Prediksi *Retweet* Berbasis Fitur *Content Similarity* dan *Content Based* Dengan Menggunakan Metode *Support Vector Machine* (SVM)

Rafi Hafizhni Anggia[1], Jondri[2], Kemas Muslim L[3]

[1,2,3] Universitas Telkom, Bandung
rafihafizhnianggia@students.telkomuniversity.ac.id[1], jondri@telkomuniversity.ac.id[2],
kemasmuslim@telkomuniversity.ac.id[3]

Abstrak

Teknologi informasi berkembang sangat pesat sehingga membantu kebutuhan manusia untuk mendapatkan sarana informasi dan komunikasi. Didalam media sosial seperti *twitter* sangat mudah sekali untuk mendapatkan informasi terkini seperti isu politik, kesehatan dan lainnya. Salah satu fitur penyebarannya yaitu *retweet* maka informasi akan cepat berpindah dari penguna satu ke pengguna lainnya. Penelitian ini berupaya untuk membangun sebuah sistem prediksi *retweet* dari isi konten pengguna dengan berbasis fitur *Content Similarity* dan *Content Based* menggunakan metode klasifikasi *Support Vector Machine*. Pembagian *dataset* menggunakan *k fold cross validation* dengan nilai k=10. Hasil akhir pada penelitian ini mendapatkan nilai akurasi rata rata sebesar 71.84%.

Kata kunci : Twitter, Retweet, Content Similarity, Support Vector Machine

Abstract

Information technology developes very rapidly so that it helps human needs to obtain information and communication facilities. In social media such as Twitter, it is very easy to obtain the latest information such as Political, Health and other issues, with one of the distribution features, called retweet, information will be very easy to move from one user to another. This research seeks to build a retweet prediction system from user content based on Content Similarity and Content Based features using the Support Vector Machine classification method. To split the dataset used k fold cross validation with value k=10. The final results in this research got an accuracy average value of 71.84%.

Keywords: Twitter, Retweet, Content Similarity, Support Vector Machine

1. Pendahuluan

Latar Belakang

Perkembangan teknologi informasi pada zaman ini merupakan suatu hal yang sangat cepat dalam perkembangannya, teknologi sangatlah penting karena merupakan kebutuhan manusia sebagai sarana untuk mendapatkan informasi, komunikasi dan juga kemudahan dalam memberikan suatu pendapat yang ingin dikemukakan pada sebuah *platform* seperti media sosial, maka dari kebiasaan manusia saat ini selalu berganting dan terbiasa untuk mengakses media sosial. Salah satu *platform* terbesar pada zaman ini adalah twitter dengan berbasis *microblog* [1], salah satu fitur penyebaran informasi twitter adalah *retweet* yaitu pengguna membuat sebuah postingan yang disebut *tweet* dan di posting ulang oleh pengguna lainnya [2] *retweet* dapat menyebar karena disukai oleh pengguna lainya [3]. Perilaku *retweet* terjadi karena seorang pengguna mengamati isi dari konten postingan kemudian tertarik dengan isi kontennya sehingga ada keinginan untuk memposting kembali. Proses *retweet* ini akan berlangsung sampai pengguna lain tidak menyebarkannya lagi. Pada perilaku ini menjadi bahan tinjauan yang menarik untuk diamati [4].

Penyebaran informasi menjadi suatu keunggulan yang dimanfaatkan untuk kepentingan suatu pihak sehingga opini dibuat berdasarkan keinginan dan pengaruh pihak tersebut [5]. Maka dengan adanya suatu data tersebut informasi biasanya dijadikan sebuah penelitian sehingga dapat dianalisis untuk tujuan membuat suatu bisnis yang dapat menguntungkan [2]. Metode ini juga dapat memahami suatu pasar saham atau dapat juga digunakan untuk mendapatkan dukungan pada saat pemilihan dan perancanangan dalam memutuskan prediksi suatu kebijakan [2], namun dari segala keunggulan yang dimiliki oleh penyebaran informasi ada suatu masalah penting karena dengan mudahnya suatu pihak memberikan opini kepada penerima informasi atau pengguna media sosial yang dapat memberikan suatu opini yang salah dan opini tersebut menyebar, misalkan penyalah gunaannya seperti isu mendalam seperti isu agama[4].

