PREDIKSI POPULARITAS ARTIKEL BERDASARKAN JUDUL DAN INTERAKSI SOSIAL

Irwan Oyong[1], Khairan Marzuki[2], Teguh Ansyor Lorosae[3] , Kusrini[4]

[1), 2), 3), 4)] Magister Teknik Informatika, Universitas AMIKOM Yogyakarta
Jl. Ring Road Utara, Condong Catur, Depok, Sleman, Yogyakarta 55281
Email : irwan.oyong@gmail.com[1)], khairanmarzuki@gmail.com[2)], ansyorlorosae95@gmail.com[3)], kusrini@amikom.ac.id[4)]

Abstrak

Perkembangan terkini membutuhkan prediksi popularitas dari sebuah item berita bahkan sebelum diterbitkan, yang memungkinkan sebuah pengambilan keputusan tepat terkait kebutuhan modifikasi dan perbaikan. Headline merupakan bagian paling menonjol dan sering kali merupakan satu-satunya yang ditampilkan dari sebuah artikel berita. Penelitian yang dilakukan berusaha untuk membuktikan dan menunjukkan pengaruh penggunaan fitur yang diambil dari judul berita, dalam melakukan prediksi popularitas yang dijangkau pada media sosial Twitter didasarkan pada interaksi sosial yang didapatkan, yaitu total reply, likes, dan retweet. Model prediksi menggunakan teknik k-Nearest Neighbor untuk mendapatkan kemungkinan popularitas sebuah judul berita, dan memberikan feedback rekomendasi perbaikan berdasarkan nilai fitur judul berita pada set data pelatihan yang memiliki nilai jarak terdekat dari data uji yang diprediksi. Fitur yang digunakan menghasilkan angka akurasi rata-rata 89.84% pada nilai k Neighbor = 9 setelah melalui uji 5-fold cross validation.

Kata kunci : *prediksi, k-Nearest Neighbor, popularitas media berita, rekomendasi keputusan*

1. Pendahuluan

Prediksi popularitas berita pada media sosial merupakan bidang riset yang menarik dan menantang pada masa sekarang. Perkembangan terkini membutuhkan prediksi popularitas dari sebuah item bahkan sebelum dia diluncurkan, yang memungkinkan sebuah pengambilan keputusan tepat terkait kebutuhan modifikasi dan perbaikan. Media sumber berita kini memusatkan sebagian besar perhatian mereka pada media *online* di mana mereka dapat menyebarkan berita mereka secara efektif kepada populasi yang besar. Oleh karena itu, merupakan hal umum bagi media sumber berita untuk memiliki akun aktif di media sosial seperti Twitter dan Facebook untuk memanfaatkan jangkauan layanan yang sangat besar ini.

Terdapat beberapa cara berbeda dalam mengartikan istilah popularitas, salah satunya adalah jumlah *click-through / pageview*, namun hal ini jarang sekali terbuka untuk diketahui oleh umum, dan sulit mendapatkan angka pasti dikarenakan kehadiran *web crawlers* dan *search engine* [1]. Kini, membaca berita telah menjadi sebuah pengalaman sosial, maka ada ukuran lain dalam hal ketertarikan pembaca, yaitu interaksi sosial seperti komentar, *likes*, *vote*, dan bagi melalui media sosial.

Sekitar 6 dari 10 orang hanya membaca berita melalui judul / *headline*-nya saja, tanpa melakukan klik pada *link* artikel penuhnya. *Headline* merupakan bagian paling menonjol dan sering kali merupakan satu-satunya yang ditampilkan dari sebuah artikel berita. Penelitian terkait *eye-tracking* telah membuktikan perilaku ini secara empiris, bahwa banyak orang merupakan "*entry-point readers*", yang hanya membaca bagian judul saja untuk kemudian menyimpulkan garis besar suatu berita [2].

