Akreditasi Ristekdikti, No:
30/E/KPT/2019 (Sinta 4)
DOI: https://doi.org/10.31294/p.v21i2

Paradigma – Jurnal Informatika dan Komputer,
Vol. 22, No. 2 September 2020
P-ISSN 1410-5063, E-ISSN: 2579-3500

Sentimen Analisis Stay Home menggunakan metode klasifikasi Naive Bayes, Support Vector Machine, dan k-Nearest Neighbor

Ikhwanul Hakim[1], Arifin Nugroho[2], Sulaeman Hadi Sukmana[3], Windu Gata[4]

[1]STMIK Nusa Mandiri
e-mail: 14002322@nusamandiri.ac.id

[2]STMIK Nusa Mandiri
e-mail: 14002306@nusamandiri.ac.id

[3]STMIK Nusa Mandiri
e-mail: sulaeman.sdu@nusamandiri.ac.id

[4]STMIK Nusa Mandiri
e-mail: windu@nusamandiri.ac.id

Abstrak *- Dunia sedang dilanda pandemi Corona Virus, virus yang berasal dari kota Wuhan di negara Cina sebagai awal pusat dari pandemi virus tersebut. Virus tersebut menyerang pernafasan akut dan menyebar dengan cepat hampir keseluruh dunia karena proses penularannya yang relatif mudah. Pemberitaan terkait virus tersebut terjadi dengan saat masif baik dimedia nasional maupun internasional. Hampir seluruh media memberitakan tentang penyebaran virus tersebut. Salah satunya melalui media sosial, twitter adalah salah satu media sosial yang cukup banyak penggunanya dan cukup digemari. Banyak pengguna twitter membagikan informasi, mengeluarkan pendapat, maupun berbagi beberapa hal. Penelitian ini fokus pada sentimen analisis stay home pada pengguna twitter, untuk dapat melihat efek dari kebijakan tersebut terhadap kehidupan mereka. Karena hampir diseluruh negara yang terkena pandemi ini mengeluarkan kebijakan seperti itu. Data yang diperoleh akan diolah menggunakan tiga metode klasifikasi yaitu Naive Bayes (NB), Support Vector Machine (SVM), dan k-Nearest Neighbor (k-NN). Dengan ketiga metode klasifikasi tersebut, akan dicari metode mana yang akan menghasilkan akurasi yang paling baik terkait dengan stay home dari tweets para penggunanya. Setelah dilakukan percobaan, algoritma Support Vector Machine + Smote mendapatkan hasil akurasi yang paling baik jika dibandingkan dengan dua algoritma lainnya. Hasil akurasi yang didapat sebesar 80,05%.*

Kata Kunci: corona virus, twitter, stay home, SVM, sentimen analisis

Abstract *- The world is being hit by the Corona Virus pandemic, a virus that originated from the city of Wuhan in China as the beginning of the center of the virus pandemic. The virus attacks acute respiratory tract and spreads rapidly almost throughout the world because of the relatively easy transmission process. News related to the virus occurred at a massive time both in the national and international media. Almost all media reported about the spread of the virus. One of them is through social media, Twitter is a social media that has quite a lot of users and is quite popular. Many Twitter users share information, express opinions, and share several things. This study focuses on the sentiment analysis of stay home for Twitter users, to be able to see the effects of these policies on their lives. Because almost all countries affected by this pandemic have issued such policies. The data obtained will be processed using three classification methods, namely Naive Bayes (NB), Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN). With these three classification methods, we will look for which method will produce the best accuracy in terms of staying home from the tweets of its users. After an experiment, the Support Vector Machine + Smote algorithm gets the best accuracy results when compared to the other two algorithms. The accuracy results obtained are 80.05%.*

Keywords: corona virus, twitter, stay home, SVM, sentiment analysis

PENDAHULUAN

Corona Virus (COV) adalah salah satu patogen utama yang terutama menargetkan virus sistem pernapasan manusia. Sejak Desember 2019, total 41 kasus pneumonia yang tidak diketahui etiologi telah dikonfirmasi di kota Wuhan, Provinsi Hubei, Cina. Sebagian besar pasien mengunjungi pasar ikan dan hewan liar setempat bulan lalu. Pada konferensi pers nasional yang diadakan hari ini, Dr JianguoXu, seorang akademisi dari *Chinese Academy of Engineering*, yang memimpin tim ascientific mengumumkan bahwa coronavirus jenis baru, yang secara tentatif dinamai Organisasi Kesehatan Dunia

Naskah diterima:23 Juni 2020 , direvisi: 10 September 2020, disetujui: 14 September 2020

sebagai coronavirus baru 2019 (Lu et al., 2020).

