Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 10, Oktober 2019, hlm. 9762-9769 http://j-ptiik.ub.ac.id

Klasifikasi Berita Olahraga Berbahasa Indonesia menggunakan Metode BM25 dan *K-Nearest Neighbor*

Enggar Septrinas[1], Indriati[2], Arief Andy Soebroto[3]

Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]enggarseptrinass@gmail.com, [2]indriati.tif@ub.ac.id, [3]ariefas@ub.ac.id

Abstrak

Antara Sumbar merupakan sebuah portal berita yang bertujuan untuk memenuhi hak publik untuk mendapatkan informasi yang akurat dan lengkap secara seketika. Terdapat berbagai macam berita yang disajikan, salah satunya adalah berita olahraga. Dengan perkembangan informasi digital yang terus meningkat menyebabkan begitu cepatnya penyebaran dokumen berita sehingga dibutuhkan portal web yang dikelola secara profesional. Antara Sumbar merupakan portal berita yang telah dikelola secara profesional, namun ada beberapa kekurangan terhadap pengelompokkan berita olahraga, yang menyebebkan terganggunya pembaca dalam mendapatkan berita berdasarkan kategori olahraga yang ada. Kekurangan tersebut dapat diatasi menggunakan metode BM25 dan *K-Nearest Neighbor*. Proses yang dilakukan untuk mengatasi kekurangan tersebut adalah melakukan *preprocessing* terhadap dokumen berita, melakukan pembobotan dan pemeringkatan menggunakan metode BM25, dan yang terakhir proses algoritme *K-Nearest Neighbor* sebagai metode klasifikasi. Proses pengujian yang digunakan adalah metode *k-fold* sebanyak 7 kali pengujian. Data yang digunakan pada tiap-tiap pengujian yaitu sebanyak 240 dokumen latih dan 40 dokumen uji. Berdasarkan pengujian yang dilakukan, didapatkan hasil pengujian tertinggi ketika nilai ketetanggaan (*k*) sebesar 20, menghasilkan nilai *precision* = 0,921577, nilai *recall* = 0,914286 dan nilai *f-measure* = 0,917917. Proses klasifikasi dipengaruhi oleh jumlah dokumen yang digunakan serta besaran nilai *k* yang ditentukan.

Kata kunci: *berita olahraga, text mining ,klasifikasi, BM25, k-nearest neighbor*

Abstract

Antara sumbar is a news portal aims to meet the public's right to get accurate information and complete information instantly. There are various kind of news presented, one of them is sports news. Rapid improvement on digital information requires prompt news dissemination. As result, web portal needs to be professionally managed. Antara Sumbar is a professionally managed web portal, however there are some shortfall, particularly on sport news categorization that causes difficulties for readers to find the news by its category. The problem can be addressed by BM25 and K-Nearest Neighbor methods. The steps to address the problem are preprocessing document news, calculate the BM25 score of each document news, and classification process using the K-Nearest Neighbor method. The testing process followed 7 k-fold method. Data used on test is 240 training documents and 40 testing documents. Performed test obtaines results in k's value as 20, with precision value = 0,921577, recall values = 0,914286, and f-measure values = 0,917917. The process of classification is affected by the number of document used and k value.

Keywords: *sport news, text mining, classification, BM25, k-nearest neighbor*

1. PENDAHULUAN

Informasi tercepat yang mengandung ide baru atau fakta, yang dimana ide baru dan fakta tersebut menarik bagi sebagian besar orang merupakan salah satu pengertian berita (Sumadiria, 2005:65). Berita juga disajikan menggunakan beberapa media antara lain yaitu televisi, koran, radio atau media *online internet.* Begitu cepatnya perkembangan dunia teknologi, menyebabkan penggunaan teknologi tidak dapat dipisahkan dari kehidupan masyarakat pada zaman *modern* ini. Penyebaran dokumen berita begitu cepat dikarenakan perkembangan

