Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 2, No. 10, Oktober 2018, hlm. 3849-3855 http://j-ptiik.ub.ac.id

Penerapan Klasifikasi Tweets Pada Berita Twitter Menggunakan Metode K-Nearest Neighbor Dan Query Expansion Berbasis Distributional Semantic

Galih Nuring Bagaskoro[1], M. Ali Fauzi[2],Putra Pandu Adikara[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]galihnuring2@gmail.com, [2]moch.ali.fauzi@ub.ac.id,[3]adikara.putra@ub.ac.id

Abstrak

Penggunaan teks pendek berbasis digital sampai saat ini masih berkembang dan meluas hingga diberbagai media sosial. Media sosial Twitter memiliki fitur kategori jenis informasi melalui *tweets* yang di unggah. Setiap pengelompokan jenis informasi dilakukan agar mempermudah pengguna untuk memanfaatkannya. Tujuan dari penggunaan kategori dalam hal ini klasifikasi, untuk mengevaluasi dan meningkatkan kualitas media sosial dalam pengelompokan kategori isi dari konten yang disediakan. Klasifikasi tradisional sampai saat ini masih digunakan, namun hasil yang diperoleh terkadang tidak maksimal, perlu dilakukan ekspansi kata untuk menambahkan kata kedalam teks agar dapat meningkatkan akurasi. Ekspansi kata digunakan dengan berbasis *distributional semantic* dengan teknik *euclidean distance* untuk menemukan kata terdekat dari sumber eksternal agar menjadi kueri yang akan ditambahkan ke teks data uji. Dengan menggunakan data uji 105 dan data latih 400, klasifikasi yang menggunakan *K-Nearest Neighbor* dapat memperoleh hasil 90% dengan tetangga terdekat K=5. Hasil tersebut sama halnya dengan hasil pengujian yang dilakukan dengan tanpa menggunakan teknik ekspansi kata. Sedangkan pengujian yang dilakukan dengan menambahkan ekspansi kata dengan *threshold* 0,5 dan nilai tertangga terdekat *K-Nearest Neighbor* K=5 memperoleh hasil akurasi 92%.

Kata kunci: *twitter, tweet, ekspansi kata, distributional semantic, euclidean distance, klasifikasi, k-nearest neighbor*

Abstract

The use of short text based on digital to date is still growing and extending to various social media. Twitter has news features in tweets to represent information representing each type. Each categorization of this type is done to make it easier for users to use it. The purpose of the use of categories in this classification, to evaluate and improve the quality of social media in grouping categories of content of the content provided. Traditional classification is still used today, but the results are sometimes not maximal, it is necessary to expand the word to add words to the text in order to improve the accuracy. Word expansion is used with a semantic-based distributional euclidean distance technique to find the closest word from an external source to be a query to be added to the test data text. Using test data 105 and training data 400, the classification using K-Nearest Neighbor can obtain 90% results with nearest neighbor K=5. These results are similar to the results of tests conducted without using word expansion techniques. While the test is done by adding the expansion of words with threshold 0.5 and the nearest immediate value K-Nearest Neighbor K=5 obtained an accuracy of 92%.

Keywords: *twitter, tweet, word expansion, distributional semantic, euclidean distance, classification, k-nearest neighbor*

1. PENDAHULUAN

Banyak pengguna internet pada dunia maya, bersamaan dengan pemilik akun Twitter di berbagai negara semakin meningkat dari tahun-ketahunnya. Twitter adalah salah satu dari berbagai jenis jejaring sosial yang membantu penggunanya untuk memperoleh berbagai informasi melalui post yang disebut kicauan (*tweets)* dengan jumlah maksimal karakter yaitu 280. Jika seseorang telah menjadi pengguna Twitter maka dapat memiliki hak akses untuk mendapatkan informasi *tweets* dari suatu akun tertentu,Pengguna harus menjadi pengikut (*follower*) akun tersebut terlebih dahulu (Kwak et al.,2010). Selain itu, apabila terdapat suatu

akun terpilih oleh twitter bahwasanya biasa memberikan info mengenai politik, namun sesekali akun tersebut mengomentari bahkan memberikan *tweets* tentang hiburan, maka oleh twitter masih dianggap informasi tentang politik.

