

BAB 6 PENGUJIAN DAN ANALISIS

Bab ini menerangkan tentang pengujian yang dilakukan terhadap sistem yang telah dibangun. Selain itu juga pada bab ini menjelaskan analisis dari hasil implementasi dan pengujian yang telah dilakukan.

6.1 Data yang digunakan

Penelitian ini dilakukan dengan berdasarkan data set yang diperoleh dari sumber penelitian sebelumnya yang dilakukan oleh Soekarno, Fauzi, Rizal(2016). Data set yang digunakan berasal dari *tweets* akun detik.com dan kompas, serta berita eksternal yang berasal dari *website* berita kompas dan detik.com. Pemilihan *tweets* setiap kategori diambil dari sumber tersebut, disesuaikan dengan kategori yang dipilih oleh peneliti untuk klasifikasi yang dilakukan. Jumlah data yang digunakan yaitu 105 dokumen berita, 100 data uji dan 400 data latih yang berupa *tweets*. Dokumen berita yang digunakan terbagi rata untuk 8 jenis kategori konten, sama halnya dengan data latih dan data uji yang masing-masing terbagi dari 8 kategori berita *tweets*.

6.2 Pengujian Nilai k Pada *K-Nearest Neighbor*

Pengujian hasil akurasi nilai k tetangga terbaik dari *K-Nearest Neighbor* dilakukan bertujuan untuk mengetahui pengaruh nilai k dan mendapatkan nilai k terbaik pada klasifikasi *tweets* menggunakan *KNN*.

6.2.1. Skenario Pengujian Nilai k Pada *K-Nearest Neighbor*

Pengujian variasi nilai k dilakukan dengan membandingkan akurasi pada beberapa variasi nilai k . Pengujian ini dilakukan tanpa menggunakan teknik ekspansi kata. Pengujian ini dilakukan dengan nilai k yang ganjil. Dalam hal ini, peneliti menggunakan nilai k pada *K-Nearest Neighbor* yang ke 1, 3, 5, 7, dan 9. Berikut merupakan hasil dari pengujian tersebut pada Tabel 6.1 dan pada Tabel 6.2 dan Tabel 6.3 merupakan contoh *tweets* yang ditentukan nilai k terbaiknya.

Tabel 6.1 Hasil Pengujian *K-Nearest Neighbor*

Nilai K	1	3	5	7	9
Akurasi	86%	87%	90%	86%	83%

6.2.2. Analisis Pengujian Nilai k Pada *K-Nearest Neighbor*

Hasil pengujian ini menunjukkan jika pada nilai k yang kecil diperoleh akurasi cukup baik. Namun, jika nilai k ditambahkan ke 3 dan ke 5 hasil pengujian menunjukkan hasil yang lebih baik. Akan tetapi, jika semakin besar nilai k hasil akurasi justru menurun. Hal ini karena semakin besar nilai k , maka semakin banyak tetangga yang tidak relevan. Dipilihnya batas nilai k hanya ke 9 karena dari tetangga terdekat ke 5 sudah diperoleh nilai tertinggi, dan nilai k setelahnya menunjukkan hasil yang terus menurun drastis. Nilai k ganjil dipilih untuk menghindari adanya hasil klasifikasi yang memiliki kesamaan jumlah jenis kategori jika digunakan nilai k genap. Hasil yang diperoleh dari pengujian tersebut menunjukkan bahwa nilai k terbaik yang diperoleh adalah $k=5$ dengan akurasi 90%. Pada Tabel 6.2 dan Tabel 6.3 menunjukkan contoh *tweets* yang dipilih untuk menghasilkan nilai k maksimal dari keseluruhan *tweets*. Gambar 6.1 menunjukkan grafik yang diperoleh dari pengujian nilai k pada *K-Nearest Neighbor*.