informasi digital yang terus meningkat. Dalam rentang waktu yang sangat cepat, terlebih dengan banyaknya portal berita *online* menyebabkan banyaknya dokumen berita yang diunggah melalui internet. Berita yang disajikan dalam portal berita pun sangat beragam, salah satunya adalah berita olahraga. Berita olahraga sendiri memiliki kategeri berita yang cukup banyak sesuai dengan jenis olahraga yang ada. Pada alexa.com menjelaskan situs portal berita seperti bola.net masuk pada 25 *top site* di Indonesia (Pramudita, Putro dan Makhmud, 2018). Hal tersebut menunjukkan bahwa minat masyarakat Indonesia mengikuti portal berita olahraga sangat tinggi. Dengan minat masyarakat yang sangat tinggi tentunya harus diiringi dengan portal berita yang dikelola secara baik.

Antara Sumbar (sumbar.antaranews.com) merupakan salah satu portal berita turunan dari portal berita Antara News (antaranews.com) yang bertujuan untuk memberikan informasi terbaru yang lengkap dan bermanfaat agar hak publik terpenuhi. Terdapat berbagai macam berita yang disajikan, mulai dari berita nasional, ekonomi, otomotif, politik, hukum, hingga berita olahraga. Antara Sumbar merupakan portal berita yang telah dikelola secara baik, namun ada beberapa kekurangan terhadap pengelompokkan berita olahraga yang belum dikelompokkan secara otomatis, menyebabkan berita olahraga belum tersusun secara rapi berdasarkan kategori dan jenis dari berita. Kekurangan tersebut tentunya akan berpengaruh kepada penerima informasi. Portal berita yang telah melakukan pengelompokkan terhadap berita olahraga yang disajikan, tentunya lebih digemari oleh para penerima informasi, dikarenakan memudahkan penerima informasi dalam hal akses pencarian berita berdasarkan kategori olahraga yang diinginkannya. Dibandingkan dengan portal berita yang belum melakukan pengelompokkan, yang akan mengakibatkan penerima informasi mengalami sedikit kesulitan dalam menemukan berita berdasarkan ketegori olahraga yang ingin dicari. Oleh karena itu, dibutuhkan sebuah teknik klasifikasi untuk melakukan pengelompokkan berita olahraga berdasarkan jenis dari kategori berita itu sendiri agar memudahkan penerima informasi dalam melakukan pencarian berita berdasarkan jenis olahraga yang diinginkannya.

Klasifikasi teks merupakan sebuah teknik yang digunakan untuk mengelompokkan suatu dokumen. Pada penelitian yang dilakukan oleh Samuel, Rachmat, dan Delima (2015) menjelaskan bahwa metode yang biasanya digunakan untuk mengelompokkan suatu dokumen adalah metode *K-Nearest Neighbor*. Besarnya nilai ketetanggaan yang ditentukan pada proses klasifikasi KNN sangat mempengaruhi hasil klasifikasi. Penelitian terdahulu yang pernah dilakukan oleh Binawan, Indriati, dan Adikara (2019) menunjukkan pengaruh nilai *k* pada hasil klasifikasi. Selain besarnya nilai *k* yang ditentukan, pembagian antara dokumen latih dan uji, serta banyaknya dokumen yang dijadikan sebagai data penelitian juga akan memengaruhi kemampuan dari *K-Nearest Neighbor*. Penelitian yang dilakukan oleh Claudya, et al. (2019) menunjukkan pengaruh sedikitnya dokumen yang dijadikan sebagai data penelitian akan mendapatkan hasil akurasi yang rendah dengan menggunakan metode klasifikasi KNN. Banyaknya data penelitian akan memengaruhi kemampuan dari metode klasifikasi yang digunakan pada penelitian kali ini

Proses pemeringkatan serta pembobotan dokumen sangat dibutukan didalam proses *text mining*. Salah satu metode pembobotan serta metode pengurutan dokumen pada *text mining* adalah metode BM25. Penelitian terdahulu dilakukan oleh Pardede, Husada, dan Riansyah (2018) menunjukkan bahwa metode BM25 merupakan metode yang efektif serta memiliki ketepatan dalam pemeringkatan dokumen berdasarkan kedekatan dokumen yang diujikan dengan dokumen latih, sehingga pada kelas *best match* metode ini merupakan metode terbaik. Penelitian terdahulu yang pernah dilakukan oleh Chen, et. al. (2012) menjelaskan bahwa proses pembobotan pada metode BM25 lebih baik dibandingkan proses pembobotan pada metode TF-IDF.

