Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN: 2548-964X
Vol. 3, No. 3, Maret 2019, hlm. 2768-2773 http://j-ptiik.ub.ac.id

Prediksi Rating Novel Baru Berdasarkan Sinopsis Menggunakan Genre Based Collaborative Filtering dan Text Similarity

Rhevitta Widyaning Palupi[1], Yuita Arum Sari[2], Putra Pandu Adikara[3]

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

Abstrak

Novel merupakan suatu cerita yang memiliki alur panjang yang bersifat imajiantif. Berdasarkan pilihan editor pada situs Amazon.com, 50 dari 100 buku dengan penjualan terbaik merupakan novel. Hal tersebut menunjukkan bahwa ketertarikan masyarakat terhadap novel cukup tinggi sebagai salah satu jenis bacaan. Namun saat ingin memilih novel yang hendak dibaca, pembaca terkadang merasa bingung untuk mengetahui kualitas dari novel tersebut. Salah satu acuan dalam melihat kualitas suatu produk yaitu rating. Situs Goodreads merupakan salah satu situs yang memungkinkan peninjau amatir menuliskan ulasan serta rating untuk membantu pembaca dalam memilih buku yang relevan. Namun terkadang pengguna Goodreads tidak memberikan rating terhadap suatu buku sehingga pengikut dari pengguna tersebut ingin mengetahui rating yang diberikan pengguna pada buku tersebut. Penelitian ini menggunakan metode *Genre Based Collaborative Filtering* sebagai penghitungan prediksi rating dan *Text Similarity* untuk mengetahui nilai kesamaan antara dokumen yang satu dengan lainnya. Data yang digunakan pada penelitian ini sebanyak 31 user dan 90 sinopsis sebagai data latih dan 35 sinopsis sebagai data uji. Akurasi sistem yang diperoleh dari hasil klasifikasi dengan menggunakan nilai kesamaan pada text similarity sebesar 45,714286% dan nilai MAE sebesar 0,27742857 sehingga dapat disimpulkan bahwa metode *Genre Based Collaborative Filtering* dan *Text Similarity* dapat digunakan untuk melakukan prediksi rating.

Kata kunci: *prediksi rating, novel, Goodreads, Genre Based Collaborative Filtering, Text Similarity*

Abstract

The novel is a story that has a long imaginary plot. Based on the editor's choice on the Amazon.com website, 50 of the 100 best-selling books are novels. This shows that public interest in the novel is quite high as one type of reading. But when you want to choose a novel that you want to read, readers sometimes feel confused to know the quality of the novel. One reference in looking at the quality of a product is rating. The Goodreads site is one site that allows amateur reviewers to write reviews and ratings to help readers choose relevant books. But sometimes Goodreads users don't give ratings to a book so followers from that user want to know the rating given by the user in the book. This study uses the Genre Based Collaborative Filtering method as a calculation of rating predictions and Text Similarity to determine the value of similarity between documents with each other. The data used in this study were 31 users and 90 synopsis as training data and 35 synopsis as test data. System accuracy obtained from the classification results by using the similarity value on text similarity of 45,714286% and MAE value of 0,27742857 so that it can be concluded that the method of genre based collaborative filtering and text similarity can be used to make rating predictions.

Keywords: *rating predictions, novels, Goodreads, Genre Based Collaborative Filtering, Text Similarity*

1. PENDAHULUAN

Novel adalah suatu cerita yang memiliki alur panjang yang bersifat imajinatif, (Tarigan, 2011). ISBN mencatat terdapat 50.498 buku fiksi baru telah terbit di tahun 2013 (ISBN, 2014). Berdasarkan pilihan editor pada situs Amazon.com menunjukkan bahwa 50 dari 100 buku dengan penjualan terbaik sepanjang masa merupakan novel (Amazon.com, 2018). Hal tersebut membuktikan tingginya ketertarikan masyarakat terhadap novel sebagai salah satu

jenis bacaan. Namun ketika ingin membaca novel baru, pembaca terkadang merasa kebingungan untuk menentukan dan mengetahui kualitas dari novel tersebut.

