Prediksi retweet berdasarkan feature user-based menggunakan metode klasifikasi *Support Vector Machine*

Rakes[1], Jondri[2], Kemas Muslim Lhaksamana[3]

[1,2,3] Universitas Telkom, Bandung

[1]rakesh@students.telkomuniversity.ac.id, [2]jondri@telkomuniversity.ac.id,

[3]kemasmuslim@telkomuniversity.ac.id

Abstrak

Ada banyak sosial media yang digunakan masyarakat dalam menyebarkan informasi secara cepat. Twitter merupakan salah satu unggulan dalam kategori ini, yang mana hanya dengan menekan *retweet* informasi dapat kita teruskan. Akan tetapi bukan berarti apapun informasi yang kita punya orang akan tertarik menyebarkannya. Penelitian ini bertujuan untuk menghasilkan program prediksi *Retweetability* sebuah tweet dan mengamati performansi dan akurasi dari machine learning *Support Vector Machine* dengan menjadikan feature user-based sebagai atribut. Metode k-fold cross validation digunakan setelah preprocessing data. Penelitian berhasil menghasilkan algoritma yang dapat memprediksi *Retweetability* sebuah tweet dengan f1-score sebesar 66,05%.

Kata kunci : Twitter, tweet, retweet, *retweetability*, *Support Vector Machine*, k-fold cross validation

Abstract

There are many social media that people use to spread information quickly. Twitter is one of the best in this category, which only by pressing retweet we can pass the information. But that doesn't mean for any information we have people will be interested on spreading it. This study aims to produce a Retweetability prediction program for a tweet and observe the performance and accuracy of machine learning Support Vector Machine by using user-based features as a model. The k-fold cross validation method is used after preprocessing the data. The research has succeeded in producing an algorithm that can predict the retweetability of a tweet with f1-score 66,05%.

Keywords: Twitter, tweet, retweet, retweetability, Support Vector Machine, k-fold cross validation

1. Pendahuluan

Latar Belakang

Di antara berbagai sistem *microblogging*, Twitter adalah layanan yang paling populer sejauh ini. Karena kemudahannya dalam berbagi informasi secara real-time, Twitter banyak mempengaruhi wacana publik di masyarakat. Di Twitter, banyak informasi dibagikan melalui jejaring sosialnya yang terstruktur, akan tetapi hanya sedikit yang mengetahui tentang bagaimana dan mengapa informasi tertentu menyebar lebih luas daripada yang lain (Bongwon Suh, 2010).

Secara struktural, retweet sendiri merupakan fitur Twitter yang setara dengan fitur *forward email* yang mana pengguna memposting pesan yang awalnya diposting oleh orang lain[1]. Melakukan *retweet* membawa orang baru kedalam sebuah topik tertentu, mengundang mereka untuk terlibat tanpa mengundang mereka secara langsung[1].

Pada paper acuan penulis beranggapan bahwa *Retweet* adalah mekanisme kunci untuk penyebaran informasi di Twitter. Peneliti sebelumnya percaya akan pentingnya untuk mengeksplorasi bagaimana retweet berfungsi untuk memahami bagaimana informasi disebarkan di Twitter jaringan dan untuk memahami mengapa tweet tertentu menyebar lebih banyak luas dari yang lain[2]. Maka dari itu peneliti sebelumnya gunakan 74 Juta tweet pada paper acuan sebagai data untuk mengidentifikasi faktor faktor yang mempengaruhi *retweet rate*. Berdasarkan hasil penelitian terdapat berbagai macam *feature* yang ada dalam data, ditemukan berbagai macam hal menarik yang dapat maupun tidak mempengaruhi *Retweetability* dari sebuah *tweet*[2]. Principal Component Analysis (PCA), Generalized Linear Model (GLM), Regular Expression Method, Feature Retweet Method digunakan penulis dalam penelitiannya untuk membantu mengkonversi data menjadi hasil prediksi[2]. Diharapkan penelitian ini juga dapat membantu mengetahui factor yang mempengaruhi penyebaran informasi, terutama dalam media social twitter.