Dari persoalan yang sudah diceritakan diatas, fokus pada penelitian yang dilakukan adalah bagaimana membuat suatu sistem untuk melakukan pengelompokkan dokumen berita olahraga berbahasa Indonesia berdasarkan topik pembahasan dan judul berita menggunakan metode BM25 dan *K-Nearest Neighbor* (KNN). Harapan dari penelitian yang dilakukan yaitu mendapatkan rancangan untuk membuat sistem klasifikasi berita olahraga berbahasa Indonesia menggunakan BM25 dan *K-Nearest Neighbor*, serta mengetahui kinerja dari sistem itu sendiri, dan juga memudahkan penerima informasi mendapatkan berita di portal web Antara Sumbar.

2. DASAR TEORI

2.1 Antara Sumbar

Antara Sumbar (sumbar.antaranews.com) merupakan salah satu portal berita turunan dari portal berita Antara News (antaranews.com) yang bertujuan untuk memberikan informasi terbaru yang lengkap dan bermanfaat agar hak publik terpenuhi. Terdapat berbagai macam berita yang disajikan mulai dari berita nasional, ekonomi, otomotif, politik, hukum, hingga berita olahraga.

2.2. *Text Mining*

Menurut Feldman dan Sanger (2006) proses untuk menemukan informasi baru dari sebuah dokumen menggunakan sumber dokumen lainnya yang berbeda, merupakan pengertian dari *text mining*. Seperti yang telah dijelaskan oleh Feldman dan Sanger (2006) terdapat 4 tahapan penting dalam proses *text mining*. Tahapan tersebut diantaranya adalah *Text Preprocessing, Text Transformation, Feature Selection*, dan *Pattern Discovery*.

2.3. *Text Pre-processing*

Menurut Hadna, Santoso, dan Winarno (2016), salah satu proses yang terdapat dalam *text mining* adalah *text preprocessing*. Dengan adanya *text preprocessing* nantinya akan didapatkan *term* atau token yang dianggap penting sehingga siap untuk dianalisa.

Feldman dan Sanger (2006) juga menjelaskan bahwa *text preprocessing* bertujuan untuk memudahkan proses komputasi, dengan cara melakukan pemecahan suatu dokumen menjadi potongan token/*term*/kata. Pemecahan dokumen berfungsi untuk memudahkan proses komputasi dengan dokumen yang telah dipersiapkan dan dapat diolah lebih lanjut.

Beberapa tahapan pada proses *text preprocessing* agar menemukan token/*term*/kata yang dianggap penting dan telah siap untuk dianalisa antara lain:

2.3.1. *Case Folding*

Menurut Sanjaya (2018) pada proses *casefolding* dilakukan pengubahan terhadap semua huruf besar menjadi huruf kecil. Selain itu penghapusan angka dan karakter lainnya yang bukan huruf juga dilakukan pada proses *casefolding* ini.

2.3.2. Tokenisasi

Menurut Nayak et al. (2016) pada proses tokenisasi terdapat proses pemisahan kalimat menjadi token. Pemisah yang digunakan pada proses ini berupa spasi. Dengan kata lain tokenisasi melakukan tahap pemotongan kalimat berdasarkan token atau kata yang menyusunnya.

2.3.3. *Filtering*

Filtering dilakukan dengan metode *stopwords*, yaitu menghilangkan semua kata sambung, kata ganti, dan lain-lain. Menurut Sanjaya (2018) contoh kata *stopword* dalam Bahasa Indonesia antara lain yaitu dia, kami, kamu, seperti, untuk, dan lain-lain. *Stoplist* yang digunakan untuk proses *filtering* ini menggunakan *stoplist* Sastrawi.