Didapatkan empat fitur utama yang mencakup spektrum informasi dari sebuah konten berita pada penelitian yang dilakukan oleh Roja Bandari, Sitaram Asur, dan Bernardo A. Huberman pada tahun 2012, yaitu sumber artikel, kategori, subjektivitas bahasa, dan entitas ternama yang disebutkan dalam berita. Hasil penelitian menunjukkan bahwa meskipun empat fitur tersebut belum dapat dikatakan cukup untuk memprediksi jumlah pasti Tweet yang akan dikumpulkan oleh sebuah artikel, mereka dapat dengan efektif menyajikan rentang kisaran popularitas yang akan didapatkan pada Twitter. Didapatkan hasil yang kurang memadai ketika melakukan prediksi menggunakan pendekatan regresi, sedangkan pengklasifikasi (*classifiers*) mencapai akurasi keseluruhan sebesar 84% [3].

Menggunakan serangkaian fitur terekstraksi yang luas (kata kunci, konten media digital, popularitas berita sebelumnya yang diacu dalam artikel, dan lain-lain), IDSS (*Intelligent Decision Support System*) yang diusulkan Kelwin Fernandes, Pedro Vinagre, dan Paulo Cortez, melakukan prediksi popularitas sebuah artikel dan kemudian melakukan optimalisasi *subset* dari fitur artikel yang dapat dilakukan modifikasi oleh penulis untuk kemudian meningkatkan probabilitas popularitas yang diprediksi [4].

Alicja Piotrkowicz et al. menyimpulkan bahwa judul berita memainkan peran penting di media sosial. Pada pendekatan baru prediksi menggunakan fitur yang berasal dari judul berita, didapati perbaikan yang signifikan pada beberapa hal dasar. Fitur yang diambil dari judul berita (umumnya dapat diubah oleh penulis

utama) terbukti berdampak positif pada performa prediksi. Hal ini menunjukkan bahwa penilaian editorial tentang isu baru dan wawasan dari riset NLP (*Natural Language Processing*) perihal tata tulis dapat diberlakukan untuk memprediksi popularitas sebuah judul berita pada media sosial [5].

Model prediksi yang digunakan pada penelitian ini dibangun berdasarkan teknik k-Nearest Neighbor. k-Nearest Neighbor (kNN) termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learner*. kNN dilakukan dengan mencari kelompok k objek dalam data pelatihan yang paling dekat (mirip) dengan objek pada data baru atau data *testing*.

kNN menggunakan set data pelatihan yang telah disimpan untuk kemudian menentukan kelas / label tujuan dari set data uji [6]. Persamaan Euclidean Distance multi-dimensi pada formula (1) digunakan untuk mendapatkan nilai jarak terdekat dengan tetangga (*neighbors*). Akurasi dari teknik ini dapat berkurang apabila terdapat *noise* atau data yang tidak berhubungan pada set data yang digunakan.

$$d(p,q) = \sqrt{(p_i - q_i)^2 + (p_n - q_n)^2} \quad \ldots(1)$$

Keterangan :

d = jarak (*distance*)

p = data pelatihan

q = data uji

p_i = nilai fitur ke-i dari data pelatihan

q_i = nilai fitur ke-i dari data uji

p_n = nilai fitur ke-n dari data pelatihan

q_n = nilai fitur ke-n dari data uji

Penelitian yang dilakukan berusaha untuk membuktikan dan menunjukkan pengaruh penggunaan fitur yang diambil dari judul berita, dalam melakukan prediksi popularitas yang dijangkau pada media sosial Twitter didasarkan pada interaksi sosial yang didapatkan, yaitu total *reply*, *likes*, dan *retweet*. Model prediksi menggunakan teknik k-Nearest Neighbor untuk mendapatkan kemungkinan popularitas sebuah judul berita, dan memberikan *feedback* rekomendasi perbaikan berdasarkan nilai fitur judul berita pada set data pelatihan yang memiliki nilai jarak terdekat dari data uji yang diprediksi. Koleksi data awal yang digunakan ada dalam bahasa Indonesia dan diambil dari akun Twitter @detikcom selaku akun media berita terbesar di Indonesia dengan jumlah *followers* 14.7 juta per tanggal 12 Desember 2017.