Tabel 6.2 Contoh Hasil Penentuan Dari Pemilihan Nilai k *Tweet* ekonomi

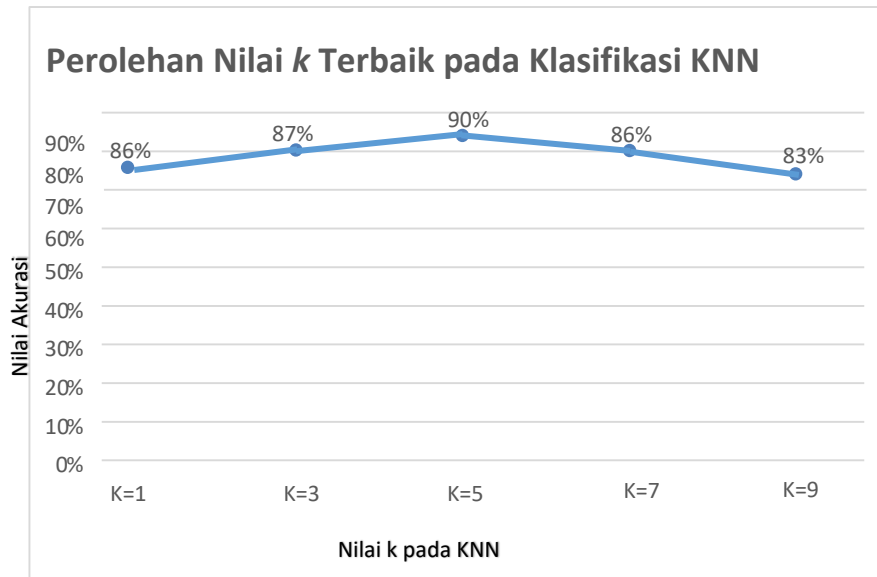
Tambah 23 pesawat, Garuda Indonesia Group proyeksikan pertumbuhan penumpang (Kategori Ekonomi)					
Nilai k	1	3	5	7	9
Urutan	Ekonomi	Ekonomi	Ekonomi	Ekonomi	Ekonomi
Rangking	-	Ekonomi	Ekonomi	Ekonomi	Ekonomi
dari	-	Ekonomi	Ekonomi	Ekonomi	Ekonomi
tertinggi	-	-	Ekonomi	Ekonomi	Ekonomi
(teratas)	-	-	Ekonomi	Ekonomi	Ekonomi
hingga	-	-	-	Teknologi	Teknologi
terendah	-	-	-	Teknologi	Teknologi
(kebawah)	-	-	-	-	Kesehatan
	-	-	-	-	Olahraga
Hasil Voting	Ekonomi	Ekonomi	Ekonomi	Ekonomi	Ekonomi

Melihat dari Tabel 6.2 disimpulkan bahwa nilai k ke 1 hingga ke 5 sesuai dengan hasilnya, karena semakin tinggi semakin baik juga karena mayoritas mempunyai kesamaan jenis. Namun ketika nilai k lebih tinggi dari nilai $k=5$, tetangga yang diperoleh mulai menunjukkan adanya kategori yang semakin tidak relevan.

Tabel 6.3 Contoh Hasil Penentuan Dari Pemilihan Nilai k Pada *Tweet* olahraga

1 Miliar Orang Tiap Hari Buka Facebook, 78 Persen via Mobile (Kategori Teknologi)					
Nilai k	1	3	5	7	9
Urutan	Ekonomi	Ekonomi	Ekonomi	Ekonomi	Ekonomi
Rangking	-	Ekonomi	Ekonomi	Ekonomi	Ekonomi
dari	-	Teknologi	Teknologi	Teknologi	Teknologi
tertinggi	-	-	Teknologi	Teknologi	Teknologi
(teratas)	-	-	Teknologi	Teknologi	Teknologi
hingga	-	-	-	Teknologi	Teknologi
terendah	-	-	-	Ekonomi	Ekonomi
(kebawah)	-	-	-	-	Ekonomi
	-	-	-	-	Ekonomi
Hasil Voting	Ekonomi	Ekonomi	Teknologi	Teknologi	Ekonomi

Berbeda dengan Tabel sebelumnya, melihat Tabel 6.3 memperlihatkan bahwa nilai k terkecil tidak selalu benar hasilnya, namun ketika hingga nilai $k=5$ menunjukkan hasil kebenarannya dikarenakan kategori mayoritas Teknologi terbanyak. Pada saat nilai k lebih tinggi hingga $k=9$ mulai menunjukkan hasil yang tidak sesuai, karena terlalu banyak hasil yang diperoleh, maka semakin banyak juga hasil yang tidak relevan.



Gambar 6.1. Grafik Perolehan Pengujian Nilai k Pada KNN

Gambar 6.1 merupakan nilai yang diperoleh dari pengujian yang telah dilakukan. Nilai $k=1$ menunjukkan akurasi 86%, namun sampai pada $k=5$ diperoleh hasil terbaik dengan akurasi 90%. Seperti yang dijelaskan pada analisis di atas, ketika semakin tinggi nilai k diperoleh hasil terbaik, namun jika tinggi nilai k terlalu tinggi maka semakin turun hasil akurasi yang disebabkan banyaknya hasil kategori tweets yang tidak relevan dengan hasil aslinya.