2.3.4. *Stemming*

Menurut Sanjaya (2018) proses *stemming* dilakukan untuk menghilangkan *term* yang mengandung sisipan, akhiran, dan awalan. Proses *stemming* akan menghasilkan *term* yang lebih optimal dari proses sebelumnya.

2.4. Metode BM25

Metode pembobotan dan pengurutan dokumen yang digunakan pada penelitan kali ini adalah metode BM25. Penentuan hasil kemiripan dilakukan berdasarkan *term* yang dimiliki dari hasil pemeringkatan yang dilakukan. Kemampuan metode BM25 dalam melakukan pembobotan kata lebih baik dibandingkan dengan metode TF-IDF (Chen dkk, 2012). Pada persamaan di bawah ini menunjukkan persamaan metode BM25 secara umum.

$$BM25(q,d)\sum_{i=1}^{|q|} idf(q_i)\frac{tf(q_i,d).(k_1+1)}{tf(q_i,d)+k_1.(1-b+b.\frac{|d|}{dl_{avg}}} \quad (1)$$

Keterangan:

$idf(.q_i.)$ = *Inverse document frequency* terhadap kata *query i*

$tf(.q_{i.},d)$ = Banyak *term frequency* q_i terhadap dokumen

k_1 = $2{,}0 \geq k_1 \leq 1{,}2$

$|d|$ = Panjang isi dokumen

dl_{avg} = Rata-rata dokumen yang didapatkan dengan membagi total panjang isi dokumen dengan total dokumen

Pada persamaan di atas, kita melihat fungsi IDF yang belum dijabarkan pada proses sebelumnya. Pada persamaan di bawah ini

dijelaskan persamaan IDF.

$$IDF(qi) = \log_{10}\left(\frac{N - df(qi) + 0.5}{df(qi) + 0.5}\right) \quad (2)$$

Keterangan:

$df(qi)$ = *Document frequency* dari kata q_i

N = Total dokumen latih

2.5. *K-Nearest Neighbor*

Salah satu metode yang berfungsi untuk menentukan kategori suatu data adalah metode *K-Nearest Neighbor*. Banyak dokumen akan memengaruhi sifat *self-learning* yang dimiliki oleh metode KNN. Menurut Streemathy dan Balamurungan (2012) metode yang biasanya digunakan untuk menentukan kategori dari sebuah data adalah metode KNN. Hasil klasifikasi ditentukan dengan cara menentukan besaran dari nilai ketetanggaan (*k)*. Dengan melihat jarak paling dekat antara data latih dengan data uji nantinya akan dapat mengetahui kategori dari dokumen yang diujikan. Kategori dari dokumen yang diujikan ditentukan dari ketetanggaan terdekatnya.

Penghitungan jarak antara dokumen menggunakan metode BM25. Setelah jarak didapatkan, dilakukan proses klasifikasi menggunakan metode KNN dengan cara mengambil data sebanyak nilai ketetanggaan data latih beserta label yang dimiliki. Kemunculan label terbanyak akan menentukan kategori dokumen yang diujikan.

2.6. Evaluasi

Precision, *recall*, dan *f-measure* adalah suatu teknik yang digunakan untuk melakukan evaluasi terhadap sistem klasifikasi berita olahraga dengan melakukan pengujian terhadap hasil klasifikasi yang didapat.

2.6.1. *Precision*

Menurut Hripsack dan Rostchild (2005) *precision* adalah nilai ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem. Rumus perhitungan nilai *precision* ditunjukkan pada persamaan di bawah ini.

$$precision = \frac{tp_i}{tp_i + fp_i} \quad (3)$$

Keterangan:

tp_i = Banyak dokumen relevan yang dikategorikan secara benar oleh sistem

fp_i = Banyak dokumen relevan yang diklasifikasikan secara salah oleh sistem

2.6.2. *Recall*

Nilai kebenaran sistem dalam melakukan prediksi adalah pengertian recall menurut Hripcsak dan Rostchild (2005). Rumus perhitungan nilai *recall* ditunjukkan pada persamaan di bawah ini.