Penelitian didalam twitter dalam memprediksi sebuah retweet telah banyak dilakukan seperti pada penelitian [3][4]. Didalam peneltian sebelumnya terdapat fitur seperti *content similarity*, *user similarity* dan *random walks* dengan menerapkan metode klasifikasi *machine leaning* seperti SVM, *Random Forest*, *Naïve Bayes* dan *Logistik regression*. Hasil dari penelitian tersebut pada dataset *atrocities* mendapatkan nilai terbesar dengan menggunakan SVM[4]. Adapun penelitian lainnya yang serupa prediksi retweet dengan menggunakan tiga fitur yaitu berbasis pengguna, berbasis konten dan berbasis waktu dengan menggunakan metode klasifikasi *random forest* [3].Maka dari penelitian tersebut penulis tertarik melakukan penelitian pada twitter untuk memprediksi sebuah *retweet* bagaimana seorang pengguna twitter dapat melakukan sebuah *retweet* dengan sebuah topik tertentu yang sedang berkembang dan banyak diperbincangkan. Maka disini penulis akan membangun sebuah sistem prediksi *retweet* menggunakan metode klasifikasi *Support Vector Machine* dengan menggunakan fitur *Content Similarity* dan *Content Based* yang dapat dijadikan faktor dalam memprediksi sebuah tweet akan di retweet atau tidak. Pemilihan *Support Vector Machine* karena hasil dari pada akurasi penelitian sebelumnya yang menghasilkan hasilkan akurasi yang baik ketika metode klasifikasi ini bekerja [4].

Topik dan Batasan

Berdasarkan apa yang telah disampaikan pada latar belakang penelitian tugas akhir ini. Maka rumusan masalah yang diambil adalah untuk mengetahui penyebaran informasi yang sedang berkembang dengan memprediksi isi konten tweet pengguna apakah akan di *retweet* oleh pengguna lainnya atau tidak,kemudian melihat *performance* dan akurasi dari metode yang telah dipilih yaitu *SVM* dalam memprediksi *retweet*. Batasan masalah yang terdapat dalam penelitian tugas akhir ini berupa dataset untuk fitur *content similarity* yang hanya mengambil 5 tweet berbeda dari setiap pengguna untuk menggambar isi konten keseluruhan.

Tujuan

Penelitian pada tugas akhir ini memiliki tujuan membangun sistem untuk dapat memprediksi *Retweet* berbasis fitur *Content Similariy* dan *Content Based Twitter* dengan menggunakan metode *Support Vector Machine (SVM).* .

Organisasi Tulisan

Pada tulisan selanjutnya akan menjelaskan mengenai studi yang terkait dalam penelitian yang akan dibuat berupa penelitian penelitian yang serupa. Lalu pada bagian selanjutnya akan menjelaskan keseluruhan gambaran sistem dari penelitian ini dengan teori yang mendasar dalam pembuatan sistem yang akan dibangun.kemudian bagian evaluasi berisikan tentang penjelasan dari hasil penelitian yang telah dilakukan berupa pengujian nilai dari hasil SVM yang didapatkan berupa akurasi. Selanjutnya penjelasan mengenai analisis dari hasil pengujian. Bagian yang terakhir dari penelitian ini adalah sebuah kesimpulan dan saran dari penulis untuk penelitian selanjutnya.

2. Studi Terkait

Penelitian yang dilakukan oleh Syeda Nadia Firdaus, Chen Ding, Alireza Sadeghian menjelaskan cara memprediksi sebuah tweet dan bagaimana seseorang itu akan meretweet atau memposting ulang sebuah tweet, kemudian memahami sebuah informasi itu bisa tersebar dengan inti penelitian ditujukan untuk memberikan sebuah gambaran tentang prediksi sebuah *retweet.* Didalam penelitian ini menggunakan fitur *Author of the tweet, User of the tweet, Content of the tweet* didalam penelitian menggunakan metode rumus seperti *cosine similarity* dan *jaccard similarity* untuk mendapatkan fitur yang akan dibuat.data dibagi menjadi data train dan data train dan memasukan data tersebut ke dalam klasifikasi mengguanakan SVM linear dan Random Forest untuk mencari performa yang paling baik. [2].

Penelitian yang dilakukan oleh T. B. N. H. J. Mothe yang memprediksi sebuah difusi informasi dengan menggunakan data dari sebuah twitter, fitur yang digunakan pada penelitian ini yaitu berupa user based, time based dan content based data yang digunakan pada penelitian ini menggunakan dataset Sandy *FirstWeek SecondWeek* yang akan di bagi menjadi kelas 0 sampai 3.. Pada proses klasifikasi data menggunakan *Binary classification* dan *Multi-class classification* dan untuk mendapatkan nilai yang terbaik maka pada penelitian ini memakai berbagai metode klasifikasi seperti *Naive Baiyes* (NB), *Support Vector Machine* (SMO) and *Random Forest* (RF).[3]