Wabah (COV) sebelumnya termasuk sindrom pernafasan akut yang parah (SARS) –Co V dan Sindrom pernapasan Timur Tengah (MERS) -CoV yang sebelumnya telah ditandai sebagai agen yang merupakan ancaman kesehatan masyarakat yang besar (Rothan & Byrareddy, 2020).

Pemberitaan Covid-19 sangatlah ramai di salah satu media social, seperti : *Twitter* merupakan salah satu media sosial yang ada dan digemari oleh para penggunanya, pengguna *twitter* dapat dengan bebas berpendapat, maupun sharing berbagai macam hal.

Situs *microblogging* kaya akan sumber untuk berbagai jenis informasi. Ini adalah tempat yang biasa di mana orang bertukar pendapat tentang berbagai masalah yang mungkin terjadi pada tren yang sedang berlangsung. Berdasarkan pengalaman mereka, mereka berbagi komentar atau keluhan pada produk apa pun dan mengekspresikan pemikiran mereka dalam hal sentimen positif atau negatif. Banyak organisasi yang akan datang memerlukan analisis umpan balik pada produk mereka untuk meningkatkan lebih lanjut. Sebagian besar waktu, Penyelenggara menganalisis tanggapan pengguna dan menjawabnya kembali di media sosial. Jadi, inilah tantangan untuk menganalisis atau mendeteksi dan menyelesaikan sentimen global (Trupthi et al., 2017).

Twitter adalah *microblogging* yang banyak digunakan platform dan layanan jejaring sosial yang menghasilkan luas jumlah informasi (Saad & Yang, 2019). Dalam beberapa tahun terakhir, para peneliti lebih menyukai untuk memanfaatkan data sosial agar dapat digunakan untuk sentimen analisis pendapat orang tentang suatu produk, topik, atau peristiwa. Sentimen analisis, juga dikenal sebagai penambangan opini, merupakan proses bahasa alami yang penting tugas. Proses ini menentukan orientasi sentimen teks sebagai positif, negatif, atau netral (Kim & Hovy, 2004; Whitelaw et al., 2005). Penambangan teks umumnya mencakup kategorisasi informasi atau teks, pengelompokan teks, ekstraksi entitas atau konsep, pengembangan dan perumusan taksonomi umum (Hashimi et al., 2015).

Sentimen analisis, juga disebut penambangan opini, adalah bidang studi yang menganalisis opini, sentimen, evaluasi, penilaian, sikap, dan emosi orang terhadap entitas seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya. Ini mewakili ruang masalah yang besar. Ada juga banyak nama dan tugas yang sedikit berbeda, mis., Analisis sentimen, penambangan opini, ekstraksi opini, penambangan sentimen, analisis subjektivitas, analisis pengaruh, analisis emosi, tinjauan penambangan, dll (Liu, 2012).

Pada sentimen analisis kali ini menggunakan algoritma klasifikasi yang terdiri dari *Naive Bayes* (NB), *Support Vector Machine* (SVM), dan *k-Nearest Neighbor* (k-NN) karena banyak dari penelitian-penelitian sebelumnya yang menggunakan algoritma tersebut. Ketiga algoritma tersebut digunakan untuk nantinya sebagai suatu metode klasifikasi yang dapat digunakan untuk memprediksi sentimen analis pada yang ada di dalam *twitter* terkait dengan *stay home*.

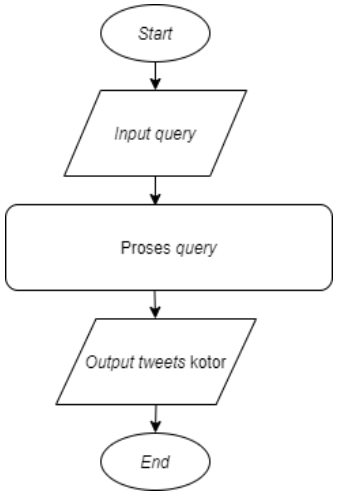
Pada penelitian sebelumnya yang pernah dilakukan berkaitan dengan masalah yang dihadapi penulis dimana dengan judul "Eksperimen Sistem Klasifikasi Analisa Sentimen *Twitter* pada Akun Resmi Pemerintahan Kota Surabaya Berbasis Pembelajaran Mesin". penelitian dilakukan oleh Faradhillah, Nuke Y. A..,dkk pada tahun 2016. Pada penelitian tersebut menggunakan 2 metode algoritma yang digunakan pada tahap klasifikasi yaitu *Naive Bayes* dan *Support Vector Machine* (SVM) didapatkan hasil bahwa nilai akurasi yang lebih tinggi dihasilkan dengan metode *Support Vector Machine* (SVM). Untuk perbandingan penggunaan kernel yaitu linear dan RBF, kernel RBF mempunyai hasil yang lebih bagus. Untuk nilai C dan γ terbaik yaitu C=20 dan γ=2-1,25 (Faradhillah et al., 2016).