Rating merupakan salah satu bagian dari *review* yang menggunakan simbol bintang dalam mengekspresikan pendapat dari pelanggan. Rating dapat diartikan sebagai penilaian pengguna pada suatu produk tertentu berdasarkan pengalaman pengguna saat menggunakan dengan produk tersebut (Li & Zhang, 2002). Dikutip dari advertising-indonesia.id rating digunakan sebagai acuan dalam melihat kualitas suatu program atau produk berdasarkan respon masyarakat terhadap program atau produk tersebut (Advertising-Indonesia, 2017). Salah satu situs terbesar di dunia yang fokus pada pembaca dan rekomendasi buku yaitu goodreads.com. Situs Goodreads memungkinkan peninjau amatir untuk menuliskan ulasan tentang buku serta memberikan kritik pedas kepada penulis novel karena tinjauan dari professional cenderung positif (Jane, 2014). Selain ulasan, Goodreads juga memberikan fasilitas rekomendasi dan rating yang dapat membantu pembaca untuk memilih buku yang relevan (Thelwall & Kousha, 2016). Namun terkadang pengguna Goodreads hanya membaca tanpa memberikan rating sehingga pengikut dari pengguna tersebut ingin tahu berapa rating yang diberikan oleh pengguna terebut pada suatu novel tertentu.

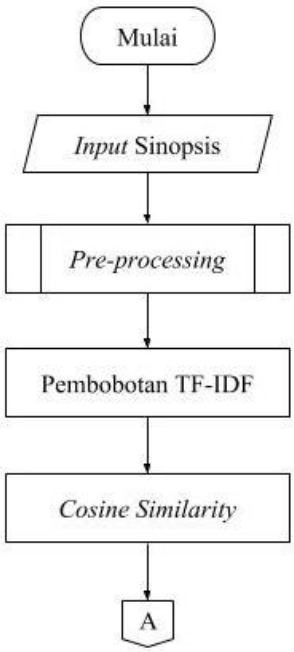
Ming-jia Wang dan Jin-ti Han (2012) melakukan penelitian dengan menggunakan algoritme *item based collaborative filtering*. Metode pada penelitian ini juga digunakan untuk menghitung nilai rating dari item yang belum dinilai oleh pengguna. Hasil dari penelitian ini menunjukkan bahwa algoritme *collaborative filtering* yang mengambil informasi karakteristik pada *item* dapat meningkatkan akurasi prediksi dan rekomendasi (Wang & Han, 2012).

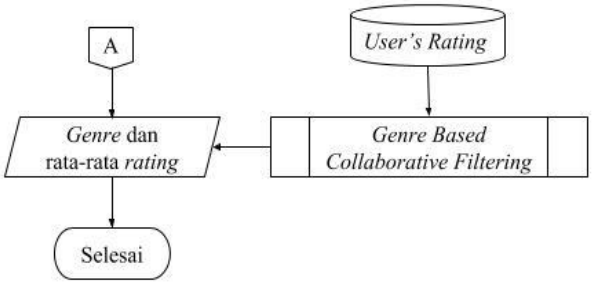
Selain itu Bening (2018) melakukan penelitan mengenai klasifikasi berita *online* dengan menggunakan pembobotan TF-IDF dan *Cosine Similarity*. Untuk melakukan proses tersebut dilakukan poses *preprocessing* diantaranya tokenisasi, penghapusan *stopword*, dan *stemming* untuk memperkecil term sehingga mempercepat proses perhitungan pembobotan. Klasifikasi yang dilakukan oleh Bening memiliki tingkat akurasi sebesar 91,25% (Herwijayanti, et al., 2018).

Berdasarkan penjelasan tersebut maka pada penelitian ini akan dilakukan prediksi rating pada novel baru berdasarkan sinopsis menggunakan *genre based collaborative filtering* dan *text similarity*. Diharapkan dengan menggunakan metode tersebut, pembaca novel dapat mendapatkan informasi mengenai kualitas buku yang hendak dibaca berdasarkan rating yang telah diprediksi. Oleh karena itu, maka dilakukan penelitian dengan judul "Prediksi *Rating* Novel Baru Berdasarkan Sinopsis Menggunakan *Genre Based Collaborative Filtering* Dan *Text Similarity*", dimana penelitian ini ditujukan untuk melakukan prediksi rating pada novel baru dengan menggunakan *genre based collaborative filtering* dan *text similarity*.

