatif, nilai matriks performansi yang dipakai adalah akurasi dan F1-Score, serta word embedding yang digunakan adalah Word2Vec.

Tujuan dari penelitian ini untuk merancang dan Membangun Sistem Informasi Transaksi Tabungan Nasabah Bank Mini IQTI Berbasis Android yang diharapkan dapat memperbaiki sistem yang sudah ada menjadi lebih baik.

2. Tinjauan Pustaka dan Landasan Teori

2.1 *Tinjauan Pustaka*

Informasi dapat didapatkan secara masif dan mudah melalui internet dengan memanfaatkan berbagai media sosial yang ada. Peningkatan informasi yang dihasilkan dari media sosial sangatlah besar setiap harinya. Setiap terjadi suatu peristiwa atau kejadian di dunia nyata akan cepat tersebar ke seluruh dunia melalui media sosial. Setiap pengguna media sosial dapat memberikan opininya terhadap suatu peristiwa yang terjadi dan dari fenomena tersebut dapat menimbulkan suatu sentimen yang terbentuk. Sentimen yang muncul dapat berupa sentimen yang positif maupun negatif. Analisis sentimen yang terbentuk dapat menjadi acuan terhadap suatu peristiwa yang terjadi, apakah peristiwa yang terjadi tersebut merupakan suatu peristiwa yang positif atau negatif.

Terdapat berbagai riset yang telah dikembangkan pada analisis sentimen. Dalam melakukan analisis sentimen, terdapat banyak teknik word embedding yang bisa dipergunakan dalam melakukan analisis sentimen, seperti metode Word2vec, CCA, Word2Vec, Doc2vec, dan lain sebagainya. Pada riset sebelumnya mengenai analisis sentimen kolom komentar kuisioner evaluasi dosen oleh mahasiswa yang dilakukan menggunakan metode Word2vec menghasilkan akurasi mencapai 70%, hasil tersebut didapatkan karena data yang digunakan sedikit dan akurasi tertinggi didaptakan saat menggunakan BOW+TF-IDF dengan akurasi 85% [4]

Pada riset [5]dengan memanfaatkan average base Word2vec dan bag of centroid Word2vec dengan klasifikasi Support vector machine (SVM) didapatkan 85,3% saat dua model digabungkan.Pada riset lainnya, yaitu analisis sentimen calon gubernur DKI Jakarta tahun 2017 yang memanfaatkan Lexicon Based Method untuk menentukan class sentiment dan menggunakan dua teknik klasifikasi yaitu Naïve Bayes dan Support vector machine (SVM) didapatkan akurasi terbaik saat menggunakan Naïve Bayes membentuk akurasi hingga 95%[6]

Pada riset [2], dengan memanfaatkan metode yang sama metode klasifikasi Support Vector Machine (SVM) dan Lexicon Based Features mencapai akurasi 79%, sementara sistem analisis sentimen yang tanpa Lexicon Based Features mencapai akurasi yang lebih besar yaitu 84% pada parameter yang sama.

Pada riset [7], dengan memanfaatkan feature expansion digunakan beberapa metode word embedding untuk klasifikasi topik dari sebuah tweet. Pada riset tesebut dijelaskan bahwa panjang kalimat dari sebuah data tweet yang terbatas dapat mempengaruhi kalimat aslinya sehingga mengurangi makna sebenarnya dari sebuah kalimat sehingga dengan metode tersebut dapat mengurangi ketidakcocokan kosakata yang diakibatkan dari kosakata yang hilang atau kurang didalam kalimat terbatas. Pada riset ini, menunjukkan metode yang digunakan tersebut dapat meningkatkan akurasi sebesar 0,38% pada algoritma klasifikasi.