Penelitian lainnya adalah "Analisis Sentimen *Twitter* Menggunakan *Text Mining* dengan Algoritma *Naive Bayes Classifier*". Penelitian dilakukan oleh Sudiantoro, Adhi V..,dkk pada tahun 2018. Penelitian ini menghasilkan Algoritma *Naive Bayes Classifier* sangat efektif untuk digunakan sebagai proses klasifikasi *tweets* yang dibutuhkan dalam sistem sentimen analisis ini dimana nilai yang di dapatkan dalam pengujian sampai 84%. Metode *Naive Bayes Classifier* dapat digunakan untuk melakukan klasifikasi *tweets* dengan cukup baik pada sistem sentimen analisis. 300 data yang dibagi menjadi 2 yaitu data latih sebanyak 200 data dan data uji berjumlah 100 data. Hasil dari klasifikasi diketahui bahwa 100 data yang diuji masuk dalam kategori sentimen negatif (Sudiantoro et al., 2018).

METODOLOGI PENELITIAN

1. Tahap Pengumpulan Data

Pada proses awal pengumpulan data, menggunakan *twitterscrapper* yang tersedia pada *anaconda promt*. Dengan menggunakan aplikasi tersebut proses penarikan data *tweets* pada *twitter* dapat dilakukan. Adapun tahapan dalam melakukan proses tersebut dapat dilihat pada Gambar 1.



Gambar 1. Tahap Pengumpulan Data

Dari Gambar 1 dapat dilihat jika hasil akhir dari tahap pengumpulan data untuk mendapatkan *tweets* yang berkaitan dengan *stay home* yang untuk selanjutnya dapat digunakan di tahap *preprossesing*.

2. Tahap *Preposessing*

Proses *preposessing* ini dilakukan melalui beberapa proses agar dapat menghasilkan data *tweets* yang baik, yang nantinya akan digunakan pada proses pengolahan data. Proses tersebut dilakukan karena pada proses pengumpulan data *tweets*, masih didapati *tweets* yang kurang sempurna. Seperti masih adanya simbol, *url*, dan beberapa data lain yang harus dilakukan pernyempurnaan agar mendapatkan hasil yang maksimal. Adapun tahapan dalam melakukan *preposessing* ini adalah :

a. *Regex Removal*

Pada tahap ini akan dilakukan proses pembersihan kata-kata didalam *tweets* yang masih terdapat simbol-simbol, seperti (!””,.`~ dst).
Contoh : *We're "extreme distancing" because no one knows who may have it*
Menjadi : *Were extreme distancing because no one knows who may have it*

b. *Remove URL*

Pada tahap ini akan dilakukan proses pembersihan *url* yang masih ada didalam *tweets*, seperti (http, .com, dst).
Contoh : *ones flying. https://www.wsj.com*
Menjadi : *ones flying*

c. *Remove Annotation*

Pada tahap ini akan dilakukan proses pembersihan *tweets* yang masih mengandung *username* ‘@’ didalamnya.

Contoh : *@TIMOTHY45231830 Stay home stay safe stay healthy*
Menjadi : *Stay home stay safe stay healthy*

d. *Remove Duplicate Tweets*

Pada tahap ini *tweets* yang memiliki isi yang sama akan dihapus, hal tersebut dilakukan untuk menghindari duplikasi isi dari *tweets* yang sudah ada sebelumnya agar dapat meningkatkan hasil akurasi.