Penelitian yang dilakukan oleh Nidhi Singha, Anurag Singha, Rajesh Sharma yang menjelaskan bagaimana cara memprediksi sebuah *information cascade* bagaimana sebuah tweet di *retweet* dan pada metode yang digunakan untuk solusi permasalahan pada kasus ini dengan menggunakan fitur *content similarity* dan *user similarity*. dataset yang digunakan ada 3 yaitu *Atrocites, Oscar* dan *GST* Untuk mendapatkan hasil prediksi yang lebih tinggi pada penelitian ini menggunakan metode *Random Walks* yaitu sebuah proses statistika yang dapat memprediksi sebuah probabilitas. Setelah mendapatkan fitur fitur, data akan dibagi menjadi data train dan data

test maka dimasukan kedalam klasifikasi seperti *Logical Regression* (LR) *klasifikasi Naive Baiyes* (NB), *Support Vector Machine* (SVM) dan *Random Forest* (RF). Menggunakan berbagai klasifikasi bertujuan untuk membandingkan klasifikasi satu dengan yang lainnya. Dari hasil penelitian akurasi paling tinggi pada dataset *atrocities* adalah SVM.[4]

2.1. Twitter

Media social twitter merupakan salah satu bentuk dari pada perkembangan zaman tentang teknologi informasi yang berbasis sebuah web dan berbasis internet, bertujuan dibuatnya media social ini adalah untuk interaksi antara penggunanya sehingga mereka akan saling terhubung satu sama lain dengan sebuah jaringan secara *online*. Pengguna twitter dapat memposting suatu *tweet* dan postingan akan dilihat oleh pengguna lainnya[7]. Twitter adalah suatu situs *microblogging* yang dapat dikategori sebagai sosial media. Pada twitter terdapat sebuah konten tweet berisikan postingan pengguna berupa kalimat dengan panjang maksimal 280 karakter pada satu postingannya, karena *tweet* itu sendiri berupa pesan singkat. Didalam twitter jugaa terdapat kata kunci yang dapat mengelompokan suatu postingan, nama dari fitur tersebut adalah sebuah # (*hashtag*) [7].

Twitter dapat dimanfaatkan untuk keperluan pribadi sebagai sarana berkomukasi atau digunakan untuk berbisnis seperti halnya media promosi sebuah produk. Media sosial ini juga dimanfaatkan sebagai jejaring sosial yang sangat efisien dan efektif.[8] Fitur dalam twitter yang sering digunakan oleh pengguna yaitu *retweet* merupakan suatu perilaku pengguna dengan cara memposting ulang suatu pesan atau tweet yang mereka anggap suka, fitur ini adalah bentuk mekanisme penyebaran informasi sangat popular dan hanya terdapat dalam dalam twitter, jika penyebarannya meluas dalam kurun waktu yang cepat maka kemungkinan besar informasi tersebut akan tersebar dan semakin besar pula topik yang dibicarakan maka informasi tersebut bisa menjadi trending topik. [9]. Didalam twitter juga dapat mengambil suatu data dengan cara *crawling* data dibantu dengan API Twitter yang dapat memberikan akses, Data biasa digunakan untuk sebuah penelitian[10].

2.2. *Support Vector Machine*

Support Vector Machine adalah algoritma klasifikasi berisikan data yang berpola pada inputannya lalu data tersebut akan diolah kedalam bentuk vektor dengan metode pada SVM ini maka *hyperplane* dapat dicari dengan *hyperplane* paling baik . *Support Vector Machine* merupakan algoritma *supervised* yang dapat mengetahui perbedaan dari tipe data berupa kelas kelas. *Support vector* merupakan dua data yang memiliki jarak terdekat dan terdapat sebuah garis pemisah antara dua kelas yang disebut dengan *hyperplane*, pada metode SVM juga memaksimalkan jarak antara *support vector* dan *hyperplane* yang disebut margin[11]. Pada proses klasifikasi biasanya terdapat data *linear* dan *non linear* maka jika terdapat data non linear dapat menggunakan sebuah metode yang disebut kernel *trick* seperti kernel *RBF*,*Linear* sigmoid dan polinom.[12] Fungsi kernel digunakan untuk mengubah proses permodelan SVM linier ke non liniear. Kernel RBF adalah suatu kernel yang dapat digunakan untuk segala jenis data. Rumus dari perhitungan kernel *RBF* ini sebagai berikut [13].

$$K(x, x^1) = \exp\left(-\frac{||x-x^1||^2}{2\sigma^2}\right) \qquad \textbf{(2.1)}$$

2.3.Penggunaan Fitur

Dalam penelitian ini untuk mendapatkan hasil prediksi sebuah *retweet* maka harus adanya sebuah fitur yang dapat membantu atau mempengaruhi sebuah tweet tersebut akan diretweet kembali. Fitur yang akan digunakan dalam penelitian ini yaitu *Content similarity* dan *Content Based* :

