BAB II
TINJAUAN PUSTAKA

Pada penelitian ini, permasalahan yang diangkat bukanlah hal yang baru. Terdapat beberapa penelitian terdahulu yang memiliki topik hampir sama dengan penelitian ini. Dari penelitian yang ditemukan maka penulis mengambil beberapa yang dapat digunakan sebagai acuan agar penelitan ini menghasilkan karya yang lebih baik.

Pada penelitian berjudul "Penerapan Metode *Clustering Text Mining* Untuk Pengelompokan Berita Pada *Unstructured Textual Data*" oleh Nyoman Gede Yudiarta, Made Sudarma, dan Wayan Gede Ariastina memiliki kesamaan yaitu mengimplementasikan *clustering* dalam penelitian. *Dataset* yang digunakan pada penelitian tersebut, menggunakan data judul berita dengan jumlah berbeda yaitu 50, 100, 200, 300, 400, dan 500 data. Dalam proses *clustering*, penelitian tersebut menggunakan algoritma *K-means*. Hasil yang diperoleh dalam penerapan *K-means* dalam *clustering* mampu memberikan hasil yang memuaskan, dengan rata-rata *Precision* sebesar 70,76% sedangkan *Recall* sebesar 70,86% serta *Purity* sebesar 0,76 untuk semua data uji[5].

Berbeda dengan sebelumnya, penelitian berjudul "Implementasi *Text Mining* Dalam Pengelompokkan Data Tweet Pertanian Indonesia Dengan *K-means*" oleh Hafiz Irsyad dan M Rizky Pribadi, memiliki persamaan yang ada adalah penelitian yang bertujuan untuk mengelompokkan data. *Dataset* yang digunakan adalah data yang diambil dari twitter. Data tersebut adalah tweet para pengguna twitter yang diambil dalam rentan waktu januari 2019 hingga oktober 2019 dengan jumlah 350 tweet. Penelitian tersebut menggunakan menerapkan algoritma *K-means* dalam pengelompokkan data. Hasil yang diperoleh dalam penerapan algoritma *K-means* terdapat 5 *cluster* yang telah diperoleh yaitu Pangan, Produksi, Lahan, Ekspor dan Teknologi. Dari 5 *cluster* yang diperoleh pengukuran rata-rata akurasi *cluster* menggunakan %performace operator menunjukkan bahwa rata-rata dari *cluster* tersebut adalah 0.334%[6].

Pada penelitian berjudul "Perbandingan Pengelompokan *K-means* dan *K-medoids* Pada Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas" oleh Athifaturrofifah, Rito Goejantoro, dan Desi Yuniarti. *Dataset* yang digunakan merupakan data potensi kebakaran hutan dan lahan. Penelitian tersebut mempunyai kesamaan dengan eksperimen ini yaitu terletak pada metode *clustering* yang menggunakan metode *K-means*. Dalam penelitian ini membahas dua perbandingan yaitu metode *K-means* dan *K-medoids*. Hasil dari penelitian tersebut menunjukan bahwa hasil menggunakan metode *K-means* pada proses *cluster* lebih baik daripada menggunakan metode *K-medoids*. Hasil tersebut didukung oleh nilai *silhouette score* global yang dimana *K-means* 0.558 sedangkan metode *K-medoids* nilai *silhouette score* global adalah 0.529. Hasil dari penelitian tersebut metode *K-means* memberikan hasil pengelompokkan yang lebih baik[7].

Pada penelitian yang lain yang berjudul "*Social Media Mining to Study Social User Group by Visualizing* Tweet *Clusters using Word2vec*, *PCA and K-means*" yang dilakukan oleh Md Morshed Jaman Adnana, Matthias L. Hemmje, and Michael A. Kaufmannc yang diterbitkan pada tahun 2021. *Dataset* yang digunakan merupakan *dataset* dari twitter. Pada penelitian tersebut melakukan *clustering* menggunakan metode *K-means*. Dalam penelitian tersebut, mempunyai kesamaan dengan eksperimen ini yaitu pada proses *word embedding* menggunakan metode *Word2vec* dan proses *clustering* yang menggunakan metode *K-means*. Pada penelitian tersebut metode *K-means* dikombinasikan dengan teknik *PCA*. Teknik tersebut mengubah vektor 100-D menjadi 2-D. Hasil dari penelitian tersebut dapat dikatakan bahwa metode *K-means* yang dikombinasikan dengan teknik *PCA* lebih baik dari pada tidak dikombinasikan. Hasil tersebut didukung oleh nilai *silhouette score* yang dimana *K-means* mempunyai nilai 0.0641 sedangkan *K-means* yang dikombinasi dengan teknik *PCA* yaitu 0.3619[8].

Pada penelitian kedua oleh Song, Wan-Li yang berjudul "*Sentence Similarity Calculating Method Based on Word2vec and Clustering*" yang diterbitkan pada tahun 2020. Pada penelitian ini mengalkulasi kemiripan teks dengan menggunakan metode *Word2vec* dan *clustering*. Data yang digunakan pada penelitian ini adalah data teks. Penelitian tersebut mempunyai kesamaan dengan eksperimen ini yaitu

memproses data menggunakan metode *Word2vec* dan *Word2vec* yang ditambahkan metode *clustering K-means*. Hasil dari penelitian ini menunjukkan bahwa metode *Word2vec* yang ditambahkan *K-means* lebih baik daripada hanya menggunakan metode *Word2vec*. Hasil tersebut berdasarkan nilai *silhouette score* yang diketahui bahwa metode *Word2vec K-means* memiliki 1-20% hasil yang lebih baik dari pada hanya menggunakan *Word2vec*[9].

UNIVERSITAS ATMA JAYA YOGYAKARTA