

Analisis Metode Wordtovec Pada Performa Klasifikasi Kategori Berita Menggunakan Metode SVM dan KNN

Dwi Utami¹, Kusri², Sudarmawan³
Universitas Amikom Yogyakarta^{1,2,3}

dwiutami.1264@students.amikom.ac.id, kusri@amikom.ac.id, sudarmawan@amikom.ac.id

Abstrak – Perkembangan teknologi saat ini meningkatkan kemudahan dalam mendapatkan berbagai informasi salah satunya adalah berita dari media internet. Berita yang ada di internet selain dapat memberikan informasi kepada masyarakat juga dapat digunakan sebagai bahan pembelajaran dan analisa seperti dapat digunakan untuk klasifikasi kategori berita. Terdapat tahapan sebelum melakukan klasifikasi yaitu *preprocessing* yang merupakan salah satu tahapan terpenting dalam klasifikasi. Penelitian ini bertujuan untuk melihat pengaruh proses *stopword* pada algoritma *wordtovec*, serta melihat performasinya pada klasifikasi yang dilakukan komparasi algoritma SVM dan KNN untuk melihat metode yang menghasilkan nilai akurasi tertinggi. Hasil penelitian menunjukkan bahwa metode kombinasi yang menghasilkan akurasi tertinggi adalah metode *wordtovec* dengan SVM dengan akurasi sebesar 92.89%. Proses *stopword* sedikit meningkatkan akurasi sebesar 1.55%.

Kata kunci: Klasifikasi; Wordtovec; SVM; KNN;

Abstract - Current technological developments increase the ease of obtaining various information, one of which is news from the internet media. News on the internet, apart from being able to provide information to the public, can also be used as learning material and analysis, as can be used to classify news categories. There are stages before doing the classification, namely *preprocessing* which is one of the most important stages in classification. This study aims to see the effect of the *stopword* process on the *wordtovec* algorithm, as well as to see its performance on the classification carried out by comparing the SVM and KNN algorithms to see the method that produces the highest accuracy value. The results showed that the combination method that produces the highest accuracy is the *wordtovec* method with SVM with an accuracy of 92.89%. The *stopword* process slightly increases the accuracy by 1.55%.

Keyword : Classification; Wordtovec; SVM; KNN;

1. Latar Belakang

Perkembangan teknologi saat ini meningkatkan kemudahan dalam mendapatkan informasi, salah satunya adalah informasi berbagai macam berita melalui media internet. Banyaknya berita yang ada di internet selain dapat memberikan informasi kepada masyarakat juga dapat dimanfaatkan sebagai bahan analisa dan pembelajaran seperti dapat di gunakan untuk proses klasifikasi berita. Klasifikasi teks adalah upaya untuk mengatur dokumen teks ke dalam kategori berdasarkan properti dan atribut yang dimiliki setiap teks. Tugas dalam klasifikasi adalah sebagai identifikasi kategori dokumen baru berdasarkan probabilitas yang disarankan oleh korpus pelatihan tertentu dari dokumen yang sudah berlabel atau teridentifikasi. Klasifikasi teks digunakan di beberapa domain, salah satunya adalah kategorisasi berita. Secara konsep terlihat sederhana, dengan sedikit dokumen, dimungkinkan untuk menganalisis setiap dokumen dan didapatkan gambaran tentang

kategori di mana dokumen itu berada. Berdasarkan pengetahuan ini, dimungkinkan untuk pengelompokan dokumen serupa ke dalam kategori atau kelas.

Tantangan dalam klasifikasi dokumen salah satunya adalah belum ada metode yang optimal pada *preprocessing* teks dokumen untuk mengubah data tidak terstruktur menjadi terstruktur [1]. *Preprocessing* teks merupakan hal penting yang harus dilakukan sebelum proses klasifikasi, sehingga diperlukan penelitian lanjutan untuk memilih metode *preprocessing* yang lebih baik. Dalam penelitian ini akan dilakukan analisis metode *Wordtovec* serta melihat pengaruh proses *stopword* pada algoritma *Wordtovec* dengan melihat performansinya terhadap hasil klasifikasi. Proses klasifikasi akan dilakukan komparasi algoritma SVM dan KNN untuk melihat metode yang menghasilkan nilai akurasi tertinggi. Berikut merupakan beberapa penelitian yang mendasari pemilihan metode dalam penelitian:

Analisis Metode Wordtovec Pada Performa Klasifikasi Kategori Berita Menggunakan Metode SVM dan KNN

Penelitian terdahulu dilakukan oleh Rosiello, melakukan penelitian dengan menggunakan metode *Wordtovec* dalam mempelajari hubungan semantik antar kata, kemudian di klasterkan menjadi beberapa *centroid* dan terkumpul beberapa kata berdasarkan kemiripan makna yang dianggap sama. Penelitian tersebut menghasilkan kinerja yang lebih baik dibandingkan dengan model pembelajaran yang lebih kompleks [2].