Content Similarity

Content Similarity merupakan suatu kesamaan tweet dengan membandingkan isi konten dari pengguna satu dengan pengguna lainnya. Untuk mendapatkan fitur ini menggunakan 2 metode yaitu *TF IDF* untuk mengetahui nilai dari bobot suatu kata. Pembobotan kata dari *TF IDF* sangat efisien, mudah serta dapat memberikan akurasi yang tinggi[14] dan metode *Cosine Similarity* ini digunakan sebagai rumus perhitungannya sehingga mendapatkan nilai kesamaan antara 2 buah tweet yang dibandingkan sehingga dapat menentukan sebuah tweet tersebut akan diretweet atau tidak. *Cosine Similarity* menghasilkan akurasi yang tinggi[4]. Rumus dari perhitungan fitur ini sebagai berikut :[4]

$$Cs_{f1f2} = \frac{\langle \widehat{t1}.\widehat{t2} \rangle}{\|f1\|\|f2\|} \qquad \textbf{(2.2)}$$

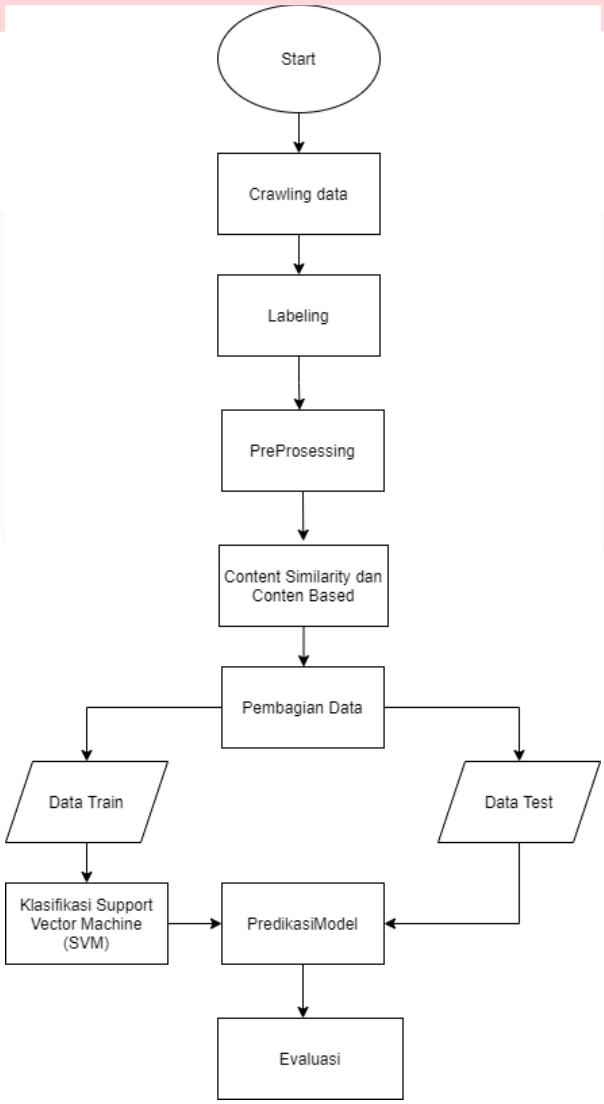
Content Based

Content Based merupakan sebuah fitur yang memiliki sebuah informasi penting dari sebuah *tweet* yang dapat menjadikan daya tarik untuk pengguna lainnya untuk menyebarkan informasi tersebut [15]. Maka dari itu fitur content based ini menggunakan data data sebagai berikut :

- *Hastag* merupakan sebuah kata kunci tertentu yang dibiasa digunakan dalam sebuah *tweet* untuk menunujukan kategori.
- *URL* : Merupakan link yang dapat memberikan sebuah informasi sumber dari sebuah tweet.
- *Length of text*: Merupakan jumlah panjang kata dari sebauh tweet pengguna.
- *Media* : Terdapatnya sebuah foto atau video dalam suatu tweet agar dapat memberikan informasi yang jelas.
- *Mention* : Merupakan tindakan seorang *user* yang menuliskan *username* pengguna lainnya dari sebuah tweet.