2. Pembahasan

Koleksi Data

Korpus berita didapatkan melalui proses pengambilan data memanfaatkan Twitter Search API pada akun @detikcom, yang berisi judul berita dan dipublikasi selama bulan Oktober-November 2017 sebagai set data awal, sebanyak 33640 *record*. Kepada setiap teks judul berita yang dikumpulkan, dilakukan pemrosesan teks untuk menghilangkan URL, seleksi data, dan pengukuran popularitas berdasarkan interaksi sosial yang didapatkan (*reply*, *likes*, dan *retweet*). Proses *cleaning* teks membuang cukup banyak *record* karena mengandung karakter simbolis dan URL yang menjadi pokok artikel, sehingga tidak dapat digunakan untuk pelatihan dan pengujian, menyisakan 16465 *record* data. Dilakukan pelabelan nilai popularitas 1 (True) kepada set data pelatihan dengan jumlah interaksi sosial >= 50, dan nilai popularitas 0 (False) kepada sisanya. Tabel 1 menunjukkan contoh teks judul dengan skor interaksi sosial dan label popularitasnya.

Tabel 1. *Contoh teks judul dengan skor interaksi sosial dan label popularitas*

Label Popularitas	**Teks**	**Jumlah Interaksi**
1 (True)	Anies: Kini Saatnya Pribumi Jadi Tuan Rumah di Negeri Sendiri	4778
0 (False)	Ini Baru Laki! Tradisi Balap Sapi di Sumatera Barat	49

Fitur Judul Berita

Ditetapkan beberapa fitur judul berita yang digunakan untuk melakukan prediksi popularitas pada penelitian ini, yang merupakan *journalism-inspired news values* dan *linguistic style*, diadopsi dan dimodifikasi dari penelitian sebelumnya [1][2][4][5], sebagai berikut :

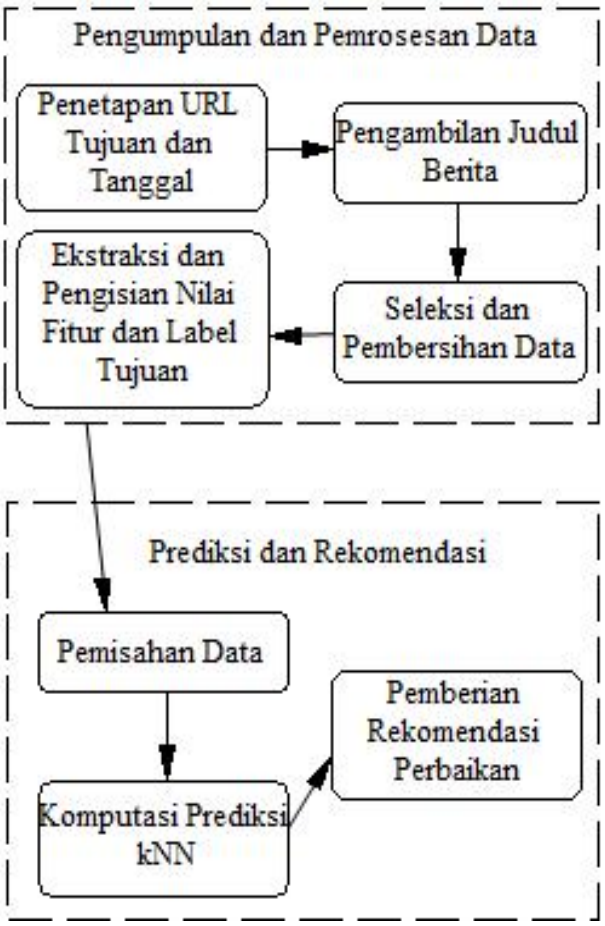
1. Entitas Ternama. Hal ini mengacu pada popularitas tempat, orang, atau organisasi yang dikenal (dan terkenal). Nilai diisikan dengan jumlah kecocokan setiap frasa yang ada pada teks judul dengan temuan dari Stanford-NER (alat bantu ekstraksi nama entitas).
2. Jumlah Kata. Jumlah total kata yang terdapat pada judul berita. Nilai diisikan dengan jumlah kata yang dipisahkan oleh spasi.
3. Jumlah Kata Kunci. Nilai diisikan dengan jumlah kecocokan kata yang terdapat pada judul berita dengan kumpulan kata kunci yang populer digunakan.
4. Kedekatan Geografis. Nilai kedekatan geografis berita dengan Indonesia, kecenderungan pembaca adalah lebih tertarik dengan berita dalam negeri

terkecuali berita luar negeri dengan tingkat urgensi yang tinggi.