Sampai saat ini terdapat banyak sekali *tweets* yang terus bermunculan dan tersebar oleh pengguna Twitter di seluruh negara. *Tweets* yang ada pada beranda Twitter tercampur menjadi satu dan tidak terkelompokkan berdasarkaan jenis beritanya, diantara lain yaitu olahraga, kesehatan, politik, ekonomi, teknologi , wisata dan lain sebagainya. Tidak adanya pengkategorian *tweets* membuat pengguna twitter kesulitan untuk membaca dan memilahnya berdasarkan informasi yang diinginkannya. Contohnya jika terdapat pengguna yang ingin mencari informasi tentang politik, maka pengguna tersebut harus mencari satu persatu *tweets* berita yang berkaitan dengan politik. Terkadang pengguna jika ingin mendapatkan jenis informasi yang sama, harus menjelajah dalam satu akun Twitter informasi atau berita yang memiliki konten satu jenis. Oleh karena itu, dibutuhkan sebuah sistem yang dapat melakukan klasifikasi tweets sesuai dengan kategorinya.

Teknik *Text Mining* di beberapa tahun terakhir menjadi sangat populer yang dilatarbelakangi semakin banyaknya jumlah teks digital yang luas dan tidak terstruktur, oleh karena itu perlu dilakukan analisis isi dari konten tersebut dengan cara yang fleksibel (Hearst, 1999). Salah satu metode pengklasifikasian teks yang sering digunakan adalah K-Nearest Neighbor. Penelitian yang dilakukan oleh Ramadhan dan Zeniarja (2016) menunjukkan bahwa K-Nearest Neighbor memiliki performa yang bagus dalam klasifikasi teks dengan akurasi 80%. Selain itu penelitian lainnya yang menggunakan K-Nearest Neighbor adalah Winda dan Rizal (2017) yang memiliki akurasi 82%. Dengan melihat hasil tersebut dapat dipahami bahwa penggunaan metode *K-Nearest Neighbor* memang baik dan bagus digunakan pada pengklasifikasian teks.

Metode klasifikasi teks secara umum masih memiliki kekurangan jika diterapkan pada *short-text* seperti Twitter (Liu dan Fan,2012). Kelemahan yang dimiliki oleh *short text* adalah adanya ambiguitas dan sedikitnya kata pada teks (Tang dan Wang.,2017). Penyebab dari ambiguitas adalah teks tersebut bersifat pendek dan berisi hanya beberapa kata yang mungkin akan ditemukan kesamaan kata dengan lebih dari satu kategori jika diklasifikasikan. Selain itu, *short text* hanya berisi sedikit kata yang terkadang jarang digunakan pada data latih, yang pada akhirnya sistem tidak dapat mengklasifikasikan kategori mana yang tepat. Teknik yang baik guna membantu permasalahan tersebut adalah dengan menambahkan fitur atau *query* baru atau yang disebut *query expansion*(QE). Caranya adalah mengoptimasi pesan atau teks singkat (*short text*) dengan menambahkan beberapa kata yang memiliki kesamaan atau kedekatan secara semantik. Dengan menggunakan *query expansion* maka perbendaharaan kata sebelumnya akan bertambah lebih banyak lagi.

Terdapat juga banyak penelitian yang menggunakan *query expansion.* Penggunaan ini cukup populer diterapkan pada penelitian. Penelitian yang dilakukan oleh Agung dan Ali (2016) menggunakan metode tersebut guna memperbaiki tingkat akurasi, dari hasil asli klasifikasi yang hanya 80% meningkat menjadi 82% setelah ditambahkan metode tersebut. Selain itu juga melalui penelitian Roi dan Ali (2016) dengan menggunakan metode tambahan *query expansion* dapat menghasilkan 96% dari yang sebelumnya 93% ketika masih belum ditambahkan metode tersebut. Dengan demikian metode tersebut terbukti mampu untuk menambahkan tingkat akurasi yang dilakukan pada hasil klasifikasi.