6.2 Pengujian Variasi *Threshold* nilai kedekatan pada *Query*

Expansion

Query expansion dilakukan dengan menambahkan beberapa kata dari informasi eksternal ke dalam data teks yang akan diklasifikasi. Kata-kata yang ditambahkan adalah kata-kata kedekatan secara semantik. Kedekatan semantik tersebut dapat dihitung dengan model semantik terdistribusi atau MST dan dimasukkan ke dalam sebuah kamus besar sebagai sumber pengetahuan eksternal untuk sistem. MST adalah sumber semantik eksternal yang dibangkitkan otomatis dan hal tersebut dilatarbelakangi dengan asumsi bahwa kata-kata yang secara semantik mempunyai arti yang sama akan muncul dalam konteks kata-kata yang sama. Kedekatan yang dimaksud adalah ketika terdapat kata pada data uji yang terdeteksi ada di *term* unik informasi eksternal. Dalam hal ini yang dipakai sebagai sumber informasi eksternal adalah data berita. Setelah ditemukan, maka

dihitung kedekatan antar kata yang ada di seluruh *term* unik berita. Perhitungan kedekatan kata tersebut menggunakan *Euclidean distance*. *Threshold* digunakan untuk memberi batas atas pada kata yang dipilih berdasarkan nilai kedekatannya. Kata yang nilai kedekatannya terhitung di bawah nilai *threshold* akan digunakan sebagai kata yang akan ditambahkan ke data teks uji agar lebih panjang. Tujuan dari *threshold* sendiri adalah untuk membatasi kata yang akan ditambahkan, jika semakin kecil nilai *threshold* yang ditemukan kedekatannya, maka akan semakin tinggi kedekatannya. Namun, apabila nilai *threshold* semakin tinggi akan semakin banyak kata yang terpilih dan kata yang tidak relevan semakin banyak.

6.2.1 Skenario Pengujian Variasi *Threshold* pada *Query Expansion*

Pengujian ini dilakukan dengan membandingkan akurasi pada beberapa variasi nilai *threshold*. Pengujian dilakukan dengan menggunakan nilai $k=5$ akurasi terbaik pada pengujian sebelumnya, yaitu $k=5$. Data uji yang digunakan berjumlah 100 data dan data latih yang terdiri 400 data dengan 8 kategori yang membawa 50 data pada setiap jenisnya. *Threshold* yang digunakan adalah 0,0, hingga 1,5 dengan data uji dan data latih yang sama. Tabel pengujian dapat dilihat pada Tabel 6.4.

Tabel 6.4 Hasil Pengujian *Query Expansion*

Data Uji	Data Latih	<i>Threshold</i> yang digunakan	Akurasi Menggunakan <i>Preprocessing</i> dan <i>Query Expansion</i>
100	400	0,0	90%
100	400	0,1	90%
100	400	0,2	90%
100	400	0,3	90%
100	400	0,4	90%
100	400	0,5	92%
100	400	0,6	92%
100	400	0,7	90%
100	400	0,8	90%
100	400	0,9	90%
100	400	1,0	90%
100	400	1,1	92%
100	400	1,2	92%
100	400	1,3	92%
100	400	1,4	92%
100	400	1,5	88%
100	400	2,0	86%

6.2.2 Analisis Pengujian Variasi *Threshold* pada *Query Expansion*

Melihat dari hasil akurasi asli yang diperoleh sebelumnya tanpa menggunakan *query expansion* adalah 90% dengan $k=5$. Setelah ditambahkan ekspansi kata, akurasi yang diperoleh naik menjadi 92% yang artinya metode ini mampu memengaruhi peningkatan akurasi. Akurasi hasil setelah dilakukan *query expansion* menjadi naik karena terdapat data teks yang telah ditambahkan kata baru yang memiliki relevansi kedekatan, agar lebih panjang jumlah teksnya. Pada akhirnya ketika dilakukan klasifikasi, kalimat atau *tweets* tersebut menjadi lebih mudah ditemukan relevansinya dengan data latih karena lebih panjang dan tidak ambigu ketika diproses. Tabel 6.5 merupakan contoh *tweets* yang dilakukan ekspansi *threshold* 0.6 .