5. Nilai Kejutan. Judul berita yang mengejutkan cenderung lebih menarik perhatian. Nilai diisikan dengan jumlah kecocokan kata yang terdapat pada judul berita dengan kumpulan kata bernilai kejut.
6. Jumlah Tanda Baca. Panduan tata tulis *The Guardian* menyarankan untuk tidak menggunakan tanda baca seperti tanda petik, tanda tanya, dan tanda seru.

Model Prediksi

Model konseptual dari prediksi dan pengambilan keputusan optimasi dijelaskan melalui Gambar 1 yang mencakup proses pengambilan data artikel (berita), seleksi data, seleksi fitur untuk mengambil dan mengisikan nilai fitur yang dipakai. Set data yang dikumpulkan dipisahkan berdasarkan rasio interaksi sosial yang didapatkan. Prediksi dilakukan menggunakan teknik kNN untuk menemukan probabilitas popularitas dari calon judul berita (data uji). Jika hasil prediksi memberikan nilai populer 0 (False), maka akan diberikan rekomendasi perbaikan berdasarkan data pelatihan yang memiliki nilai populer 1 (True) dan jarak terdekat.

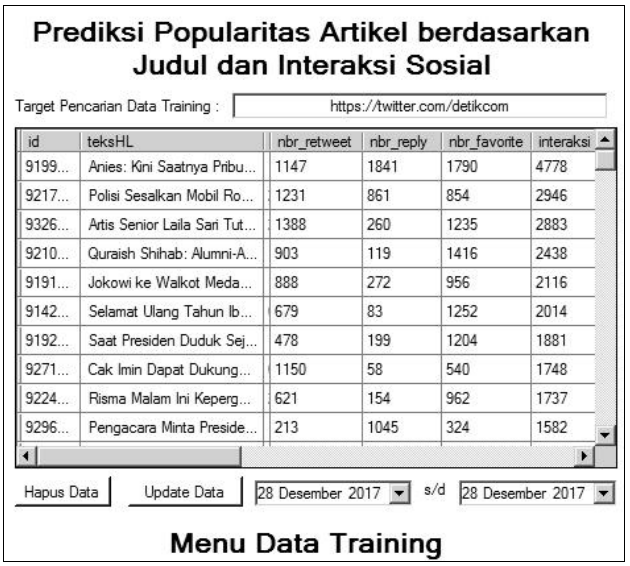


Gambar 1. *Model Konseptual Prediksi dan Pemberian Rekomendasi*

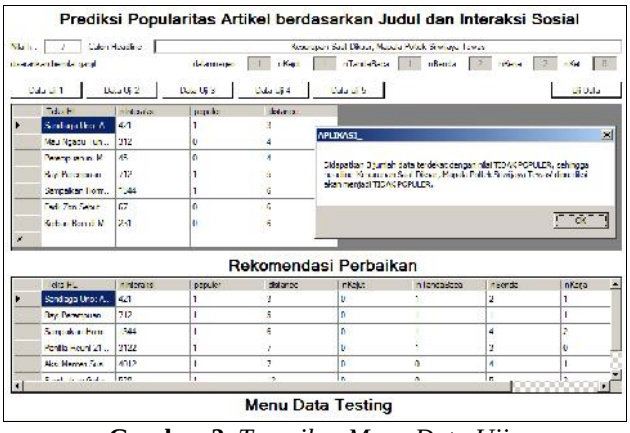
Hasil dan Diskusi

Terdapat beberapa fitur lain yang dinyatakan berdampak positif terhadap prediksi yang dilakukan oleh peneliti sebelumnya, seperti jumlah kata kerja, jumlah kata benda, dan jumlah kata keterangan cara [5]. Namun belum dapat digunakan secara optimal pada penelitian ini dikarenakan sangat bergantung pada alat bantu pemrosesan bahasa alami (Natural Language Processing), yang mana masih sangat minim ketersediaannya untuk Bahasa Indonesia. Pemaksaan penggunaan beberapa fitur tersebut membuat model prediksi mendapatkan nilai akurasi yang sangat rendah, yaitu 20% - 60% untuk nilai k Neighbor dalam rentang 3 – 10, sehingga kemudian tidak dilanjutkan sebagai fitur yang digunakan dalam penelitian.