$$recall = \frac{tp_i}{tp_i + fn_i} \quad (4)$$

Keterangan:

tp_i = Jumlah dokumen relevan yang diklasifikasikan secara benar oleh sistem

fn_i = Jumlah dokumen tidak relevan yang diklasifikasikan secara salah oleh sistem

2.6.3. *F-Measure*

Menurut Hripcsak dan Rostchild (2005) Gabungan antara *recall* dengan *precision* merupakan pengertian dari *f-measure*. Persamaan di bawah merupakan rumus untuk mengetahui nilai dari *f-measure*.

$$f - measure = 2\,x\,\frac{precision\,x\,recall}{precision + recall} \quad (5)$$

Keterangan:

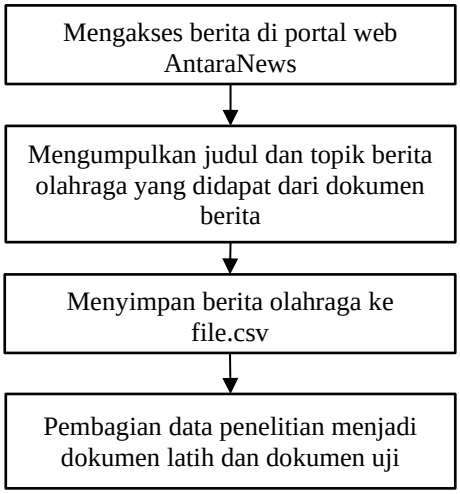
$precision$= Perbandingan antara jawaban benar dari sistem dengan informasi dari user

$recall$ = Nilai kebenaran sistem terhadap prediksi yang dilakukan

3. METODOLOGI

3.1. Pengumpulan Data

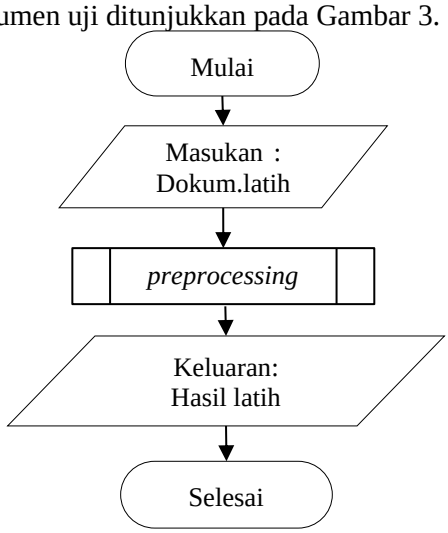
Dokumen penelitian didapatkan dari portal berita Antara News dan diolah sendiri. Terdapat 280 berita olahraga yang dijadikan sebagai data pada penelitian kali ini. Dari 280 data, dilakukan pembagian data menjadi 30 dokumen uji serta 250 dokumen latih. Kategori berita yang akan diklasifikasikan adalah berita sepakbola, bulutangkis, basket, balap, dan beladiri. Gambar 1 merupakan alur dari pengumpulan dokumen berita yang akan dijadikan sebagai data penelitian.