Dalam penelitian ini akan digunakan *feature user-based* dalam data set hari pemilihan presiden amerika serikat yang berjumlah hampir 400 ribu data oleh Chris Albon dari website kaggle[11]. Akan digunakan 1000 data teratas sebagai data latih dan data uji menggunakan *Support Vector Machine* di bantu oleh k-fold cross validation untuk membangun sistem pada penelitian ini.

Topik dan Batasan

Sesuai dengan apa yang sudah disampaikan pada Latar Belakang, rumusan masalah dalam penelitian ini adalah memprediksi apakah suatu tweet akan di retweet atau tidak berdasarkan *feature user-based* yang dipilih dalam data. Dilihat juga performansi dan akurasi dari *Support Vector Machine* sebagai metode klasifikasi yang dipilih. Batasan masalah dari penelitian ini merupakan sebagian data yang digunakan dan *Feature User-Based* yang dipakai.

Tujuan

Tujuan dari penelitian ini adalah untuk membangun sebuah sistem yang diharapkan dapat memprediksi *retweet* berdasarkan *feature user-based* dengan menggunakan metode *Support Vector Machine*.

Organisasi Tujuan

Akan dijelaskan pada bagian bagian selanjutnya mengenai studi terkait penelitian ini pada bagian kedua. Pada bagian ketiga akan dipaparkan sistem yang dibangun dalam penelitian ini pada dan juga gambaran besar system. Pada bagian keempat evaluasi dari penelitian. Dan yang terakhir merupakan kesimpulan dan juga saran untuk penelitian selanjutnya.

2. Studi Terkait

Penelitian ini berdasarkan penilitian terkait sebagain acuan. Berdasarkan penelitian terkait, dengan judul "*Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network*"[2] yang mana berfokus kepada penemuan fitur fitur yang dianggap berpengaruh terhadap *Retweetability* sebuah tweet. Peneliti sebelumnya menggunakan *content feature-based* yakni *URL, hashtag, mention* lalu *follower, followee/following, favorite, day, status* sebagai *user feature-based*. Penelitian ini mendapatkan kesimpulan yakni *follower, followee/following* sebagai aspek paling berpengaruh pada *user feature-based* sementara untuk *content feature-based* dalam porsi besar dipengaruhi oleh trend.

2.1 Twitter

Twitter adalah platform komunikasi berbasis web yang menggabungkan Pesan Instan dan *SMS* yang memungkinkan pelanggan layanannya untuk mengirim 'pembaruan status' singkat kepada orang lain[3]. Pada awal kemunculannya twitter banyak membawa hal hal baru dalam dunia platform komunikasi. Twitter memperkenalkan *markup culture* untuk fitur postingnya yang kita kenal dengan sebutan *Tweet.* Beberapa diantaranya adalah RT disingkat dari kata *retweet*, '@' yang biasa dilanjutkan dengan *username user* sebagai fitur mention, dan '#' yang diikuti oleh kata – kata yang biasa mempresentasikan *hashtag*[4].

Tingkat Popularitas platform ini meningkat cepat, pada juli 2014 Twitter mencetak 500 juta *tweets* per hari dan 255 juta pengguna aktif bulanan[5]. Alasan dari popularitas ini adalah keunikan twitter, yakni mengizinkan user yang memiliki ketertarikan pada bidang yang sama untuk bertukar pikiran ke seluruh dunia tanpa biaya dan juga sebagai ruang debat topik itu sendiri[6]. Pada bidang Pendidikan dan Komunikasi, layanan ini memegang peranan penting untuk sebagain tempat dan juga media penyebaran kemanjuan penelitian untuk berbagai peneliti dan juga organisasi[7].