Berdasarkan berbagai riset sebelumnya dan sepengetahuan penulis, belum terdapat riset mengenai analisis sentimen menggunakan dataset dari kumpulan data tweet pada media sosial twitter berbahasa Indonesia yang menerapkan feature expansion dengan metode Word2Vec. Maka dari itu, pada penlitian ini penulis mencoba untuk membuat sebuah Sistem analisis sentimen twitter dengan feature expansion dan memanfaatkan metode word embedding Word2Vec dengan algoritma klasifikasi Support Vector Machine (SVM).

3. METODOLOGI

1) Alat dan Media percobaan

Alat yang digunakan adalah seperangkat komputer yang terhubung dengan Google Collabs.

Sistem dibangun dengan bahasa python dengan library numpy,pandas, sklearn dan mathplotlib pada platform KERAS.

Algoritma pembelajaran embedding kata yang mendominasi didasarkan pada hipotesis distribusi yang menyatakan bahwa representasi kata dapat direfleksikan oleh konteksnya. Cara efektif untuk menyandikan konteks kata menjadi representasi kata adalah "prediksi konteks". Pada kata target w_i dan kata konteksnya h_i, "Prediksi konteks" bertujuan untuk memprediksi w_i berdasarkan h_i, yang dapat dilihat sebagai pemodelan bahasa. Konteks kata target bisa jadi sebelum ceding, kata-kata berikut atau sekitarnya

ISSN: 2477-3042 SURYA INFORMATIKA, VOL. 11 No.1, November 202

terjadi dalam sebuah teks.

$$h_i = \{w_{i-c}, w_{i-c+1}, \dots, w_{i-1}, w_{i+1}, w_{i+c}\} \quad (1)$$

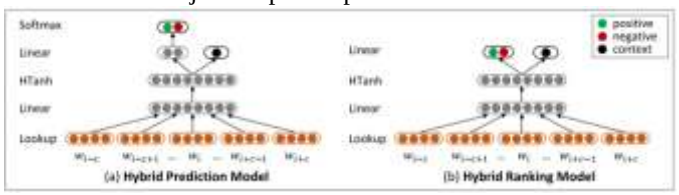
. Lapisan pencarian (juga disebut sebagai lapisan proyeksi) berisi tabel pencarian $LT \in R^{dx|V|}$ yang memetakan setiap kata ke vektor kontinu, di mana d adalah dimensi dari setiap vektor kata dan |V| adalah ukuran kosa kata. Operasi pencarian dapat dilihat sebagai fungsi proyeksi yang menggunakan file vektor biner idx saya yang nol di semua posisi kecuali di *i* index.

$$e_i = LT.idx_i \in R^{1xd} \quad (2)$$

Dimana e_i adalah embeddings kata konteks $\{w_{i-c}, w_{i-c+1}, \dots, w_{i-1}, w_{i+1}, w_{i+c}\}$ sebagai keluaran dari lapisan pencarian, yang diformalkan seperti di bawah ini

$$O_{lookup} = [e_{i-c}, \dots, e_{i-1}, e_{i+1}, \dots, e_{i+c}] \quad (3)$$

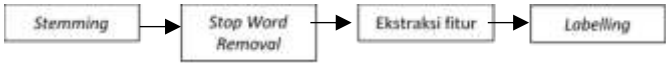
menggunakan pendekatan peringkat berpasangan untuk menangkap konteks kata untuk mempelajari embeddings kata. Ini memegang ide yang sama dengan estimasi kontrastif kebisingan tetapi tujuan pengoptimalannya adalah untuk menetapkan kata nyata pasangan konteks. Gambar 2 menunjukkan proses prediksi model.



