BAB V
KESIMPULAN DAN SARAN

Pada bab ini akan memaparkan mengenai kesimpulan yang didapatkan dari penelitian. Penulis juga akan menyebutkan hal apa saja atau saran yang dapat dilakukan untuk penelitian yang akan mendatang.

5.1 Kesimpulan

Setelah melakukan penelitian mengenai implementasi *K-Nearest Neighbor* dengan *cosine similarity* pada klasifikasi abstrak jurnal internasional ilmu komputer, maka penulis mendapatkan beberapa kesimpulan yang selaras dengan tujuan penelitian. Berikut kesimpulan yang didapatkan oleh penulis.

1. Penelitian ini berhasil membuat sebuah sistem klasifikasi untuk abstrak jurnal internasional ilmu komputer yang menggunakan *K-Nearest Neighbor* dengan *cosine similarity*. Tahapan dari perancangan sistem diantaranya pengumpulan data, *text preprocessing*, *feature selection* dengan pembobotan TF-IDF, menghitung kemiripan suatu dokumen dengan *cosine similarity* dan implementasi algoritma KNN dengan nilai *k* (jumlah tetangga) 3,5,7 dan 9.
2. *K-Nearest Neighbor* bergantung pada penggunaan jumlah kelas tetangga (k) yang berakibat pada performansi yang dihasilkan. Dalam penelitian ini semakin besar jumlah kelas tetangga (k) yang digunakan maka semakin tinggi performansi yang didapat. Hal tersebut dapat terlihat pada *holdout method* di skenario 2,3 dan 5 sedangkan pada metode 10 *fold cross validation* terdapat pada skenario 1,2,7 dan 10 yang terus mengalami peningkatan nilai performasi.
3. Penambahan jumlah *data training* mempengaruhi nilai *F1-Measure* yang didapat dan mempengaruhi seluruh *k* dalam peningkatan nilai *F1-measure*. Hasil akhir terbaik dari *f1-measure* yang didapatkan pada *holdout method* terdapat pada skenario ke 5 dengan 20% data training dan 80% data testing sebesar 63,91%. Sedangkan untuk metode 10 *fold cross validation* terlihat

pada skenario 1 dengan 90% data *training* dan 10% data *testing* sebesar 62,42%. Jika dilihat dari sistem klasifikasi teks semakin jumlah data training tinggi tentu akan berdampak pada kenaikan nilai f1-measure. Tetapi dari penelitian yang dilakukan menunjukan bahwa dengan bertambahnya jumlah data *training* tidak mempengaruhi terjadinya peningkatan nilai *f1-measure* pada metode *holdout method,* hal tersebut karena penyebaran id asbtrak yang tidak seimbang. Maka pada penelitian ini pemodelan *classifier* terbaik terdapat pada metode *splitting* atau pembagian data 10 *fold cross validation* dengan data *training* sebesar 90% yaitu 62,42%. Karena selain memiliki data *training* yang besar dalam pengujiannya pun penyebaran id asbtrak lebih merata.

4. Jika dilihat secara keseluruhan hasil perhitungan klasifikasi berdasarkan beberapa skneario yang dibuat menggunakan tabel *confusion matrix*, untuk kategori *Computer and Education* memiliki hasil yang lebih dominan dalam mengklasifikasikan data dengan benar dibandingkan dengan kategori *Computer and Security* dan *Computer in Human Behavior*. Dari hasil keseluruhan performansi sistem pada peningkatan nilai *k,* hasil performansi maksimal terdapat pada nilai $k = 9$.
5. Perbedaan hasil penelitian dengan penelitian sebelumnya sesuai dengan rancangan sistem yang dilakukan, dimana penelitian yang dilakukan oleh penulis menggunakan algoritma klasifikasi *supervised learning* dengan *K-Nearest Neighbor*. Sedangkan untuk penelitian sebelumnya menggunakan algoritma klasifikasi *unsupervised learning* dengan algoritma SentiStrength. Hasil menunjukan algoritma klasifikasi *supervised learning* dengan *K-Nearest Neighbor* dapat memberikan hasil performansi yang lebih maksimal dibandingkan dengan algoritma klasifikasi *unsupervised learning* dengan algoritma SentiStrength dalam pengkategorian dokumen teks berbahasa Inggris.

5.2 Saran

Dalam pelaksanaan penelitian penulis menyadari bahwa masih banyak kekurangan yang dilakukan oleh penulis di dalam penelitian ini. Oleh karena itu

penulis menyampaikan beberapa saran yang dapat dilakukan di kemudian hari agar penelitian selanjutnya dapat menghasilkan analisis yang lebih baik. Berikut beberapa saran yang penulis anjurkan.

1. Pada penggunaan metode *splitting data* dengan *holdout method* untuk pendistribusian *id* data *training* dan data *testing* dilakukan sesuai dengan urutan *id* data awal atau dengan kata lain dalam pendistribusiannya tidak dilakukan secara *random*. Hal tersebut agar diketahui perbedaan untuk pemodelan yang lebih optimal dengan penelitian yang penulis lakukan pada pendistribusian data *holdout method* secara random.
2. Saat melakukan tahapan *stopword removal* di dalam dokumen *dataset* selain menghapus *term* yang terdapat pada *stopword list*, diharapkan juga untuk menghapus *term* yang sering muncul atau yang mendominasi di dalam keseluruhan dokumen atau kategori kelas. Sehingga *term* tersebut tidak menjadi ciri suatu dokumen kelas dan tidak dimasukan ke dalam perhitungan klasifikasi. Hal tersebut dilakukan agar setiap kategori memiliki perbedaan yang lebih signifikan, dan tidak mendominasi terhadap satu kategori kelas. Proses tersebut juga akan mempercepat saat dilakukan tahapan pembobotan TF-IDF. Karena akan mengurangi jumlah *term* yang akan dihitung sehingga waktu prosesnya menjadi lebih efisien, dimana *term* yang dihitung hanya merupakan *term* inti dari setiap dokumen.