Salah satu cara melakukan *query expansion* adalah menggunakan informasi eksternal (*Unlabeled Background Knowledge*) seperti Wikipedia, Wordnet, dan dokumen berita (Zelikovitz dan Hirsh.,2000). *Query expansion* bisa 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 yang 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 (Yudi,2016).

Berdasarkan latar belakang tersebut, maka diusulkan penelitian yang menggunakan metode klasifikasi *K-Nearest Neighbor* dan *Query Expansion* berbasis *Distributional Semantic*

untuk memperoleh hasil lebih maksimal dari penelitian sebelumnya. Diharapkan penelitian ini dapat membantu pengguna media sosial membaca suatu konten berita sesuai kebutuhannya dengan lebih mudah.

2. TINJAUAN PUSTAKA

2.1 Preprocessing

Pada tahap *text mining* perlu adanya suatu proses yang dilalui agar dapat memperoleh informasi secara terstruktur dan jelas. Salah satunya adalah proses *preprocessing text.* Tahapan ini dilakukan untuk menyeleksi dan memfilter teks yang baik dan berguna dan siap untuk dianalisis (Hadna, et.al.,2016). Tahapan tersebut meliputi Case Folding, Tokenizing, Filtering, dan Stemming.

2.2 Ekspansi Kata(*Query Expansion*)

Metode ini merupakan salah satu teknik dasar pada *relevance feedback* di mana sistem akan menambahkan *query* tambahan pada pencarian pertama (Fachrudin,2011). Pada klasifikasi *short text* seringkali terjadi kendala pada kata yang muncul dalam dokumen *short text* sering tidak terdapatnya di dokumen latih. Selain itu terkadang muncul permasalahan yang baru seperti proses klasifikasi tidak berjalan dengan baik karena akan ada banyak kata yang tidak mampu terdeteksi masuk pada kategori yang mana. Proses pengelempokan dokumen pada kategori-kategori yang harus mempunyai kesamaan kata dengan dokumen latih.

Solusi dari permasalahan tersebut adalah mengunakan teknik *query expansion.* Teknik tersebut merupakan perluasan *query* dengan memformulasikan kembali *query* awal dengan melakukan penambahan kata untuk meningkatkan kinerja. Dengan kata lain, perbendaharaan kata yang ada di *dataset* semakin banyak sehingga dapat membantu kinerja pada klasifikasi untuk lebih baik lagi.

2.3 Distributional Semantic

Distributional semantic merupakan teori yang mempelajari tentang metode untuk mengukur dan mengkategorikan teks yang mempunyai makna yang sama berdasarkan sifat distribusinya dari data sampel kamus bahasa yang besar dan berasal dari eksternal. Dalam hal ini, dapat dimaksud dengan kombinasi dua kata berbeda namun memiliki arti maupun level yang sama(Nikolaos,2013). Terdapat salah satu parameter yang digunakan pada penelitian ini, yaitu *similarity measure* dengan menggunakan teknik yang dipakai adalah *euclidean distance*. Teknik ini berguna untuk menghitung jarak antar dokumen untuk mengetahui tingkat kemiripan berdasarkan nilai jarak yang terpendek.

2.3.1 Euclidean Distance

Metode ini adalah salah satu pendekatan yang bisa digunakan dalam membantu proses perhitungan jarak antar dokumen maupun teks yang terdapat pada *dataset* yang telah tersedia. *Euclidean Distance* adalah metrika yang paling sering digunakan untuk menghitung kesamaan dua vektor(titik). Rumus *euclidean distance* adalah akar dari kuadrat perbedaan dua titik data (Sendhy, 2014).

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})2}$$

Keterangan :

D_{ij} = Tingkat perbedaan (*dissimilarity degree*)

N = Jumlah vektor/titik

X_{ik}= Titik *input*

Xj_k= Titik pembanding

2.4 Klasifikasi Teks

Klasifikasi merupakan teknik bagaimana untuk mengkategorikan teks yang sesuai dengan karakteristiknya. Dengan terdapatnya teknik tersebut,dapat memberikan pandangan konseptual mengenai cara bagaimana untuk mengelompokkan dokumen yang mempunyai peran penting pada dunia nyata (Sriram eal, 2010). Dalam hal ini, penulis menggunakan metode K-Nearest Neighbor untuk teknik klasifikasi teksnya.