3. Gambaran sistem yang Dibangun

Perancangan sistem menggambarkan proses penelitian tugas akhir berbentuk *flowchart*, untuk mengetahui bagaimana tahapan yang akan dibangun. Gambar 2 merupakan gambaran umum sistem



Gambar 1. Alur sistem yang dibangun

1. *Crawling* data
   Tahapan *crawling* ini bertujuan untuk mendapatkan 2 tipe user yang pertama user yang membuat tweet dan user yang dianggap akan menyebarkan tweet atau meretweet. Dalam pengambilan data dilakukan dengan dua tahapan.
   Tahap pertama dilakukan dengan menggunakan aplikasi netlytic agar data dapat membuat sebuah network sebuah node dan edges berasal dari hubungan user tersebut karena sebuah retweet.
   Tahap kedua dilakukan dengan cara crawling pada twitter. Hal ini dilakukan untuk mendapatkan suatu isi tweet yang berbeda dari user yang akan digunakan untuk fitur *Content Similarity* dengan cara mengambil 5 tweet yang berbeda dari setiap user kemudian data akan dibandingkan dengan konten user yang pernah di *retweet* dan tidak di *retweet*.
2. Pelabelan data
   Pada tahap ini dilakukan pemberian label dengan melihat isi tweet tersebut apakah di *retweet* atau tidak. Jika di *retweet* makan nilai labelnya adalah 1 kalau tidak maka labelnya adalah 0.
3. *Preprossesing*
   Tahap preprosessing dilakukan untuk menghilangkan suatu atribut yang tidak relavan, pada tahap ini menggunakan beberapa proses seperti *stemming, tokenizing, cleaning, case folding* dan *stopword*.
4. Fitur Penelitian
   Content similarity yang digunakan untuk membandingkan isi konten pengguna didalam fitur content similarity menggunakan dua metode yaitu *TF IDF* guna untuk mendapatkan bobot dari suatu kata dan cosine similarity untuk perhitungan dalam membandingkan kesamaan dari isi tweet pengguna. Langkah yang diterapkan untuk mendapatkan hasil dari *content similarity* dengan cara membandingkan 5 tweet berbeda dari pengguna dengan 1 tweet yang akan di cek kesamaannya kemudian hasil tersebut akan dibandingkan dan hasil akhirnya di rata ratakan. Data yang dibandingkan dari tweet hubungan user. Pengambilan data 5 tweet yang berbeda karena untuk mendapat validasi dari isi konten penggunanya memiliki kemiripan lebih tinggi, jika data hanya 1 maka masih berupa asumsi dan kemungkinannya kemiripannya lebih kecil. Contoh dari perbandingan data seperti pada tabel berikut :

Tabel 1. Perbandingan tweet fitur *content similarity*

Username	Tweet	Username	Tweet	Hasil *content similarity*
A	Tweet A ke 1	B	Tweet B	Numerik
A	Tweet A ke 2	B	Tweet B	Numerik
A	Tweet A ke 3	B	Tweet B	Numerik
A	Tweet A ke 4	B	Tweet B	Numerik
A	Tweet A ke 5	B	Tweet B	Numerik
			Hasil rata rata	Numerik

   Content based digunakan untuk menarik penggguna dengan isi konten tweet orang tersebut seperti hastag,url,media dan lainnya. Fitur fitur tersebut dapat membantu untuk memprediksi apakah sebuah tweet tersebut akan di diretweet atau tidak. Data fitur ini diambil dengan cara pada penelitian [3].
5. Pembagian data
   Pada tahap pembagian pada peneltian ini dibagi ke data test dan train,menggunakan *k fold cross validation* dengan nilai k = 10 untuk menguji validasi dari metode klasifikasi *Support Vector Machine* dalam memprediksi retweet. 10 *fold validation* sangat dianjurkan karena dapat memberikan nilai uji validasi terbaik jika dibandingkan dengan nilai k lainnya.[16]
6. Klasifikasi SVM
   Pada tahapan ini digunakan untuk mendapatkan hasil akurasi yang didapatkan dari hasil klasifikasi telah digunakan yaitu menggunakan SVM.
7. Evaluasi

Evaluasi berisikan hasil berupa data pengujian dan analisis terhadap sistem yang telah dibangun

4. Evaluasi

Bagian ini berisikan hasil daripada pengujian kemudian analisis yang telah dilakukan pada sistem prediksi *retweet* yang telah dibangun berdasarkan metode yang terlah dipilih yaitu dengan menggunakan klasifikasi SVM. Pada pengujian dan analisis dilakukan dengan cara yang terdapat pada tujuan dan pendahuluan.

4.1 Dataset

Dataset yang digunakan untuk penelitian sebesar 1200 hubungan antar *user* dengan 6000 tweet yang akan dibandingkan pada fitur *content similarity* dan data akhir yang dijadikan sebagai model klasifikasi sebesar 1200. Data data tersebut diambil dengan cara *crawling* dari data netlytic sebagai data utama dan juga *crawling* dari twitter yang digunakan untuk kebutuhan fitur *content similarity*.