2. METODE USULAN

Prediksi rating novel baru berdasarkan sinopsis menggunakan *genre based collaborative filtering* dan *text similarity* memiliki bebarapa tahapan dalam perancangan sistemnya. Tahap pertama yang dilakukan yaitu prediksi rating user terhadap novel yang kosong. Tahap selanjutnya yaitu pengolahan data dengan menggunakan *pre-processing*. Kemudian perhitungan bobot term, hingga pengklasifikasi dokumen menggunakan nilai kedekatan pada *cosine similarity*. Dalam klasifikasi dapat diketahui prediksi genre dan rating pada dokumen yang diuji. Deskripsi umum sistem dapat dilihat pada Gambar 1.





Gambar 1. Deskripsi umum sistem

2.1. *Genre Based Collaborative Filtering*

Tahap pertama dalam melakukan prediksi rating dalam sistem ini yaitu melakukan perhitungan dengan menggunakan *genre based collaborative filtering*. *Genre based collaborative filtering* merupakan salah satu metode dalam memberikan rekomendasi berdasarkan kemiripan antar genre. Terdapat dua tahap untuk melakukan prediksi dengan *genre based collaborative filtering*. Tahap pertama yaitu menghitung nilai persamaan antar genre dengan menggunakan satu jumlah *similarity measure* (Mustafa, et al., 2017). Untuk menemukan kemiripan antar genre digunakan rumus persamaan (1).

$$sim(i,j) = \frac{\Sigma u \in U\ (R_{u,i} - \overline{R_u})(R_{u,j} - \bar{R}_u)}{\sqrt{\Sigma u \in U\ (R_{u,i} - \overline{R_u})^2}\sqrt{\Sigma u \in U\ (R_{u,j} - \overline{R_u})^2}} \quad (1)$$

Keterangan:

$sim(i,j)$: nilai kemiripan

$u \in U$: *user u* yang melakukan rating pada *genre i* dan *j*

$R_{u,i}$: rating *user u* pada genre *i*

$R_{u,j}$: rating *user u* pada genre *j*

$\bar{R}_u$: nilai rata-rata rating *user u*

Selanjutnya menghitung prediksi rating untuk genre yang belum diketahui nilai ratingnya. Untuk menghitungnya digunakan persamaan (2)

$$P(u,j) = \frac{\Sigma i \in j\ (R_{u,i} \times S_{i,j})}{\Sigma i \in j\ |S_{i,j}|} \quad (2)$$

Keterangan:

$P(u,j)$: nilai prediksi *user u* pada genre *j*

$i \in j$: genre yang mirip dengan genre *j*

$R_{u,i}$: rating yang diberikan *user u* pada genre *i*

$S_{i,j}$: nilai kemiripan genre *i* dan genre *j*

2.2. *Pre-processing*

Pre-processing merupakan suatu proses awal yang dilakukan dalam pemrosesan teks terhadap data dalam bentuk teks untuk menghasilkan data berbentuk angka (Wahyuni, et al., 2017). Terdapat lima proses dalam *pre-processing* diantaranya:

1. *Tokenization*
2. *Cleaning*
3. *Case Folding*
4. *Filtering*
5. *Stemming*

2.3. Pembobotan TF-IDF

TF-IDF merupakan proses perhitungan bobot term dengan cara mengalikan frekuensi kemunculan term dalam sebuah dokumen dengan *inverse* frekuensi dokumen yang mengandung term tersebut. Untuk menghitung bobot term digunakan persamaan (3).

$$TF - IDF = \log(1 + tf_{t,d}) \times log_{10}(\frac{N}{df_t}) \quad (3)$$

Keterangan:

TF-IDF : bobot term *t* pada dokumen *d*

$Tf_{t,d}$: jumlah kemunculan term *t* pada dokumen *d*

N : jumlah dokumen

df_t : jumlah dokumen yang mengandung *term t*

2.4. *Cosine Similarity*

Cosine similarity merupakan suatu perhitungan kesamaan antara dua vektor dengan mencari kosinus sudut antara kedua vektor dimana nilai terkecil yaitu nol dan nilai terbesar adalah satu (Kesuma & Pribadi, 2016). Untuk menghitung nilai *cosine similarity* digunakan persamaan (4).