2.2 Support Vector Machine

Support Vector Machine (SVM) pertama kali diperkenalkan oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep-konsep unggulan dalam bidang pattern recognition. Sebagai salah satu metode pattern recognition, usia SVM terbilang masih relatif muda. SVM adalah metode learning machine yang bekerja atas prinsip Structural Risk Minimization (SRM) dengan tujuan menemukan hyperplane terbaik yang memisahkan dua buah class pada input space[8].

SVM (Support Vector Machine) digunakan sebagai rangkaian harmonis konsep - konsep unggulan dalam bidang pattern recognition. Konsep dasar SVM merupakan salah satu metode yang terbimbing. Metode terbimbing yang dimaksud adalah metode yang membutuhkan data training dan data testing dalam uji coba[9].

Pada SVM, untuk memisahkan kelas algoritma ini akan berusaha menemukan hyperplane yang terbaik pada input space, Hyperplane pemisah terbaik antara kedua kelas dapat ditemukan dengan mengukur margin hyperplane tsb. dan mencari titik maksimalnya. Margin adalah jarak antara hyperplane tersebut dengan pattern terdekat dari masing-masing class. Pattern yang paling dekat ini disebut sebagai support vector[10].

2.3 Feature User-Based

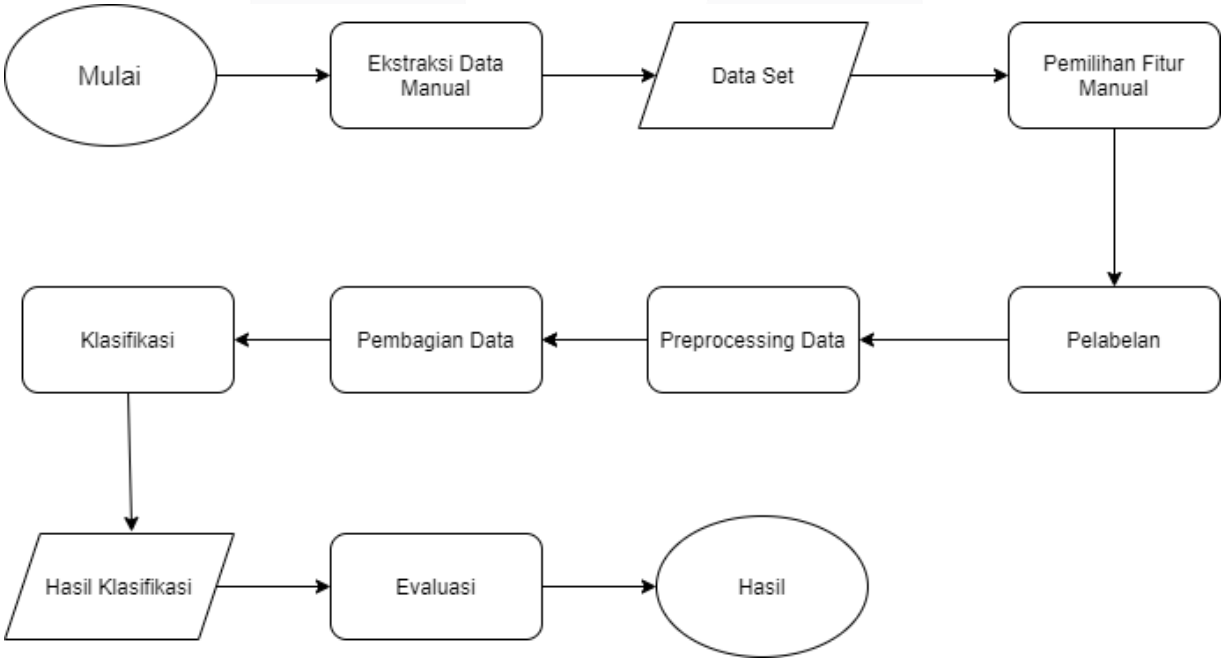
Dalam Penelitian ini akan digunakan fitur yang dianggap termasuk didalam *Feature User-Based* yaitu