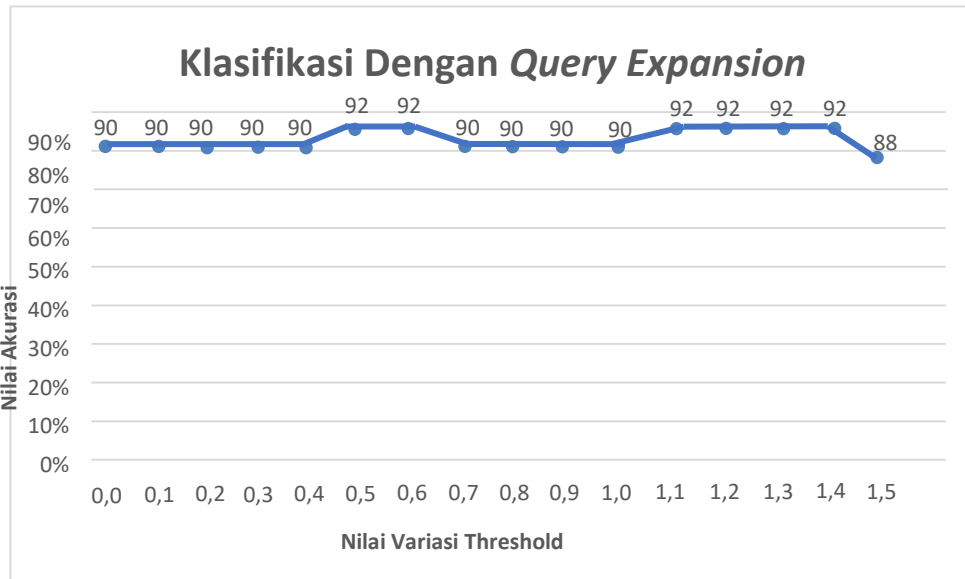
Tabel 6.5 Contoh Hasil *Tweets* yang Diekspansi

Doc	Isi	Keterangan	Kategori
1	Menpar Janji Promosikan Pariwisata untuk berlibur Banten	Sebelum Ekspansi Kata	Teknologi
2	menpar infrastruktur janji Promosikan pariwisata pantai wisata kebun hemat travel untuk berlibur pelni bunga Banten	Setelah Ekspansi Kata	Travel

Berdasarkan Tabel 6.5, menunjukkan penambahan atau ekspansi kata yang terjadi mampu membuat *tweets* tersebut menjadi sesuai kategori sebenarnya. Tulisan berwarna merah adalah kata yang terpilih dari perhitungan ekspansi kata yang terdapat pada dibawah *threshold* yang telah ditentukan. Pengaruh dari *threshold* tersebut mampu membantu ekspansi kata yang dilakukan untuk memiliki batas agar tidak seluruh kata terpilih dan dianggap sebagai kedekatan kata. Selain itu, apabila menggunakan ketika ditemukan kata yang di bawah batas nilai *threshold* saat batas nilai kecil, maka teks tersebut dianggap memiliki kedekatan secara semantik yang paling mirip. Namun ketika saat *threshold* semakin tinggi, maka akan semakin banyak pula kata yang diperoleh menjadi ekspansi kata, dan semakin banyak juga kata yang tidak memiliki hubungan kedekatan atau relevansi yang berakibat menurunnya tingkat akurasi saat proses klasifikasi teks. Berikut merupakan contoh ekspansi kata yang menggunakan *threshold* tinggi bernilai 1,5 pada Tabel 6.6.

Tabel 6.5 Contoh Ekspansi Kata Menggunakan *Threshold* Tinggi

Doc	Isi	Keterangan	Kategori
1	Menpar Janji Promosikan Pariwisata untuk berlibur Banten	Sebelum Ekspansi Kata	Teknologi
2	Menpar infrastruktur target akses jokowi dana aset jokowi janji Promosikan pariwisata pantai wisata kebun hemat travel murah ponsel indonesia untuk berlibur pelni bunga gunung bali eksotis bencana Banten	Setelah Ekspansi Kata	Ekonomi



Gambar 6.2 Grafik Perolehan nilai Pengujian Variasi *Threshold* Pada KNN

Gambar 6.2 menunjukkan hasil dari pengujian variasi *threshold* pada angka *threshold* 0,5 dengan 92%. Hasil yang diperoleh jika menggunakan *threshold* 0.0 hingga 1.5 menunjukkan perolehan nilainya yang naik dan juga turun. Pengujian *threshold* hanya sampai pada titik 1.5 karena mulai dari nilai *threshold* tersebut nilai akurasi menunjukkan grafik yang turun hingga pada titik ke 2.0.