Gambar 2 dan 3 merupakan tampilan antarmuka dari rancangan yang digunakan untuk mendapatkan hasil pengujian dan pemberian rekomendasi pada penelitian ini.



Gambar 2. *Tampilan Menu Data Pelatihan*



Gambar 3. *Tampilan Menu Data Uji*

Pengembangan dilakukan dengan modifikasi fitur hingga mendapatkan hasil yang optimal. Percobaan terakhir mendapatkan hasil terbaik dengan menggunakan fitur

seperti Entitas Ternama, Jumlah Kata, Jumlah Kata Kunci, Kedekatan Geografis, Nilai Kejutan, dan Jumlah Tanda Baca. Rancangan berusaha memberikan rekomendasi perbaikan untuk setiap data uji yang mendapatkan nilai label tujuan populer 0 (False), dengan cara menampilkan sejumlah *k* data latih dengan nilai populer 1 (True) terdekat darinya. Diharapkan dengan sajian rekomendasi tersebut, pihak redaksi / penulis dapat melakukan penyesuaian dan perbaikan sesuai dengan isian nilai fitur yang dikatakan mendapat jumlah interaksi sosial lebih tinggi.

Uji Validitas

Untuk menghindari *overfitting* (model hanya mendapatkan akurasi tinggi pada data yang sudah ada saja, namun tidak pada data yang benar-benar baru), uji validitas pada penelitian ini dilakukan dengan teknik k-*fold cross validation*, sebanyak 5 kali iterasi. K-*fold cross validation* merupakan salah satu metode praktis populer dalam mendapatkan estimasi tingkat akurasi dan *error* yang baik dari sebuah model pembelajaran [7]. Pengujian dilakukan dengan pengacakan baris-baris dataset, yakni set data sebanyak 16465 dipartisi menjadi 5 bagian dengan proporsi yang sama, yakni 20% dan menghasilkan 3293 *record* pada setiap partisinya. Pada setiap partisi dilakukan pengujian masing-masing hanya satu kali terhadap 4 partisi lainnya. Melalui pengujian sebanyak 5 kali iterasi, didapatkan akurasi rata-rata seperti yang ditunjukkan pada tabel 2.

Tabel 2. *Hasil uji 5-fold cross validation*

Kali	Cocok	Akurasi
Iterasi 1	2897	87.97%
Iterasi 2	2977	90.40%
Iterasi 3	2924	88.80%
Iterasi 4	3010	91.41%
Iterasi 5	2984	90.62%
Rata-rata		**89.84%**

Uji validitas dilakukan terhadap nilai label tujuan populer yang diisikan menggunakan model prediksi kNN yang dibangun dengan nilai k Neighbor 9. Nilai k Neighbor 9 menunjukkan hasil prediksi dengan nilai akurasi terbaik dibandingkan dengan nilai k Neighbor lebih rendah maupun lebih tinggi (1, 3, 5, 7, 10, 11, 13, dan 15).

3. Kesimpulan dan Saran

Seperti yang telah diungkapkan oleh Piotrkowicz et al., fitur-fitur dalam judul berita / headline menunjukkan pengaruh positif yang signifikan dalam melakukan prediksi popularitas sebuah judul berita pada media sosial / online, dan hal tersebut juga berlaku untuk media berita dalam bahasa Indonesia. Hal ini memberikan pembuktian bahwa pemrosesan teks judul berita untuk pengambilan fitur membantu penyelesaian tugas prediksi popularitas dengan rata-rata akurasi 89.84% dalam uji validitas 5-*fold cross validation*.