Atribut	Deskripsi
User.favourites_count	Jumlah Tweet yang sudah pernah User favoritkan/like pada akun
User.statuses_count	Jumlah Tweet(termasuk retweet) yang sudah pernah diposting oleh user
User.verified	Status akun pengguna apakah verified atau tidak
User.listed_count	Jumlah public list dimana User adalah anggotanya
User.followers_count	Jumlah followers User
User.friends_count	Jumlah Followings User
Retweet_count	Jumlah Retweet pada sebuah Tweet

Tabel 1. Deskripsi fitur

3. Sistem yang Dibangun

Pada bagian ini akan perlihatkan gambaran besar dari rancangan sistem yang dibangun pada penelitian ini.



Gambar 1. Alur Rancangan Sistem

1. Ekstraksi Data

Pada tahapan ini adalah proses ekstraksi data set yang diambil dari tweet pada hari pemilihan presiden amerika serikat tahun 2016 oleh Chris Albon dari website kaggle[11]. Data memiliki 34 fitur dan hampir 400 ribu data

dalam bentuk file csv yang akan direduksi menjadi 1000 data teratas untuk digunakan sebagai data latih dan data uji. Pada tahapan ini juga akan diimport library yang akan digunakan untuk mendukung proses penelitian.

2 Pemilihan Fitur

Dalam proses Pemilihan Fitur, akan dipilih User-Based Feature yang dianggap penting pada penelitian sebelumnya yakni jumlah Followers dan juga jumlah Following. Selanjutnya fitur fitur tersebut akan di kombinasikan dengan fitur fitur baru yang dianggap dapat membantu kinerja dari fitur fitur pada penelitian sebelumnya. Fitur fitur yang ditambahkan antara lain favourites_count, statuses_count, verified, dan juga listed_count.

3 Pelabelan

Pada tahapan ini, kolom retweet yang ada pada data set akan dirubah valuenya. Value 1 berarti tweet di retweet oleh pengguna lain setidaknya satu kali, dan Value 0 yang berarti tweet tidak di retweet.

4 Preprocessing Data

Setelah fitur fitur yang tidak digunakan dieliminasi, akan dilakukan Preprocessing Data pada fitur fitur Dataset yang dipilih. Preprocessing Data yang dilakukan antara lain pengecekan data yang memiliki value Null/NaN, mengubah tipe data beberapa Fitur, dan menentukan jumlah data yang akan digunakan dalam penelitian.

5. Pembagian Data

Proses Pembagian Data akan menggunakan 5-fold cross validation untuk menghasilkan akurasi dari klasifikasi.

6. Klasifikasi.

Pada tahapan ini akan dilakukan perhitungan akurasi, precision, dan f1-score yang berguna untuk mengukur akurasi dan performansi dari klasifikasi yang dilakukan.

4. Evaluasi

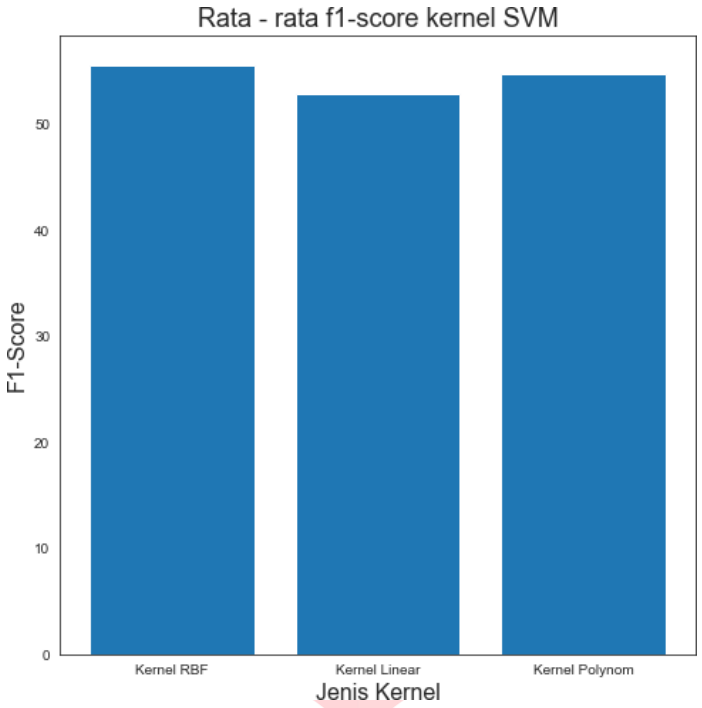
Bagian ini dibagi menjadi dua, Hasil Pengujian dan Analisis Hasil Pengujian dari sistem prediksi. Pengujian dan analisis yang dilakukan selaras dengan tujuan TA sebagaimana dinyatakan dalam Pendahuluan.