Gambar 2. Ilustrasi model yang menangkap konteks kata serta sentimen kalimat

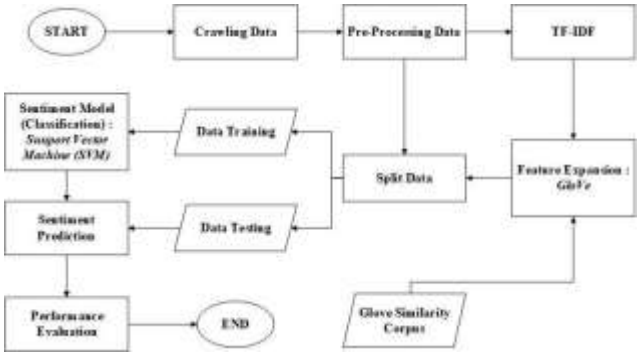
2) Data yang digunakan

Data yang digunakan adalah data twitter yang di crawl dari 1 Januari 2020 sampai 19 Desember 2020 sebanyak 3109 baris dengan 4 buah keyword, yaitu pembakaran sampah, asap kendaraan, asap polusi, dan asap rokok. Langkah – langkah pre-processing dimulai dari stemming sampai labelling dapat dilihat pada Gambar 3.



Gambar 3 Text Pre-processing

Gambaran atau deskripsi sistem menggambarkan bagaimana tahap pengerjaan dalam riset. Pada riset ini membahas tentang feature expansion Word2vec untuk analisis sentimen mengenai kebijakan publik di Twitter. Alur pengerjaan pada riset ini dapat dideskripsikan pada Gambar 4.



Gambar 4. Sistem Analisis Sentimen Menggunakan Feature Expansion Word2Vec

3.2 Crawling dan Pelabelan Data

Tahapan paling awal dari pembuatan sistem analisis sentimen adalah crawling merupakan proses pengumpulan data baik berukuran besar maupun kecil yang berada di dalam web yang dapat disimpan di penyimpanan lokal dan data diambil sesuai dengan kata kunci yang ditentukan. Tahapan ini menggunakan Application Programming Interface (API) yang sudah dipersiapkan oleh pihak twitter. API Twitter dapat diakses dengan melakukan pengajuan otentikasi. Twitter menggunakan Open Authentication (OAuth) dan setiap permintaan perlu diajukan oleh pengguna Twitter yang telah resmi terdaftar. Proses crawling data pada riset ini dilakukan kurang lebih dalam 5 bulan antara bulan Maret 2021 hingga Agustus 2021.

Pada tahapan pengumpulan data akan didapatkan dataset yang berisi data campuran yang belum memiliki kelas/label. Dalam proses pembuatan dataset diperlukan proses klasfikasi. Klasifikasi merupakan proses membangun model untuk menentukan kelas atau konsep dari data. Klasifikasi mengelompokkan data menjadi beberapa kelas/label sesuai dengan yang dibutuhkan. Pada riset ini, data tweet yang terkumpul belum memiliki kelas/label sehingga dilakukan pelabelan secara manual (Labeling). Proses pelabelan pada dataset ini menggunakan dua kelas/label yaitu positif dan negatif. Proses pelabelan untuk tiap 1 kelas tweet melibatkan 3 orang lainnya dan diambil menggunakan majority vote atau suara terbanyak. Contoh dari pelabelan dataset dijelaskan di Tabel 1.

Tabel 1. Contoh Data

Tweet	**Kelas/ Label**
Produk bagus, keren RTO	positif
Barang yang dikirm tidak sesuai pesanan, dan pakcing rusak	negatif

ISSN: 2477-3042 SURYA INFORMATIKA, VOL. 11 No.1, November 202

3.3 Pre-processing Data

Data hasil crawling merupakan data yang masih mentah sehingga perlu dilakukan tahapan untuk membersihkan data dari informasi yang tidak diperlukan. Tahapan untuk membersihkan data yaitu pre- processing. Tahapan ini bertujuan agar data menjadi bersih dan dapat menghasilkan hasil yang lebih optimal ketika dilakukan pelatihan model analisis sentimen. Berikut merupakan langkah-langkah dalam pre- processing:

1.) Data Cleaning

Data cleaning adalah suatu proses yang dilakukan untuk membersihkan data yang diterapkan untuk menghilangkan noise dan memperbaiki inkonsistensi dalam data. Data dalam dunia nyata cenderung berisi noise dan tidak konsisten. Pada data cleaning tweet, dilakukan penghapusan URL, angka, simbol, dan atribut yang mengandung missing value atau kosong.