Sistem pendukung keputusan perbaikan untuk publikasi judul artikel yang dikembangkan merupakan langkah awal pendekatan dalam bahasa Indonesia. Kendala yang dihadapi adalah minimnya alat bantu terkait pemrosesan teks NLP dan ketersediaan korpus dalam bahasa Indonesia, sehingga akurasi dari pengisian nilai fitur belum bisa mendapatkan hasil yang sangat baik sehingga harus meninggalkan beberapa fitur terkait judul berita yang dikatakan memiliki dampak positif terhadap kualitas prediksi. Pengujian dan pengambilan keputusan menggunakan model kNN memiliki risiko keperluan komputasi yang besar sesuai dengan jumlah set data yang ada.

Pekerjaan di masa yang akan datang dapat dimulai dengan perancangan alat bantu pemrosesan bahasa alami (NLP) untuk melengkapi fitur yang digunakan, dan dapat dilakukan pengujian validitas dengan nilai *k-fold* yang lebih besar untuk mendapatkan tingkat robustness yang lebih tinggi. Rancangan dapat dikembangkan ke dalam bentuk *website* yang dapat diakses oleh banyak pihak yang memiliki kepentingan dalam prediksi dan perbaikan judul berita untuk mendapatkan interaksi sosial yang lebih baik. Masukan data dari pengguna aktif juga berkemungkinan akan memiliki dampak yang baik jika kemudian digunakan untuk pengembangan model prediksi dengan teknik Machine Learning yang dapat menjadi semakin cerdas dalam melakukan prediksi berdasarkan data yang ada.

Daftar Pustaka

[1] A. Tatar, P. Antoniadis, M.D. Amorim, S. Fdida, "From Popularity Prediction To Ranking Online News", *Springer: Soc. Netw. Anal. Min. 2014*, 2014

[2] J. Holsanova, H. Rahm, K. Holmqvist, "Entry Points and Reading Paths on Newspaper Spreads: Comparing a Semiotic Analysis with Eye-Tracking Measurements", *in Visual Communication, Vol 5, Issue 1*, pp. 65 – 93, 2006

[3] R. Bandari, S. Asur, B.A. Huberman, "The Pulse of News in Social Media: Forecasting Popularity", *in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pp. 26-33, 2012.

[4] K. Fernandes, P. Vinagre, P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News", *in 17th Portuguese Conference on Artificial Intelligence, EPIA*, pp. 535-546, 2015.

[5] A. Piotrkowicz, V. Dimitrova, J. Otterbacher, K. Markert, "Headlines Matter: Using Headlines to Predict the Popularity of News Articles on Twitter and Facebook", *Association for the Advancement of Artificial Intelligence*, 2017

[6] Wu J., Gao Z., Hu C., "An Empirical Study on Several Classification Algorithms and Their Improvements. In: Cai Z., Li Z., Kang Z., Liu Y. (eds) Advances in Computation and Intelligence", *ISICA 2009. Lecture Notes in Computer Science, vol 5821. Springer, Berlin, Heidelberg*, 2009

[7] S. Kale, R. Kumar, S. Vassilvitskii, "Cross-Validation and Mean-Square Stability", *Proceedings of 2nd Symposium on Innovations in Computer Science (ICS)*, 2011

Biodata Penulis

Irwan Oyong, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK Indonesia Banjarmasin, lulus tahun 2015. Sedang menempuh pendidikan Magister Teknik Informatika di Universitas AMIKOM Yogyakarta.

Khairan Marzuki, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika STT Pelita Bangsa, lulus tahun 2013. Sedang menempuh pendidikan Magister Teknik Informatika di Universitas AMIKOM Yogyakarta.

Teguh Ansyor Lorosae, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Sistem Informasi STMIK Amikom Yogyakarta, lulus tahun 2016. Sedang menempuh pendidikan Magister Teknik Informatika di Universitas AMIKOM Yogyakarta.

Kusrini, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Ilmu Komputer Universitas Gadjah Mada, lulus tahun 2002. Memperoleh gelar Master Komputer (M.Kom), Jurusan Ilmu Komputer Universitas Gadjah Mada, lulus tahun 2006. Memperoleh gelar Doktor (Dr), Jurusan Ilmu Komputer Universitas Gadjah Mada, lulus tahun 2010. Saat ini menjadi Dosen dan Direktur Program Pascasarjana di Universitas Amikom Yogyakarta.

ISSN : 2302-3805