4.1 Dataset

Pada penelitian ini dataset diambil dari tweet pada hari pemilihan presiden amerika serikat tahun 2016 oleh Chris Albon dari website Kaggle[11], yang nantinya bagian kecil dari data akan digunakan. Dari hampir dari 400 ribu data akan diambil 1000 tweet yang nantinya akan dilakukan undersampling pada data karena terdeteksi bahwa kelas pada data *imbalanced*.Tujuh atribut yang dianggap termasuk dalam *feature user-based* antara lain User.favourites_count, User.statuses_count, User.verified, User.listed_count, User.followers_count, User.friends_count, dan Retweet_count yang akan dijadikan *class*.

4.2 Hasil Pengujian

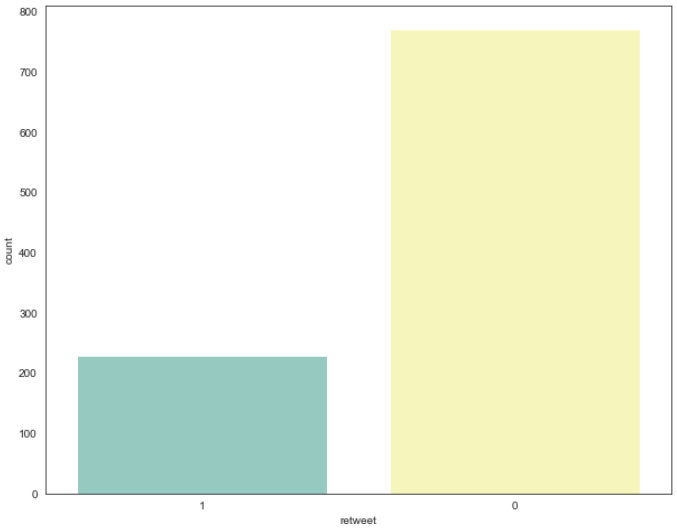
Penelitian yang dilakukan menggunakan k-fold cross validation dengan nilai k = 5, menghasilkan akurasi keseluruhan fold untuk masing masing kernelnya yang ditampilkan pada histogram berikut ini:



Gambar 2. F1-score untuk setiap Kernel SVM

Seperti yang diperlihatkan pada Gambar 2 pengujian menghasilkan rata rata performansi akurasi dari tiga jenis kernel svm yang berbeda yakni RBF, Linear, dan juga Polynom. Pada kernel RBF dihasilkan nilai rata rata f1-score sebesar 55,60 % sedangkan untuk kernel Linear menghasilkan nilai rata rata f1-score sebesar 52,93% dan yang terakhir untuk kernel Polynom sebesar 54,70%. Ketiga hasil yang ditampilkan memberikan kesimpulan bahwa SVM kernel RBF merupakan kernel dengan f1-score paling tinggi pada penelitian ini.