2.) Stop Words

Stop words merupakan proses yang dilakukan untuk penghapusan kata pada tweet yang mengandung kata yang dianggap tak berpengaruh penting dalam menentukan klasifikasi seperti konjungsi. Kata-kata tersebut dimasukkan ke dalam daftar stop words, maka kata tersebut akan dihapus dari tweet.

3.) Stemming

Stemming merupakan proses menghilangkan kata dan mengubah kata menjadi kata dasar. Proses ini dapat dilakukan dengan cara menghapus awalan atau akhiran (imbuhan) dari sebuah kata.

4.) Case Folding

Case folding ialah metode mengoversikan huruf kapital pada tweet menjadi huruf kecil. Proses ini dilakukan agar semua huruf dalam data input menjadi seragam dan mempermudah proses klasifikasi.

5.) Tokenizing

Tokenizing adalah sebuah tahapan memisahkan kata-kata yang dipisahkan oleh spasi. tahapan ini dilakukan untuk mempermudah klasifikasi

4. Skenario dan Hasil Pengujian

Pada riset ini, digunakan 3 skenario pengujian dengan algoritma klasifikasi Support Vector Machine. Skenario yang pertama adalah melakukan pengujian dengan memanfaatkan metode feature selection Principal Component Analysis (PCA) yang digunakan sebagai baseline. Skenario yang kedua adalah melakukan pengujian dengan memasukkan proses pembobotan TF-IDF untuk mengetahui pengaruh dengan adanya pembobotan TF-IDF. Skenario yang ketiga adalah dengan menambahkan feature expansion untuk mengetahui pengaruh dari adanya penambahan proses feature expansion.

Tabel 2. Hasil Performansi Baseline

Rasio	Akurasi (%)	F1-Score
70:30	75,78%	0,7584
80:20	75,42%	0,7524
90:10	**75,90%**	**0,7561**

Berdasarkan hasil yang didapatkan pada tabel 2, didapatkan hasil paling optimal pada rasio pembagian data latih dan data uji dengan rasio 90:10 dengan nilai akurasi sebesar 75,90%, sehingga untuk pengujian selanjutnya akan menggunakan rasio pembagian data latih dan data uji 90:10.

Tabel 3. Hasil Performansi Baseline + TF-IDF

Mode l	Akurasi (%)	F1-Score
Baseline	75,90%	0,7561
Baseline + TF-IDF	**78,37% (+3,25)**	**0,7818 (+3,40)**

Pada skenario kedua didapatkan hasil seperti pada tabel 3, pembobotan TF-IDF meningkatkan akurasi dari algoritma klasifikasi sebesar 3,25% dari baseline menjadi 78,37%, serta meningkatkan nilai f1-score sebesar 3,40% menjadi 0,7818.

Tabel 4. Hasil Performansi Baseline + TF-IDF + Word2Vec (TOP 1)

Random State	Akurasi (%)	F1-Score
1	78,37% (+3,25)	0,7818 (+3,40)
24	77,95% (+2,70)	0,7779 (+2,88)
38	78,73% (+3,73)	0,7867 (+4,05)
42	79,40% (+4,61)	0,7927 (+4,84)
54	**79,52% (+4,77)**	**0,7942 (+5,04)**

Pada skenario ketiga dengan fitur TOP 1 didapatkan hasil akurasi yang meningkat hingga 4,77% dari 75,90% menjadi 79,52%, serta meningkatkan nilai f1-score sebesar 5,04% menjadi 0,7942.