2.4.1 K-Nearest Neigbor

K-Nearest Neighbor adalah teknik pengelompokan data baru dengan menghitung jarak dari data baru dengan beberapa data terdekat(Santosa,2007:53). Dalam arti kata lain yaitu suatu pendekatan guna mencari permasalahan dengan menghitung jarak terdekat antara kasus baru dengan beberapa kasus yang lama, dan mencocokan berdasarkan bobot dari sejumlah data yang ada. Algoritma ini sangat mudah jika digunakan untuk teknik klasifikasi, yang sangat membantu unuk

mengklasifikasikan data baru berdasarkan atribut dan *training sample* yang selanjutnya dihasilkan titik *training* paling dekat dengan titik *querry*.

2.4.2 Cosine Similarity

Cosine Similarity merupakan metode yang digunakan sebagai perhitungan tingkat kemiripan atau kesamaan antar kedua buah teks maupun objek. Umumnya perhitungan dengan metode tersebut sebelumnya berdasarkan *vector space similarity measure*. Metode *cosine similarity* ini menghitung dua buah obyek atau dinyatakan pada dua buah *vector* dengan kata kunci (*keywords*) dari dokumen sebagai tolak ukur.

$$Cosine\ (d_i, q_i) = \frac{q_i . d_i}{|q_i||d_i|} = \frac{\sum_j^t =_1 (q_{ij} . d_{ij})}{\sqrt{\sum_{j=1}^t (q_{ij})^2} \cdot \sum_{j=1}^t (d_{ij})^2}$$

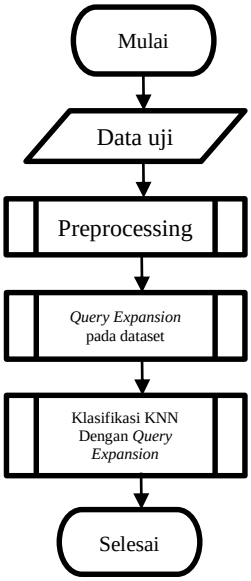
Keterangan :

q_{ij} = bobot istilah *j* pada dokumen *i*= $.idf_j$

d_{ij} = bobot istilah *j* pada dokumen *i*= $.idf_j$

3. METODOLOGI

Pada Klasifikasi tweets mengenai berita Twitter berbahasa indonesia memiliki beberapa tahap yang berurutan dan agar sesuai dengan hasil yang diperlukan.Berikut merupakan diagram alir secara umum yang digambarkan pada Gambar 1.



Gambar 1. Diagram Alir sistem

Pada Gambar 1 dijelaskan bahwa sistem yang akan dikembangkan memiliki tujuan untuk menyelesaikan permasalahan klasifikasi yang kurang maksimal pada kategori berita Twitter berbahasa Indonesia yang tersedia. Permasalahan tersebut diselesaikan dengan metode klasifikasi *K-Nearest Neighbor* yang sebelumnya dilakukan ekspansi kata yang memiliki nilai terdekat untuk dijadikan sebuah kueri yang nantinya akan ditambahkan ke dalam data uji. Proses tersebut diawali oleh proses *Preprocessing* , yaitu proses tambahan yang berguna untuk memperbaiki kata-kata yang akan digunakan. Setelah itu tahap ekspansi kata dilakukan kepada data uji dengan data set berita yang akan memproses agar kata data uji menjadi lebih panjang. Kemudian dilakukan pengklasifikasian dengan *K-Nearest Neigbor* dengan teknik *Cosine Similarity*. Hasil daripada akurasi klasifikasi yang diperoleh bergantung pada probabilitas frekuensi data latih, serta jumlah sumber berita eksternal yang dimiliki.

4. HASIL DAN PEMBAHASAN

4.1 Pengujian Nilai K Pada *K-Nearest Neighbor*

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

4.1.1 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 1.