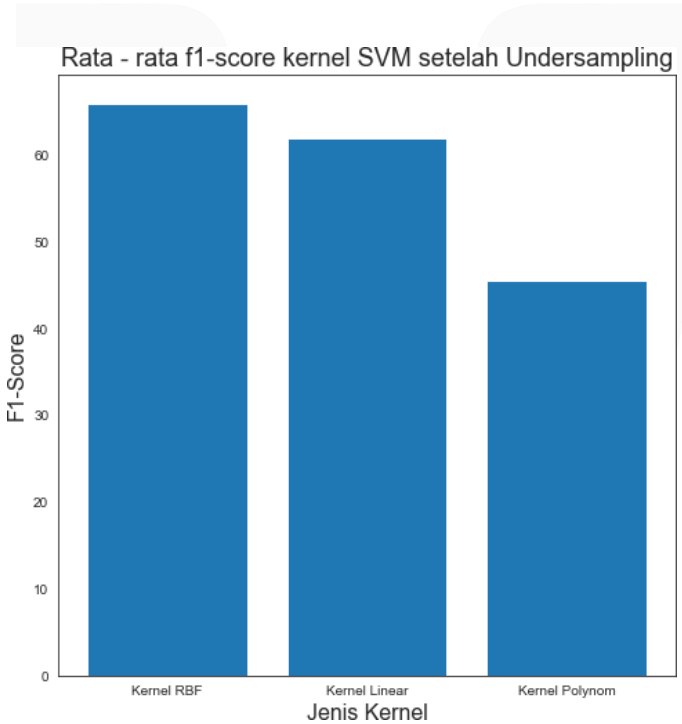
Dikarenakan kelas data yang *imbalanced*, dilakukan metode *undersampling* untuk mengatasi permasalahan ini. Kelas data yang awalnya *imbalanced* dengan jumlah kelas retweet sebanyak 228 dan kelas tidak diretweet sebanyak 772, setelah dilakukan metode *undersampling* berubah menjadi 228 baik untuk kelas retweet maupun tidak diretweet.



Gambar 3. Perbandingan value pada masing masing kelas

Gambar 4. Perbandingan value setelah proses undersampling

Selanjutnya tidak lupa dilakukan kembali pengukuran performansi f1-score untuk masing masing kernel untuk mengetahui adanya penurunan atau peningkatan pada kelas data. Didapatkan f1-score sebesar 66,05% untuk kernel RBF, lalu 61,95% untuk kernel linear dan yang terakhir f1-score sebesar 45,55% untuk kernel polynom yang mana memiliki f1-score terkecil dibandingkan 2 kernel lainnya seperti yang terlihat pada Gambar 5 dibawah.