Tabel 5. Hasil Performansi Baseline + TF-IDF + Word2Vec

Random State	Akurasi (%)	F1-Score
1	78,25% (+3,10)	0,7811 (+3,31)

ISSN: 2477-3042 SURYA INFORMATIKA, VOL. 11 No.1, November 202

24	78,19% (+3,02)	0,7798 (+3,13)
38	78,67% (+3,65)	0,7862 (+3,98)
42	**79,34% (+4,53)**	**0,79 (+4,48)**
54	79,04% (+4,14)	0,7886 (+4,30)

4.7 Analisis Hasil Pengujian

Berdasarkan dari hasil pengujian diatas, pengujian meggunakan feature expansion dapat menghasilkan hasil yang berbeda-beda tergantung daripada penggunaan corpus kata dan ukuran fitur yang digunakan. Dari pengujian yang telah dilakukan, proses pembobotan TF-IDF dan feature expansion dapat meningkatkan akurasi dengan cukup baik dibandingkan dari tanpa menggunakan pembobotan TF-IDF dan feature expansion. Nilai optimum akurasi yang didapakan dari pengujian yang dilakukan adalah 79,52% serta nilai f1-score mencapai 0,7942 pada fitur top 1 dan random state 54. Hal tersebut dikarenakan fitur top 1 akan lebih spesifik dalam mencari similaritas dari kata.

5. Kesimpulan

Pada riset ini, dilakukan analisis sentimen menggunakan feature expansion Word2Vec dan algoritma klasifikasi Support Vector Machine. Pada riset ini, diketahui bahwa dalam membangun sistem analisis sentimen terhadap data kebijakan pemerintah dibutuhkan beberapa tahap mulai dari crawling data, pre-processing, pembobotan TF-IDF, pembuatan corpus, feature expansion Word2Vec, serta proses klasifikasi. Berdasarkan hasil riset diatas juga dapat disimpulkan bahwa penggunaan feature expansion Word2Vec terbukti dapat meningkatkan akurasi dari algoritma klasifikasi dibandingkan dengan tanpa menggunakan feature expansion dengan peningkatan akurasi sebesar 4,77%. Dari pengujian diatas didapatkan juga hasil paling optimal untuk akurasi pada algoritma klasifikasi Support Vector Machine pada fitur Top 1 dengan nilai akurasi sebesar 79,52% dan nilai F1-Score sebesar 0,7942.

Daftar Pustaka

[1] S. Informatika, “MANAJEMEN RTO,” vol. 10, no. 1, pp. 10–12, 2021.

[2] E. Subowo, E. Sediyono, and Farikhin, “Ant Colony Algorithm for Determining Dynamic Travel Routes Based on Traffic Information from Twitter,” *E3S Web Conf.*, vol. 125, no. 201 9, 2019, doi: 10.1051/e3sconf/201912523012.

[3] M. G. Ozsoy, “From Word Embeddings to Item Recommendation,” 2016.

[4] S. B. Kaleel and A. Abhari, “Cluster-discovery of Twitter messages for event detection and trending,” *J. Comput. Sci.*, vol. 6, pp. 47–57, 2015, doi: 10.1016/j.jocs.2014.11.004.

[5] D. C. Wintaka, M. A. Bijaksana, and I. Asror, “Named-entity recognition on Indonesian tweets using bidirectional LSTM-CRF,” *Procedia Comput. Sci.*, vol. 157, pp. 221–228, 2019, doi: 10.1016/j.procs.2019.08.161.

[6] A. Taufik, “Komparasi Algoritma Text Mining Untuk Klasifikasi Review Hotel,” *J. Tek. Komput. AMIK BSI*, vol. IV, no. 2, pp. 69–74, 2018, doi: 10.31294/jtk.v4i2.3461.

[7] E. Subowo, I. Rosyadi, and H. H. Kusumawardhani, “Twitter Data as Decision Tree Parameter for Analysis of Tourism Potential Policies,” vol. 436, pp. 474–478, 2020, doi: 10.2991/assehr.k.200529.099.