Tabel 2. Atribut dataset dari fitur *content similarity* dan *content based*

	Atribut	Keterangan	Tipe data
Content Similarity	Tweet	Membandingkan isi tweet pengguna satu dengan pengguna lain.	Numerik
Content Based	Hastag	Berisikan sebuah hastag dari sebuah tweet	Boolean
	Url	Berisikan URL dari sebuah *tweet*	Boolean
	length of text	Panjang dari kalimat suatu *tweet*	Numerik
	Mention	Merupakan mention dari pengguna yang lain	Boolean
	Media	*Tweet* berisikan sebuah gambar atau video	Boolean

4.2 Hasil Pengujian

Pengujian data dilakukan dengan cara membagi kedalam data test dan data train yang akan dilakukan dengan nilai k = 10 menggunakan sebuah metode *k fold cross validation*. Pengujian ini dilakukan agar mengetahui sebuah prediksi *retweet* dengan klasifikasi yang dipilih yaitu *Support Vector Machine* (SVM). Dari metode tersebut maka akan mendapatkan sebuah nilai akurasi. Berikut adalah hasil dari percobaan metode klasfikasi *Support Vector Machine* dari berbagai kernel yang telah dilakukan :

Kernel RBF

Tabel 3. Hasil akurasi dan rata rata akurasi dari kernel RBF

Fold	Akurasi
Fold 1	75.00%
Fold 2	73.33%
Fold 3	71.39%
Fold 4	72.08%
Fold 5	71.33%
Fold 6	72.08%
Fold 7	71.67%
Fold 8	71.04%
Fold 9	70.83%
Fold 10	69.67%
Rata-rata	71.84%

Berdasarkan informasi pada tabel dengan pengujian kernel *RBF* (*Radial Basis Function*) pada *Support Vector Machine* menggunakan *k fold validation* dengan nilai k = 10. Hasil dari pengujian yang telah dilakukan untuk mendapatkan prediksi *retweet* mendapatkan nilai akurasi rata rata sebesar 71.84.%.

Kernel *Linear SVC*

Tabel 4. Hasil akurasi dan rata rata akurasi dari kernel *Linear*

Fold	Akurasi
Fold 1	65.00%
Fold 2	64.58%
Fold 3	63.06%
Fold 4	63.12%
Fold 5	62.67%
Fold 6	63.61%
Fold 7	62.62%
Fold 8	61.88%
Fold 9	60.46%
Fold 10	60.08%
Rata-rata	62.71%

Berdasarkan informasi pada tabel dengan pengujian kernel *Linear* pada *Support Vector Machine* (SVM) menggunakan *k fold validation* dengan nilai k = 10. Hasil dari pengujian yang telah dilakukan untuk mendapatkan prediksi *retweet* mendapatkan nilai akurasi rata rata sebesar 62.71%.

Kernel *Polynomial*

Tabel 5. Hasil akurasi dan rata rata akurasi dari kernel polynomial

Fold	Akurasi
Fold 1	66.67%
Fold 2	67.08%
Fold 3	65.00%
Fold 4	65.00%
Fold 5	65.33%
Fold 6	65.83%
Fold 7	64.64%
Fold 8	63.54%
Fold 9	62.04%
Fold 10	61.58%
Rata-rata	64.67%

Berdasarkan informasi pada tabel dengan pengujian kernel *polynomial* untuk mendapatkan hasil yang maksimal melakukan percobaan pada degree 1 sampai 10, setelah melakukan percobaan maka mendapatkan nilai degree terbaik dengan nilai 2. Pada pembagian data dilakukan sama dengan kernel lainnya dengan menggunakan *k fold validation* dengan nilai k = 10. Hasil dari pengujian yang telah dilakukan untuk mendapatkan prediksi *retweet* mendapatkan nilai akurasi rata rata sebesar 64.67%.