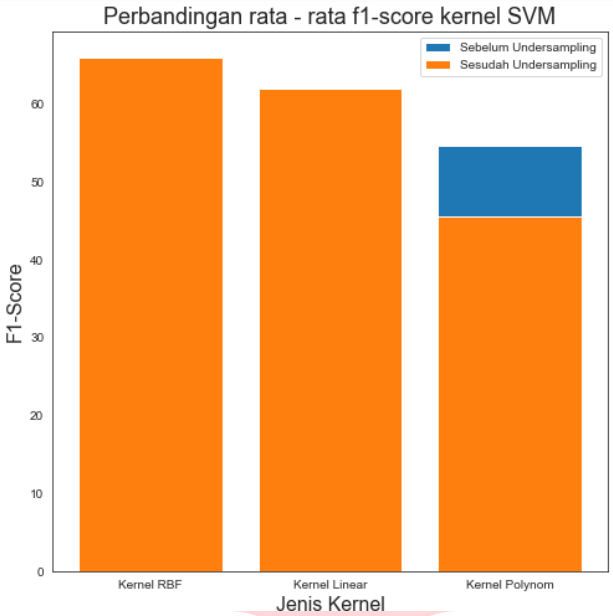


Gambar 5. F1-score untuk setiap Kernel SVM setelah undersampling

4.3 Analisis Hasil Pengujian

Dari hasil penelitian untuk prediksi retweet menggunakan metode SVM menghasilkan performansi yang memuaskan, dimana f1-score tertinggi diraih oleh kernel RBF 66,05% yang mana mengalami peningkatan sebesar 10,45% dari hasil sebelumnya yakni 55,60%. Untuk kernel linear sendiri terdapat peningkatan yang cukup signifikan yakni sebesar 9,02%, yang awalnya 52,93% menjadi 61,95%. Sementara untuk kernel Polymon mengalami penurunan performa sebesar 9,15 persen, yang awalnya 54,70% menjadi 45,55%. Yang mana dapat disimpulkan bahwa kernel RBF memiliki f1-score tertinggi, Linear mengalami peningkatan yang cukup signifikan

sementara polynom berbeda dari yang lainnya dengan mengalami penurunan setelah *undersampling*. Peubahan masing masing kernel akan ditampilkan pada *overlapping* histogram dibawah ini.



Gambar 6. Perbandingan f1-score

5. Evaluasi

Kesimpulan penelian yang bertujuan untuk menghasilkan sistem prediksi retweet. Menghasilkan performansi dan akurasi menggunakan 7 atribut yang dianggap termasuk dalam feature user-based antara lain User.favourites_count, User.statuses_count, User.verified, User.listed_count, User.followers_count, User.friends_count, dan Retweet_count yang dijadikan sebagai class. Penelitian berhasil menghasilkan sistem prediksi yang memiliki performansi yang cukup bagus, dengan hasil f1-score tertinggi yakni 66,05% yang dihasilkan oleh kernel RBF setelah *undersampling*. Penelitian juga berhasil menangani permasalahan *imbalanced* yang terdapat pada data menggunakan metode *undersampling*.

Saran yang dapat diberikan untuk penelitian selanjutnya adalah untuk mencoba menghasilkan sistem yang dapat memprediksi jumlah retweet pada suatu postingan twitter, baik menggunakan *feature user-based* maupun keseluruhan attribute dalam twitter. Disarankan juga untuk menggunakan dataset yang seimbang untuk memudahkan pembuatan sistem dan meningkatkan performansi.

REFERENSI

[1] Boyd, d., Golder, S., and Lotan, G. **"**Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter.**"** Available : https://www.danah.org/papers/TweetTweetRetweet.pdf

[2] Bongwon**, S.**, Lichan**,** H., Peter**, .**, and Ed**,** H. "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network". Available : https://ieeexplore.ieee.org/document/5590452

[3] Diaz, S. (2007, June 9). "Life, in little chirps: Introducing Twitter, a web experience in the mass appeal of mundane details."

[4] Kwak, H., Lee, C., Park, H., and Moon, S. "What is Twitter, a Social Network or a News Media?."

[5] Stefanie, H., Timothy, D., Kim , H., Andrew, T., Cassidy and Vincent, L."Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter". Available : https://arxiv.org/ftp/arxiv/papers/1410/1410.4139.pdf

[6] Fiona, M., Derek, J., Gail , C., Heather, M."Understanding Twitter" Available : https://www.researchgate.net/publication/239901618_Understanding_Twitter

[7] Jose, L. "The presence of academic journals on Twitter and its relationship with dissemination (tweets) and research impact"

[8] Anto, S., Arief, B., Dwi , H. **"**Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika". Available : https://asnugroho.net/papers/ikcsvm.pdf

[9] Suhardjono, Ganda, W., Abdul , H. "Prediksi Waktu Kelulusan Mahasiswa Menggunakan SVM Berbasis PSO"

[10] Muhammad, A., Isa, I., Elly, M., “Support Vector Machine untuk Image Retrieval”
[11] Chris, A. “Election Day Tweets”. Available : https://www.kaggle.com/kinguistics/election-day-tweets?select=election_day_tweets.csv