$$Cos\ \alpha = \frac{A \bullet B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (4)$$

Keterangan:

A : vektor *A*

B : vektor *B*

$A \bullet B$: perkalian antara vektor *A* dan vektor *B*

$|A|$: panjang vektor *A*

$|B|$: panjang vektor *B*

$|A||B|$: perkalian antara $|A|$ dan $|B|$

3. HASIL DAN PEMBAHASAN

Evaluasi yang digunakan dalam penelitian ini yaitu akurasi untuk klasifikasi genre dan *Mean Absolute Error* untuk prediksi rating.

3.1 Hasil dan Analisis Prediksi Rating

Terdapat 31 *user* dan 90 sinopsis yang digunakan sebagai data latih untuk mendapatkan

nilai prediksi rating pada tiap genre. *Genre based collaborative filtering* digunakan sebagai penghitung prediksi rating yang diberikan *user* pada masing-masing sinopsis. Setelah diketahui semua rating yang diberikan oleh *user*, dilakukan perhitungan rata-rata rating tiap genre. Hasil dari rata-rata tersebut yang akan digunakan sebagai prediksi rating per genre. Hasil dari prediksi rating per genre ditunjukkan pada Tabel 1.

Tabel 1. Hasil rata-rata rating per genre

Genre	Nomor sinopsis per genre	Rata-rata rating genre
Fiction	1 - 18	3,93
Fantasy	19 - 34	3,88
Horror	35 - 45	3,98
Sci-fic	46 - 57	3,85
Mystery	58 - 70	3,86
Thriller	71 - 82	3,72
Romance	83 - 90	3,56

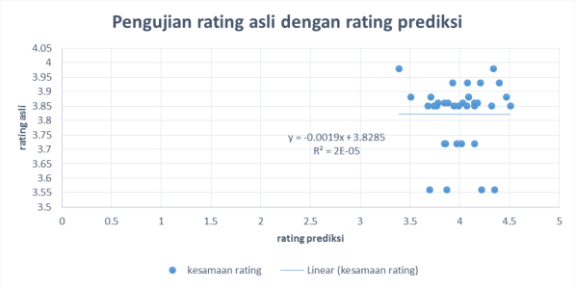
Setelah diketahui hasil dari perhitungan prediksi rating berdasarkan genre dengan menggunakan *genre based collaborative filtering*, dilakukan pengujian pada sistem dengan menggunakan data uji. Contoh hasil dari pengujian sistem ditunjukkan dalam Tabel 2.

Tabel 2. Hasil Pengujian Prediksi Rating

No	Rating Asli	Rating Prediksi
1	4,08	3,93
2	4,09	3,88
3	4,18	3,86
4	3,71	3,88
5	3,97	3,72
6	3,86	3,72
7	4,02	3,72

Dari hasil pengujian sistem terhadap prediksi rating berdasarkan genre, dapat diketahui nilai MAE yang dihasilkan termasuk kecil yaitu 0,27742857. Hal ini dikarenakan nilai rating yang sesungguhnya dengan nilai rating prediksi yang tidak terlalu jauh berbeda.

Kemudian dilihat dari diagram *scatter* dari hasil pengujian rating pada data uji terhadap sistem diperoleh bahwa nilai rating dapat diprediksi sebesar -0,0019 kali nilai rating asli ditambah dengan nilai konstan sebesar 3,8285 dengan nilai koefisien korelasi sebesar 0,00002. Grafik pengujian rating asli dengan rating prediksi ditunjukkan pada Gambar 2.



Gambar 2. Hasil pengujian rating asli dan rating prediksi

3.2 Hasil dan Analisis Klasifikasi Sinopsis

Berdasarkan dari hasil pengujian genre data uji terhadap sistem diperoleh bahwa jumlah data latih tidak memberikan pengaruh yang besar pada hasil prediksi untuk genre. Contoh hasil dari klasifikasi sinopsis ditunjukkan dalam Tabel 3.