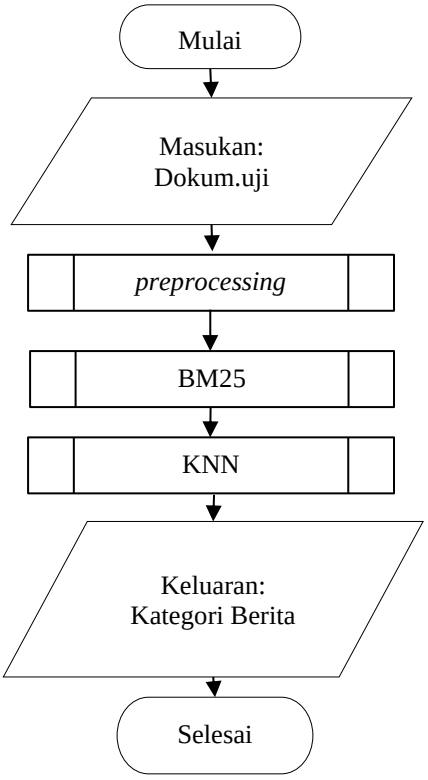
Gambar 1. Tahap Pengumpulan Data

3.2. Alur Algoritme

Pada tahapan ini dilakukan perancangan algoritme untuk dijadikan dasar saat implementasi. Rancangan pada proses dokumen latih dan dokumen uji tentunya berbeda. Pada proses dokumen latih nantinya akan mendapatkan keluaran berupa potongan kata dari suatu kalimat atau biasa disebut token unik, yang akan digunakan untuk proses *testing*. Sedangkan pada proses dokumen uji akan menghasilkan keluaran berupa kelas atau kategori dari berita yang telah ditentukan sebelumnya. Gambar 2 menunjukkan diagram alir dokumen latih sedangkan diagram alir dokumen uji ditunjukkan pada Gambar 3.



Gambar 2. Alur Dokumen *Training*



Gambar 3. Alur Dokumen *Testing*

Tahapan *pre-processing* diawali dengan proses *casefolding*, setelah itu dilanjutkan ke proses *filtering*, lalu dilakukan proses *stemming*, dan pada akhirnya dilakukan tahapan tokenisasi. Setelah didapatkan hasil dari proses *preprocessing*, dilakukan proses perhitungan nilai *term frequency* (tf), perhitungan panjang kata masing-masing dokumen (|D|), perhitungan total dokumen (N), dan melakukan perhitungan rata-rata dokumen (avgdl). Setelah masing-masing nilai telah didapatkan dilakukan perhitungan *Inverse Document Frequency* (IDF) dilanjutkan dengan menghitung nilai dari BM25. Setelah nilai BM25 pada tiap-tiap dokumen didapatkan, dilakukan proses pengurutan dokumen secara *descending*. Pengurutan dokumen berfungsi untuk melakukan penentuan kelas atau kategori dari dokumen uji menggunakan metode KNN, dengan menentukan nilai ketetanggaan *(k)*. Kelas atau kategori yang sering muncul berdasarkan nilai *k* yang telah ditentukan akan menentukan kategori dari dokumen uji.

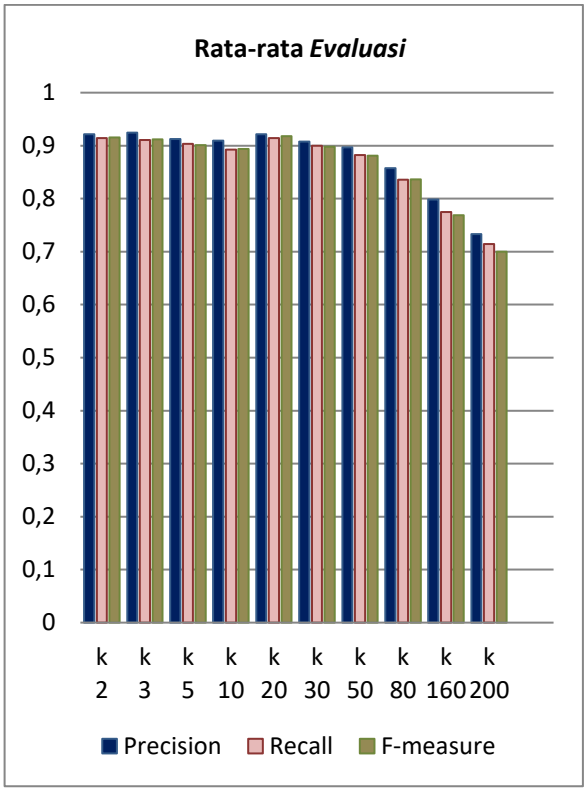
4. PENGUJIAN DAN ANALISIS

Proses pengujian dan analisa dilakukan sebagai pengukur kemampuan dari metode BM25 dan metode klasifikasi KNN. Kemampuan dari metode yang digunakan diukur