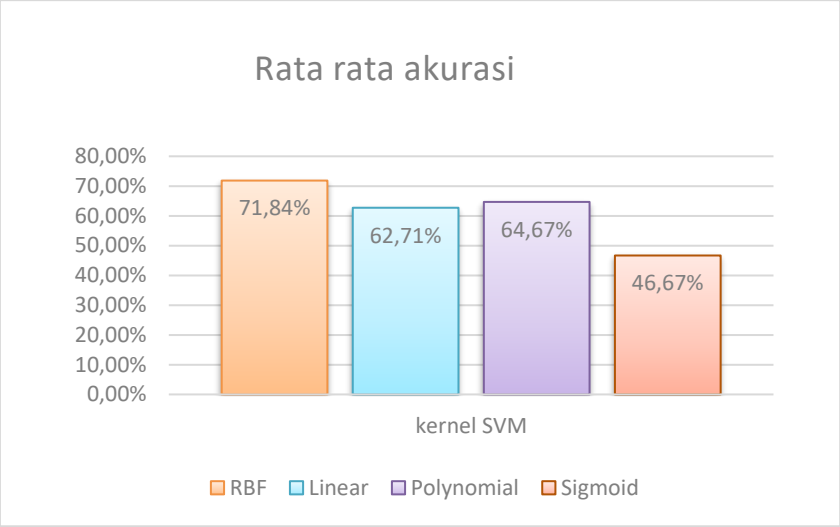
Kernel *Sigmoid*

Tabel 6. Hasil akurasi dan rata rata akurasi dari kernel *sigmoid*

Fold	Akurasi
Fold 1	50.83%
Fold 2	45.83%
Fold 3	47.22%
Fold 4	47.29%
Fold 5	47.33%
Fold 6	46.25%
Fold 7	45.95%
Fold 8	45.52%
Fold 9	45.09%
Fold 10	45.42%
Rata-rata	46.67%

Berdasarkan informasi pada tabel dengan pengujian kernel *sigmoid* pada *Support Vector Machine* menggunakan *k fold validation* dengan nilai k = 10. Hasil dari pengujian yang telah dilakukan untuk mendapatkan prediksi retweet mendapatkan nilai akurasi rata rata sebesar 46.67%.

Berdasarkan hasil yang telah didapatkan pada tabel tabel diatas nilai akurasi rata rata dari berbagai kernel *Support Vector Machine* seperti RBF (*Radial Basis Function*), *Linear*, *Polynomial* dan *Sigmod* dapat digambarkan sebagai berikut :



Gambar 2. Perbandingan rata rata akurasi dari berbagai kernel yang digunakan

4.3 Analisis Hasil Pengujian

Hasil dari pada penelitian data dibagi dengan *k fold cros validation* dengan menggunakan nilai sebesar 10, perbandingan ratio 9 untuk data train dan 1 untuk data test. Dalam memprediksi sebuah retweet dengan menggunakan metode *Support Vektor Machine* (SVM) dari berbagai kernel mendapatkan nilai performa terbaik pada kernel *RBF (Radial Basic Funtion)* mendapatkan nilai akurasi rata rata sebesar 71.84% kemudian selanjutnya pada kernel *polynomial* memakai degree terbaik mendapatkan nilai akurasi 64.67%, pada kernel Linear mendapatkan nilai akurasi rata rata sebesar 62.71% sedangkan kernel yang memiliki nilai akurasi yang paling rendah adalah kernel *Sigmoid* mendapatkan nilai akurasi rata rata sebesar 46.67%. *RBF* merupakan kernel terbaik karena *hyperplane* yang digunakan sangat cocok untuk data pada penelitian ini, Tingginya akurasi dikarenakan data uji dapat dibaca dengan benar oleh *machine* sehingga dapat menentukan nilai akhir akurasi yang baik sedangkan untuk data uji yang salah pada klasifikasi pada SVM biasanya ditandai dengan masuk suatu data ke dalam kelas yang berbeda yang dibatasi oleh *hyperplane* sehingga data tersebut dapat dianggap *noise* oleh *machine* dan tidak dapat dibaca dengan baik.

Berdasarkan hasil yang telah dilakukan pada penelitian terdapat kekurangan yang sangat mempengaruhi hasil prediksi *retweet* pada metode *support vector machine* (SVM) yaitu pada fitur *content similarity* karena kesulitan dalam mendapatkan data seperti akun pengguna yang tiba tiba ngubah akunnya menjadi *protected*,akun tidak ditemukan, akun berganti nama atau pengguna sudah menghapus isi tweet, sehingga untuk mendaparkan dataset memakan waktu yang cukup lama. Pada peneltian untuk fitur *content similarity* hanya dapat mengambil 5 isi tweet berbeda yang dapat mewakili isi konten dari setiap pengguna. Dataset pada fitur *content similarity* yang akan dibandingkan adalah sebesar 6000 tweet dan untuk mendapatkan hasil akhir dari fitur content similarity maka di rata ratakan sehingga mendapatkan data yang akan dijadikan bahan model klasifikasi adalah 1200 data.