Tabel 3. Hasil Pengujian Klasifikasi Sinopsis

No	Genre Asli	Klasifikasi Genre
1	Fiction	Fiction
2	Fantasy	Fantasy
3	Horror	Mystery
4	Sci-fic	Fantasy
5	Mystery	Thriller
6	Thriller	Thriller
7	Romance	Thriller

Dari hasil pengujian klasifikasi sinopsis, dapat diketahui bahwa genre *fiction* yang memiliki data latih yang paling banyak yaitu sebanyak 18 dengan jumlah term sebanyak 985 hanya menghasilkan 4 genre yang sama yaitu *fiction* yang memiliki data latih sebanyak delapan belas sinopsis hanya menghasilkan prediksi genre *fiction* sebanyak empat dari tiga puluh lima data uji yang mana tiga diantaranya merupakan genre *fiction* asli sedangkan satu yang lainnya merupakan dari genre *romance.* Sedangkan genre *sci-fic* yang hanya memiliki data latih sebanyak dua belas dengan jumlah term sebanyak 292 menghasilkan kesamaan genre paling banyak yaitu sejumlah Sembilan. Satu berasal dari genre *fiction*, dua dari genre *fantasy*, tiga dari genre *sci-fic*, dua dari genre *mystery*, dua dari genre *horror*, dan satu dari genre *mystery*. Akurasi yang dihasilkan dari

klasifikasi dengan *text similarity* hanya menghasilkan nilai akurasi sebanyak 0,45714286 %.

3.3 Contoh prediksi rating berdasarkan sinopsis

Untuk prediksi rating, telah diketahui dari hasil perhitungan *genre based collaborative filtering* pada data latih. Untuk klasifikasi sinopsis digunakan *text similarity* dimana nilai kedekadan antara sinopsis data uji dengan data latih berdasarkan dari hasil perhitungan *cosine similarity*. Tabel 4 Beikut ini merupakan contoh sinopsis novel baru yang akan diklasifikasi.

Tabel 4. Data uji novel baru

Genre	Sinopsis	Rating
Fantasy	Beneath the Sugar Sky returns to Eleanor West 's Home for Wayward Children. At this magical boarding school, children who have experienced fantasy adventures are reintroduced to the "real" world. Sumi died years before her prophesied daughter Rini could be born. Rini was born anyway, and now she ’s trying to bring her mother back from a world without magic.	4,09

Langkah-langkah yang diperlukan untuk melakukan klasifikasi sinopsis yaitu dengan *pre-processing*. Di dalam proses *pre-processing* sinopsis akan di *tokenization*, *cleaning*, *case folding*, *filtering*, serta *stemming*. Selanjutnya dilakukan pembobotan term pada sinopsis dengan menggunkan TF-IDF. Langkah terakhir yaitu menghitung kesamaan antara data latih dengan data uji menggunakan *cosine similarity*. Dari langkah-langkah tersebut didapatkan bahwa sinopsis novel baru pada Tabel 4 mendapatkan hasil klasifikasi genre berupa *fantasy*. Pada perhitungan prediksi yang dilakukan sistem, diketahui bahwa genre *fantasy* mendapatkan prediksi rating sebesar 3,88. Hasil dari klasifikasi dan prediksi ditunjukkan pada Tabel 5.

Tabel 5. Hasil prediksi rating dan klasifikasi

Genre Asli	Rating Asli	Klasifikasi Genre	Rating Prediksi
Fantasy	4,09	Fantasy	3,88

Dari contoh dapat diketahui bahwa nilai rating asli dengan rating prediksi memiliki nilai yang hampir mirip. Sedangkan dari klasifikasi genre yang dihasilkan menunjukkan bahwa sinopsis ini merupakan salah satu sinopsis yang memiliki kesamaan term paling banyak dengan term yang ada pada genre *fantasy* sehingga klasifikasi genre yang dihasilkan juga *fantasy*.