Penelitian lain dilakukan oleh Maulina dan Sagara melakukan penelitian tentang klasifikasi artikel hoax menggunakan Tf-Idf dan *Support Vecto Machine*. Penggunaan *support vector machine* dengan pembobotan Tf-Idf terbukti mencapai kinerja yang sangat baik yaitu 95%. Sehingga dalam penelitian ini menggunakan metode klasifikasi dengan SVM [3].

Rozi dan Sulistyawati meneliti tentang klasifikasi berita hoax dengan menggunakan pembobotan Tf-Idf dan metode klasifikasi dengan KNN. Kinerja pembobotan dengan KNN menghasilkan kinerja yang sangat baik yaitu mencapai 92%[4].

Budiman melakukan penelitian tentang ekstraksi fitur menggunakan *Wordtovec* pada pembentukan fitur berdimensi rendah. Pada penelitian tersebut juga menggunakan algoritma klasifikasi SVM dan menghasilkan akurasi kinerja yang baik hingga 87%[5].

Berdasarkan studi literatur di atas penelitian ini berfokus pada analisis preprocessing yaitu *Wordtovec* dan melihat pengaruhnya terhadap klasifikasi kategori berita news, ekonomi dan kesehatan. Klasifikasi kategori berita dilakukan dengan menggunakan komparasi dua metode yaitu SVM dan KNN.

2. Kajian Pustaka

Asiyah dan Fithriasari melakukan penelitian tentang perbandingan hasil klasifikasi berita online dengan menggunakan algoritma KNN dan SVM dengan metode *preprocesing* Tf-Idf. Pengujian algoritma pada klasifikasi dengan metode *preprocessing* Tf-idf menunjukkan perbedaan yang signifikan sehingga diperlukan analisa metode *preprocesing* dan klasifikasi yang berbeda untuk mendapatkan performa terbaik dan teruji. Perbedaan algoritma pada preprocessing yaitu *Word2vec*. Perbedaan lain juga perbandingan dengan melihat pengaruh *stopword* pada metode *wordtovec* [6].

Irfa, Adiwijaya dan Mubarak melakukan penelitian tentang klasifikasi topik-topik berita Bahasa Indonesia dengan menggunakan metode KNN. Kelemahan pada penelitian tersebut pada dataset yang hanya 360 sehingga perlu dilakukan penelitian lanjutan

dengan penambahan dataset, akurasi yang dihasilkan cukup rendah hanya 69.9%. Perbedaan dengan penelitian yang akan dilakukan dengan melihat pengaruh *stopword* pada metode *Wordtovec* serta melihat performa terhadap proses klasifikasi dengan menggunakan metode SVM dan KNN [7].

Santoso melakukan penelitian pengklasifikasian berita menggunakan algoritma *Self-Training Naive Bayes*. Metode pemilihan fitur yang digunakan dalam klasifikasi teks tersebut adalah model *Wordtovec Skip-Gram*. Perbedaan pada penelitian yang akan dilakukan adalah menggunakan metode *wordtovec* dengan melihat performa pada klasifikasi menggunakan metode SVM dan KNN serta melihat pengaruh proses *stopword* terhadap metode *wordtovec* [8].

3. Metode Penelitian

a. Sumber Data

Sumber data pada penelitian ini adalah data berita online dari situs *kompas.com*. Data berita yang digunakan terdiri dari tiga kategori yaitu news, kesehatan, dan keuangan. Data tiap kategori diambil 300 artikel sehingga data keseluruhan adalah 900 artikel berita.

b. Metode Analisis

Berikut merupakan langkah analisis data pada penelitian ini, yang di tunjukan pada gambar 1:



Gambar 1 Langkah Analisis

Pada gambar 1 menunjukan langkah-langkah analisis data yang dijelaskan sebagai berikut:

1. Menyiapkan data berupa artikel berita dengan 3 kategori yaitu news, kesehatan, keuangan.
2. Melakukan preprocessing teks yaitu stopwords. Pada penelitian ini preprocessing yang dilakukan hanya stopwords untuk mengetahui pengaruhnya pada saat pengujian.
3. Pembentukan vektor dilakukan dengan metode wordtovec.
4. Pengujian dengan menggunakan metode klasifikasi yaitu SVM dan KNN sebagai perbandingan.
5. Menganalisis performa klasifikasi berdasarkan nilai akurasi yang dihasilkan.
6. Mengevaluasi hasil dari perbandingan metode yang menghasilkan nilai akurasi tertinggi.
7. Menarik kesimpulan dari hasil penelitian

4. Implementasi Sistem dan Hasil

a. Support Vector Machine

Berikut merupakan hasil akurasi menggunakan metode wordtovec pada klasifikasi menggunakan SVM menggunakan kernel linear, dan $c=1.0$ dengan kfold cross validation sebesar 5 menghasilkan akurasi dan f-measure yang ditunjukkan pada tabel 1 sebagai berikut:

Tabel 1 hasil dengan SVM

	akurasi	f-measure
SVM	94.44%	94.44%

Pada tabel 1 hasil akurasi dengan metode wordtovec dengan algoritma klasifikasi SVM menghasilkan akurasi 94.44% dan f-measure sebesar 94.44%.

b. K-Nearest Neighbour

Berikut merupakan hasil akurasi menggunakan metode wordtovec pada klasifikasi menggunakan KNN dengan $K=2,3,5$ menggunakan kfold cross validation sebesar 5 menghasilkan akurasi dan f-measure yang ditunjukkan pada tabel 2 sebagai berikut:

Tabel 2 Hasil dengan KNN

	akurasi	f-measure
KNN k=2	91.33%	91.23%
KNN k=3	91.99%	92.02%
KNN k=5	92.44%	92.46%

Pada tabel 2 hasil akurasi dengan metode wordtovec dengan akurasi tertinggi di dapatkan dengan $K=5$ dengan nilai akurasi 92.44% dan f-measure 92.46%

Hasil perbandingan kinerja wordtovec dengan algoritma SVM dan KNN, serta pengaruh stopwords ditunjukkan pada tabel 3 sebagai berikut:

Tabel 3 Perbandingan SVM dan KNN

	Tanpa stopwords	Stopword
Wordtovec - SVM	92.89%	94.44%
Wordtovec - KNN	92.44%	92.02%

5. Kesimpulan

Berdasarkan hasil penelitian, metode yang menghasilkan akurasi tertinggi pada klasifikasi kategori berita news, kesehatan, dan keuangan adalah wordtovec dan SVM dengan nilai akurasi keseluruhan 92.89%% sedangkan untuk pengaruh stopwords meningkatkan akurasi 1.55%.

6. Pustaka

- [1] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny)*, vol. 477, pp. 15–29, 2019, doi: 10.1016/j.ins.2018.10.006.
- [2] G. Rossiello, P. Basile, and G. Semeraro, "Centroid-based Text Summarization through Compositionality of Word Embeddings," 2017, doi: 10.18653/v1/w17-1003.
- [3] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *J. Mantik Penusa*, 2018.
- [4] F. N. Rozi and D. H. Sulistyawati, "KLASIFIKASI BERITA HOAX PILPRES MENGGUNAKAN METODE MODIFIED K-NEAREST NEIGHBOR DAN PEMBOBOTAN MENGGUNAKAN TF-IDF," *KONVERGENSI*, 2019, doi: 10.30996/konv.v15i1.2828.
- [5] irwan budiman, M. R. Faisal, and D. T. Nugrahadi, "Studi Ekstraksi Fitur Berbasis Vektor Word2Vec pada Pembentukan Fitur Berdimensi Rendah," *J. Komputasi*, 2020, doi: 10.23960/komputasi.v8i1.2517.
- [6] S. N. Asiyah and K. Fithriasari, "Klasifikasi Berita Online Menggunakan Metode Support Vector Machine Dan K-

- Nearest Neighbor,” *J. Sains dan Seni ITS*, 2016.
- [7] A. A. Irfa, Adiwijaya, and M. S. Mubarak, “Klasifikasi Topik Berita Berbahasa Indonesia Menggunakan k-Nearest Neighbor,” *e-Proceeding Eng.*, vol. 5, no. 2, p. 3631, 2018.
- [8] J. Santoso *et al.*, “Self-Training Naive Bayes Berbasis Word2Vec untuk Kategorisasi Berita Bahasa Indonesia,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 2, pp. 158–166, 2018, doi: 10.22146/jnteti.v7i2.418.