menggunakan *confusion matrix* dari tiap-tiap kelas, setelah itu ditentukan nilai evaluasi dengan tiap nilai ketetanggaan yang ditentukan. Pengujian dilakukann terhadap semua data agar hasil evaluasi didapatkan. Metode yang biasanya digunakan untuk menguji semua data adalah metode *K-Fold Validation*. *K-Fold* yang digunakan yaitu sebanyak 7 kali pengujian. Data yang digunakan pada pengujian sebanyak 280 dokumen. Dari 280 data tersebut dilakukan pembagian sebanyak 240 dokumen latih dan 40 dokumen uji dengan komposisi masing-masing 8 dokumen dari tiap-tiap kelas yaitu kelas Sepakbola, Bulutangkis, Basket, Balap, dan Beladiri. Dokumen *testing* yang digunakan pada tiap-tiap pengujian menggunakan dokumen yang berbeda. Tabel 1 dan Gambar 4 menunjukkan hasil rata-rata yang didapatkan dari pengujian yang dilakukkan.

Tabel 1. Rata-rata Nilai *Precision*, *Recall*, dan *F-Measure*

Nilai k	*Precision*	*Recall*	*F-Measure*
2	0,921283	0,914286	0,915757
3	0,924786	0,910714	0,912064
5	0,912671	0,903571	0,901269
10	0,909283	0,892857	0,89372
20	**0,921577**	**0,914286**	**0,917917**
30	0,907666	0,9	0,897726
50	0,897054	0,882143	0,88114
80	0,85724	0,835714	0,836169
160	0,798314	0,775	0,76864
200	0,732889	0,714286	0,699871



Gambar 4. Rata-rata Evaluasi

Dari pengujian yang telah dilakukan sebanyak 7 *K-Fold*, didapatkan hasil pengujian tertinggi ketika nilai ketetanggaan (*k*) sebesar 20, menghasilkan nilai *precision* = 0,921577, *recall* = 0,914286 dan *f-measure* = 0,917917. Sedangkan hasil pengujian terendah diketahui ketika nilai ketetanggaan (*k*) sebesar 200, yang menghasilkan nilai *precision* = 0,732889, *recall* = 0,714286 dan *f-measure* = 0,699871. Pengujian *K-Fold* sebanyak 7 kali dilakukan agar pembagian partisi data pada tiap-tiap pengujian merata. Sebelumnya sudah dijelaskan bahwa jumlah keseluruhan data adalah sebanyak 280 data. Dengan menggunakan 7 *K-Fold*, data dapat dibagi menjadi 7 partisi, yang masing-masing partisi berisikan 40 dokumen uji. Tiap-tiap partisi yang berisikan 40 dokumen nantinya akan dilakukan pengujian sebanyak 7 kali menggunakan partisi data yang berbeda pada tiap-tiap pengujian yang dilakukan. Dari 40 partisi data berisikan masing-masing 8 data untuk tiap-tiap kategori. 8 berita sepakbola, 8 berita bulutangkis, 8 berita basket, 8 berita balap, dan 8 berita beladiri. Dengan melakukan pembagian data secara seimbang didapatkan hasil evaluasi yang cukup tinggi.

Gambar 4 juga juga menjelaskan rata-rata dari masing-masing nilai pada tiap-tiap pengujian mengalami nilai yang naik turun

terhadap nilai *k* yang diberikan. Dapat dilihat ketika nilai ketetanggaan (*k*) sebesar 2 sampai nilai ketetanggaan (*k*) sebesar 20, rata-rata dari masing-masing nilai pada tiap-tiap pengujian mengalami nilai yang fluktuatif dengan *range* nilai yang rendah. Lain halnya ketika nilai ketetanggaan (*k*) sebesar 20 sampai nilai ketetanggaan (*k*) sebesar 200, rata-rata dari masing-masing nilai pada tiap-tiap pengujian mengalami penurunan nilai. Hal tersebut terjadi dikarenakan nilai ketetanggaan (*k*) yang semakin mendekati jumlah data *training*, sehingga mengakibatkan muncul banyaknya tetangga dengan kategori yang berbeda pada saat penentuan kelas suatu dokumen. Hal tersebut menyebabkan menurunnya akurasi sistem dalam penentuan kelas suatu dokumen.