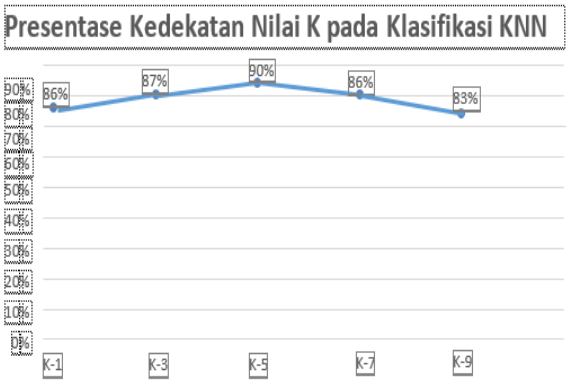
Tabel 1 Hasil Pengujian K-Nearest Neighbor

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

4.1.2 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%.



Gambar 2. Gambar grafik Pengujian nilai *k*

Gambar 2 merupakan nilai yang diperoleh dari pengujian yang elah 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.

4.2 Pengujian Variasi *Threshold* 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 kedekatanya. Kata yang nilai kedekatanya terhitung dibawah 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

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

Pengujian ini dilakukan dengan membandingkan akurasi pada beberapa variasi nilai Threshold. Pengujian dilakukan dengan menggunakan nilai K 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 2 berikut.

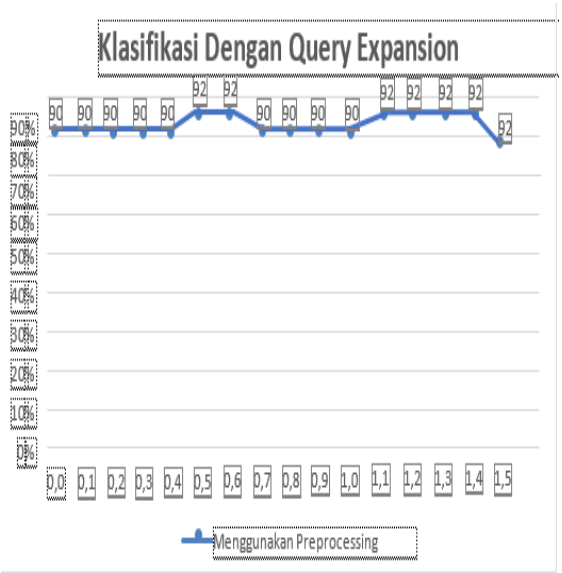
.Tabel 2. Pengujian Variasi *Threshold* Ekspansi Kata

Data Uji	Data Latih	*Threshold* yang digunakan	Akurasi Menggunakan Preprocessing dan Query Expansion
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%

4.2.2 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 mempengaruhi 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.



Gambar 3. Gambar grafik pengujian *threshold* ekspansi kata

Gambar 3 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 yang presentasenya naik dan juga turun. Pengujian *threshold* hanya sampai pada titik 1.5 karena mulai dari nilai *threshold* tersebut presentase akurasi menunjukkan grafik yang turun hingga pada titik ke 2.0.

5. KESIMPULAN

Berdasarkan dari hasil penelitian yang telah dilakukan demikian, maka dapat diambil beberapa kesimpulan sebagai berikut:

1. Klasifikasi tweet pada penelitian dilakukan melalui beberapa tahap yaitu *preprocessing*, ekspansi kata atau penambahan kata baru pada data uji, kemudian yang terakhir adalah proses klasifikasi yang menggunakan metode *K-Nearest Neighbor*.
2. Nilai K yang telah diperoleh melalui pengujian di atas dapat diketahui hasil terbaiknya dengan tingkat akurasi 90% dengan K=5. Nilai K yang diperoleh ini mempengaruhi hasil saat dilakukan ekspansi kata.
3. Menggunakan *query expansion* dapat membantu peningkatan akurasi hasil, yang sebelumnya memiliki akurasi 90% meningkat menjadi 92%. Perolehan tersebut didapat dengan menggunakan *threshold* 0.5, 0.6, 1.1, 1.2, 1.3, 1.4. Penggunaan *threshold* mampu mempengaruhi tingkat akurasi, karena *threshold* berperan memberikan batas kata kedekatan yang dimasukkan ke data teks uji. Threshold yang bernilai terlalu kecil sulit untuk mendapatkan kata yang mempunyai kedekatan, namun ketika *threshold* ditingkatkan maka akan didapatkan kata yang mempunyai kedekatan secara semantik. Kelemahan jurtru didapat ketika *threshold* nilainya terlalu besar yang mengakibatkan terlalu banyak kata yang diperoleh untuk ditambahkan ke data teks, dan terlalu banyak juga kata yang tidak relevan yang diperoleh yang akhirnya membuat akurasi klasifikasi menurun.