4. KESIMPULAN

Dalam perancangan prediksi rating novel baru menggunakan *genre based collaborative filtering* dan *text similarity* diperlukan beberapa proses tahapan yaitu yang pertama merupakan proses prediksi rating yang masih kosong dengan menggunakan *genre based collaborative filtering* untuk menghitung prediksi rating. Selanjutnya menggunakan *text similarity* untuk klasifikasi genre. Salah satu keunggulan dari sistem ini yaitu jumlah data latih yang tidak tertalu mempengaruhi hasil dari prediksi kedekatan nilai *cosine similarity*. Seperti pada genre *ficiton* yang memiliki data latih sebanyak delapan belas sinopsis hanya menghasilkan prediksi genre *fiction* sebanyak empat dari tiga puluh lima data uji yang mana tiga diantaranya merupakan genre *fiction* asli sedangkan satu yang lainnya merupakan dari genre *romance*. Hal tersebut mempengaruhi akurasi sehingga akurasi yang dihasilkan dari *text similarity* yaitu sebesar 45,714286 % . Sedangkan kekurangan dari sistem ini yaitu *user* dipaksa untuk memiliki keminatan pada semua genre dan memberikan rating pada semua jenis genre sehingga menyebabkan nilai rating yang hampir sama antara nilai yang sesungguhnya dengan nilai prediksi.

Dari penelitian yang telah dilakukan terdapat beberapa saran yang dapat digunakan sebagai saran untuk penelitian selanjutnya yaitu fitur yang digunakan dapat hanya berupa sinopsis dan rata-rata rating total yang tersedia pada halaman goodreads.com serta untuk melakukan klasifikasi dapat menggunakan metode *Support Vector Machine* (SVM) karena metode ini memiliki tingkat akurasi yang baik.

5. DAFTAR PUSTAKA

Advertising-Indonesia, 2017. *Mengenal Manfaat dan Kekurangan Rating*. [Online] Available at: http://advertising-indonesia.id/2017/09/18/mengenal-manfaat-dan-kekurangan-rating/ [Accessed 15 February 2018].

Amazon.com, 2018. *Top 20 lists in Books.* [Online] Available at: https://www.amazon.com/ [Accessed 23 November 2018].

Goodreads, n.d. *goodreads.com.* [Online] Available at: https://www.goodreads.com/about/us [Accessed 19 11 2018].

Herwijayanti, B., Ratnawati, D. E. & Muflikhah, L., 2018. Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, Volume 2, pp. 306-312.

ISBN, 2014. *Traditional Print Book Production Dipped Slightly in 2013.* [Online] Available at: http://www.bowker.com/news/2014/Traditional-Print-Book-Production-Dipped-Slightly-in-2013.html [Accessed 13 February 2018].

Jane, 2014. *On the importance of pseudonymous activity.* [Online] Available at: https://dearauthor.com/features/essays/on-the-importance-of-pseudonymous-activity/ [Accessed 9 11 2018].

Li, N. & Zhang, P., 2002. *Consumer Online Shopping Attitudes and Behavior: An Assessment of Research.* Dallas, AMCIS.

Mustafa, N., Ibrahim, A. O., Ahmed, A. & Abdullah, A., 2017. *Collaborative Filtering: Techniques and Applications.* Khartoum, s.n.

Nurgiyantoro, B., 1995. *Teori Pengkaji Fiksi.* Yogyakarta: Gadjah Mada University Press.

Puntheeranurak, S. & Chaiwitooanukool, T., 2011. An Item-based Collaborative Filtering Method using Item-based Hybrid Similarity. pp. 469-472.

Spaeth, A. & Desmarais, M. C., 2013. *Combining Collaborative Filtering and Text Similarity for Expert Profile Recommendations in Social Websites.* Canada, s.n.

Tarigan, H. G., 2011. *Pengajaran Analisis Kesalahan Berbahasa.* Bandung: Angkasa.

Thelwall, M. & Kousha, K., 2016. Goodreads: A Social Network Site for Book Readers. *Journal of the Association for Information Science and Technology*, pp. 1-12.

Wahyuni, R. T., Prastiyanto, D. & Supraptono, E., 2017. Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi. *Jurnal Teknik Elektro*, Volume 9, pp. 18-23.

Wang, M.-j. & Han, J.-t., 2012. Collaborative Filtering Recommendation Based on Item Rating and Characteristic Information Prediction. pp. 214-217.