5. KESIMPULAN

Berkaitan dengan penelitian klasifikasi berita olahraga menggunakan metode pembobotan dan pemeringkatan BM25 serta *K-Nearest Neighbor* sebagai metode klasifikasi, yang diawali dengan tahap perancangan, setelah itu dilanjutkan pada tahap implementasi, dan pada akhirnya dilakukan tahap pengujian, dapat disimpulkan beberapa hal bahwa metode pembobotan dan pemeringkatan BM25 serta *K-Nearest Neighbor* sebagai metode klasifikasi dapat dilakukan dalam penelitian kali ini. Dimulai dari tahap *preprocessing*, sampai pada tahap *sorting* nilai BM25, serta melakukan proses klasifikasi dokumen berdasarkan hasil *sorting* nilai BM25 yang telah didapatkan.

Sedangkan metode pengujian *precision, recall, dan f-measure* yang digunakan untuk evaluasi sistem didapatkan hasil kesimpulan, bahwa hasil pengujian tertinggi didapatkan ketika nilai ketetanggaan (*k)* sebesar 20, menghasilkan nilai *precision* = 0,921577, *recall* = 0,914286 dan *f-measure* = 0,917917. Sedangkan untuk hasil pengujian terendah didapatkan ketika nilai ketetanggaan (*k*) sebesar 200, menghasilkan nilai *precision* = 0,732889, *recall* = 0,714286 dan *f-measure* = 0,699871.

6. DAFTAR PUSTAKA

Binawan, D.H., Indriati dan Adikara, P.P., 2019. Klasifikasi Dokumen Abstrak Skripsi Berdasarkan Fokus Penelitian di Bidang Komputasi Cerdas Menggunakan BM25 dan *K-Nearest Neighbor*, 3(3), hal.2640-2645.

Chen, I., Du, H., Wu, S. dan Yang, C., 2012. Duplication Detection for Software Bug Reports based on BM25 Term Weighting.

Claudy, Y.I., Perdana, R.S. dan Fauzi, M.A., 2018. Klasifikasi Dokumen Twitter Untuk Mengetahui Karakter Calon Karyawan Menggunakan Algoritme K-Nearest Neighbor (KNN). 2(February), hal.2761–2765.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. USA: Cambridge University Press.

Hadna, N. M. S., Santoso, P.I. & Winarno, W.W., 2016. Studi Literatur tentang Perbandingan Metode untuk Proses Analisis Sentimen di Twitter

Hripcsak, G. dan Rothschild, A., 2005. Agreement , the F-Measure , and Reliability in Information Retrieval, 12(3), hal.269-298

Nayak, A.S., Kanive, A.P., Chandavekar, N. dan Balasubramani, 2016. Survey on Pre-Processing Techniques for Text Mining. *International Journal of Advanced Trends in Computer Science and Engineering*, 5, hal.16875–16879.

Pardede, J., Husada, M.G. dan Riansyah, R., 2018. Implementasi dan Perbandingan Metode Okapi BM25 Dan Plsa Pada Aplikasi Information Retrieval. hal.1-10.

Pramudita, Y.D., Putro S.S. dan Makhmud, N., 2018. Klasifikasi Berita Olahraga Menggunakan Metode Naïve Bayes dengan Enhanced Confix Stripping Stemmer. 5(3), hal.269-276. Tersedia melalui : Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) < http://jtiik.ub.ac.id/> [Diakses 2 Januari 2019]

Samuel, Y., Rachmat, A. dan Delima, R., 2014. Implementasi Metode K-Nearest Neighbor dengan Decision Rule untuk Klasifikasi Subtopik Berita. 10(1), hal.1-15.

Sanjaya, F., 2018. Pemanfaatan Sistem Temu Kembali Informasi dalam Pencarian Dokumen Menggunakan Metode Vector Space Model. *Journal of Information and Technology*, 5(2), hal.147–153.

Sreemathy, J. dan Balamurungan, P.S., 2012. An Efficient Text Classification Using KNN and Naive Bayesian. International Journal on Computer Science and Engineering (IJCSE), 4(3), hal. 392-396.