6. DAFTAR PUSTAKA

Chenglong ,M. dan Weiqun X. (2014) . Distributional Representations of Words for Short Text Classification . *Institute of Acoustics, Chinese Academy of Sciences , Beijing. China.*

Fachruddin, M. (2011). Analisis dan ImplementasiPseudo Relevance Feedback dengan *Query Expansion* Menggunakan

Term Selection Value. Institut Teknologi Bandung.

Hadna, N. et al., 2016. Studi literatur tentang perbandingan metode proses analisis sentimen di twitter. Fakultas Teknik, Universitas Gadjah Mada.

Hand, D. (2010). Text Mining: Classification, Clustering, and Applications edited by Ashok Srivastava, Mehran Sahami. *International Statistical Review*, 78(1), pp.134-135.

Huang, W. dan Li ,S. (2009) School of Information System and Management, *National University of Defense Technology, Changsha, China.*

Insan, P. P. (2013). Klasifikasi Emosi untuk Teks Berbahasa Indonesia dengan Menggunakan Algoritma C5.0. *Program Studi Informatika/Ilmu Komputer PTIIK Universitas Brawijaya : Malang.*

Jing, T. dan Yue ,W. (2017) . End-to-end Learning for Short Text Expansion .*School of Information, University of Michigan. Michigan*

Kwak, H., Lee, C., Park, H., dan Moon S. (2010). What is Twitter, a Social Network or a News Media?. *Department of Computer Science, KAIST, Korea.*

Liu, M. dan Fan, X. (2012). A Method for Chinese Short Text Classification Considering Effective Feature Expansion. *International Journal of Advanced Research in Artificial Intelligence (IJARAI)*, 1(1): 1-5, 2012.

Mandala, R. (2009). Relevance Feedback And *Query Expansion*. *Cambridge*

Megantara, G. dan Kurniati, P. (2010). KLASIFIKASI TEKS DENGAN MENGGUNAKAN IMPROVED K-NEAREST NEIGHBOR ALGORITHM . *Telkom University. Bandung*

Mooney, R. J. (2006). Machine Learning Text Categorization. *University of Texas, Austin, United State*.

Netbeans, 2016. Netbeans IDE 8.0. [program komputer]. Tersedia di: <http://netbenas.org/>

O'Reilly, T. dan Milstein, S. (2012). The Twitter Book. 2nd ed. *California: O'Reilly Media, Inc*.

Sarah Zelikovitz and Haym Hirsh. 2000. Improving Short Text Classification Using Unlabeled Background Knowledge to Assess Document Similarity. In Proceeding of the *Seventeenth International Conference on Machine Learning*, volume 2000, pages 1183–1190

Sendhy, R. Dan Novianto, S. (2014) .Perbandingan Euclidean Distance Dengan Canberra Distance Pada Face Recognition

Zelikovitz, S. and Marquez, F. (2005). Transductive Learning For Short-Text Classification Problems Using Latent Semantic Indexing. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(02), pp.143-163.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M. (2010). Short text classification in twitter to improve information filtering. Proceeding of the 33rd *international ACM SIGIR conference on Research and development in information retrieval* - SIGIR '10.

Twitter, 2016. *Tweets on Twitter*. Tersedia di: <http://twitter.com> [Diakses 1 Juni 2016].

Nurjanah, W. Perdana, S. Dan Fauzi, M. (2017). Analisis Sentimen Terhadap Tayangan televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode *K-Nearest Neighbor* dan Pembobotan Jumlah *Retweet.* Fakultas Ilmu Komputer. Universitas Brawijaya Malang.

Yudi, W. (2012). Model Semantik Terdistribusi Untuk Penentuan Textual Entailment .*Program Studi Doktor Teknik Elektro dan Informatika .Bandung*