Abstrak

Sebagai bahasa yang kaya dengan kosa kata, bahasa Indonesia memiliki banyak kata yang berbeda namun memiliki arti sama (sinonim). Hal ini dapat menyebabkan banyak berita masuk ke dalam kelompok atau kategori yang tidak relevan dengan isi beritanya. Untuk itu diperlukan cara untuk mengolah data untuk mendapatkan manfaat dari data itu, salah satu cara untuk mengolah berita tersebut yaitu *data mining*.Dalam data mining terdapat salah satu metode yang sering digunakan yaitu *clustering*. *Clustering* merupakan pengelompokkan objek berdasarkan karakteristiknya. Pengelompokkan berita dapat menggunakan metode clustering dengan tujuan untuk mengelompokkan artikel berita sesuai dengan topik beritanya.

Dalam tugas akhir ini mengimplementasikan suatu metode *clustering*, yaitu algoritma *Clustering based on Frequent Word Sequences* (CFWS) pada artikel berita berbahasa Indonesia. CFWS merupakan algoritma yang mempresentasikan dokumennya dengan menggunakan kata-kata yang paling sering muncul secara berurutan pada setiap dokumen. Dengan menggunakan algoritma ini dapat mengurangi dimensi dari setiap dokumen secara signifikan sehingga proses *clustering* menjadi lebih efisien. Pengujian dilakukan untuk melihat kualitas hasil *cluster* berdasarkan metode pengukuran akurasi *F-measure*.

Berdasarkan pengujian yang sudah dilakukan, algoritma CFWS dapat menghasilkan hasil kualitas hasil *cluster* yang baik. Selain itu algoritma CFWS dapat menghasilkan hasil *cluster* yang baik untuk dataset dengan topik yang berdekatan maupun topik yang sangat berbeda.

Kata Kunci : *data mining, clustering, CFWS, F-measure*

.