5. Kesimpulan

Penelitian sistem dapat membangun suatu sistem prediksi dari pengguna twitter apakah *tweet* dengan dipengaruhi isi konten akan di *retweet* atau tidak.Penelitian menggunakan fitur seperti *content similarity* yang bertujuan untuk menbanding isi konten pengguna dengan pengguna lainnya dan fitur yang dapat menarik pengguna yaitu *content based* seperti *hastag*, *url* dan *mention.* Untuk dapat mendapatkan predikasi yang paling baik penelitian telah melakukan dengan berbagai kernel yang terdapat dalam *support vector machine* (SVM).Nilai paling tinggi pada penelitian ini yaitu memakai kernel RBF *(Radial Basis Function)* mendapatkan nilai akurasi rata rata sebesar 71.84%. Dari hasil rata rata tersebut penelitian ini mendapatkan performa yang dapat dikatakan

kurang memuaskan bagi peneliti yang disebabkan oleh faktor kesulitan dalam mendapatkan dataset pada fitur *content similarity*. Karena pada saat mengambil data *tweet* terdapat beberapa pengguna yang *protected*, ganti nama akun, isi tweet yang dihapus oleh pengguna sehingga harus kembali diganti lagi dengan data pengguna lainnya. Data *sample* yang digunakan pada fitur *content similarity* hanya mengambil 5 tweet yang berbeda dari pengguna yang mewakili seluruh isi konten pengguna.

Saran yang diberikan oleh penulis untuk penelitian selanjutnya mempersiapkan data dan waktu yang banyak untuk fitur *content similarity* karena terdapat banyak kendala dalam memaksimalkan pengambilan data untuk perbandingan isi konten pengguna satu ke pengguna lainnya. Dengan cara tersebut dipastikan dapat meningkatkan hasil nilai performa prediksi *retweet* akan lebih besar.

REFERENSI

[1] E. S. Pandu Adi Cakranegara, "Analisis Strategi Implementasi Media Sosia Studi Kasus UKM "XYZ"," Journal President, pp. 1-16, 2019.

[2] C. D. A. S. Syeda Nadia Firdaus, Retweet: A popular information diffusion mechanism – A survey paper, Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada: Department of Computer Science, 2018.

[3] T. B. N. H. J. Mothe, "Predicting Information Diffusion on Twitter - Analysis of predictive features," Computational Science, pp. 1-11, 2017.

[4] A. S. ,. R. S. Nidhi Singha, "Predicting Information Cascade on Twitter Using Random Walk," Procedia Computer Science, vol. 173, pp. 201-209, 2020.

[5] H. a. M. G. Allcott, "Social Media and Fake News in the 2016 Election," Journal of Economic Perspectives, vol. 31, no. 2, pp. 211-236, 2017.

[6] M. S. M. A. Muhammad Hilman Aprilian Nurjaman, "Analisis sentimen pada ulasan buku berbahasa inggris menggunakan Information Gain dan Suport Vector Machine," e-Proceeding of Engineering , vol. 4, no. 3, p. 4900, 2017.

[7] O. Z. Tane, "Analisis Sentimen pada Twitter Tentang Calon Presiden 2019 Menggunakan Metode SVM," Telkom University, 2019.

[8] E. W. Arief Wibowo, "Paper Review: Data Mining Twitter," ResearchGate, 2018.

[9] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," WWW 2013 Companion - Proc. 22nd Int. Conf. World Wide Web, pp. 657–664, 2013, doi: 10.1145/2487788.2488017.

[10] E. B. S. Z. A. B. Jaka Eka Sembodo, "Data Crawling Otomatis pada Twitter," Researchgate, p. 12, 2

[11] E. B. S. Eias Raihandtsa Mamuri, "Mendeteksi Pesan Berita Palsu (Hoax) pada Twitter dengan Algoritma AdaBoost dan ANP," Universitas Telkom, p. 1, 2019.

[12] Neneng Rachmalia Feta, Asep Rahmat Ginanjar "Kompirasi fungsi kernel metode Support Vector Machine untuk permodelan klasifikasi terhadap penyakit tenaman kedelai", Institut Teknologi dan Bisnis Bank Rakyat Indonesia, 2019

[13] Rarasmaya Indraswari, Agus Zainal Arifin, Darlis Herumurti "RBF kernel optimization method with particle swarm optimization on SVM using the analyisis of input data's movement", Institut Teknologi Sepuluh Nopember

[14] Ria Melita, Victor Amrizal, Hendra Bayu Suseno, Taslimun Dirjam "Penerapan Metode (TF-IDF) DAN Cosine Similarity Pada system temu kembali informasi untuk mengetahui syarag hadist berbasis Web" Universitas Islam Negeri Syarif Hidayatullah Jakarta, 2018

[15] Z. Xu, Q. Yang, "Analyzing user retweet behavior on twitter," IEE ASONAM 2012

[16] Ron Kohavi " A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection" Stanford University