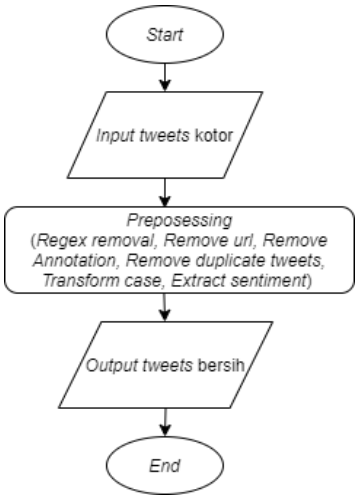
e. *Extract Sentiment*

Pada tahap ini menggunakan model *vader*, proses tersebut dilakukan untuk mengolah data agar tiap *tweets* mendapatkan label sesuai dengan *value* masing-masing.

f. *Transform Case*

Pada tahap ini semua huruf yang ada didalam *tweets* menjadi huruf kecil. Hanya huruf “a” sampai dengan “z” yang akan dilakukan perubahan.
Contoh : *Were extreme distancing because no one knows who may have it*
Menjadi : *were extreme distancing because no one knows who may have it*

Berikut adalah *flowchart* dari tahapan *preprocessing* berlangsung yang dapat dilihat di Gambar 2.



Gambar 2. Tahapan *Preprocessing*

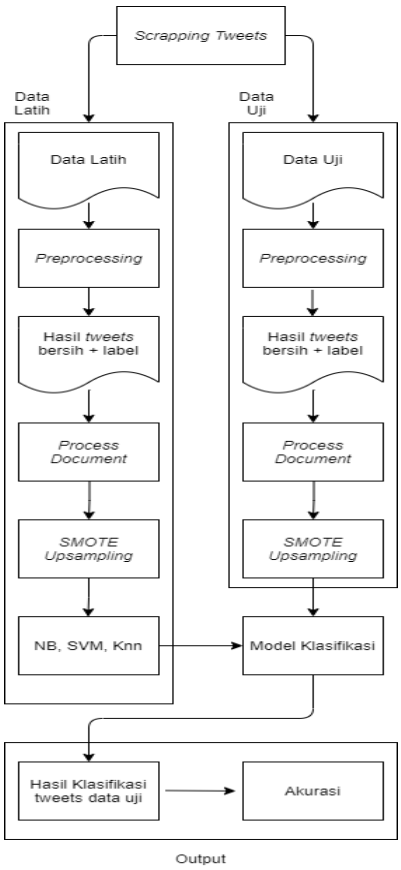
Dari Gambar 2 dapat dilihat jika hasil akhir dari tahap *preprocessing* untuk mendapatkan *tweets* bersih yang sudah dapat digunakan.

3. Tahap Pengujian Data

Tahapan ini dibuat agar dapat menggambarkan tahapan-tahapan dari proses yang dilakukan aplikasi sentimen analisis mulai dari pengambilan data, *preprosessing*, perhitungan klasifikasi sampai dengan akurasi.

Data yang digunakan adalah data *tweets* bersih hasil dari tahapan preprocessing sebelumnya yang akan dilakukan uji coba, dengan menggunakan tiga perhitungan algoritma yang terdiri dari *Naive Bayes (NB), Support Vector Machine (SVM),* dan *k-Nearest Neighbor (k-NN).*

Dengan menambahkan *operator toolbox SMOTE Upsampling*, penggunaan proses tersebut dilakukan agar data *tweets* menjadi lebih seimbang antara *tweets* yang berlabel *positive* maupun *negative*. Sehingga dengan melalui tahapan tersebut dapat meningkatkan hasil akurasi pada percobaan data *tweets*. *Flowchart* bagaimana tahap pengujian data dapat dilihat di Gambar 3.



Gambar 3. Proses Tahap Pengujian Data

Dari Gambar 3 dapat dilihat output dari sistem yang dibuat berupa 2 kategori (*positive*, dan *negative*). Hasil yang dapat diperoleh dari sistem terhadap data *tweets* yang diuji, akan mendapatkan tingkat akurasi dari tiap algoritma berbeda yang akan bentuk tersaji dalam bentuk presentase.

HASIL DAN PEMBAHASAN

1. Scrapping Tweets

Pengambilan *tweets* menggunakan *twitterscrapper* pada *anaconda promt*, data *tweets* diambil dari tanggal 1 Maret 2020 – 1 April 2020. Jumlah data yang terkumpul dalam kurun waktu tersebut sebanyak 1652 *tweets* berhasil dikumpulkan dan berikut adalah hasil dari 10 *tweets* yang terambil.

Tabel 1. Hasil *Scrapping Twitter*

No	*Tweets*
1	*We're "extreme distancing" because no one knows*
2	*Interesting the advertisement under your post*
3	*Following directives? The @CDCgov said to stay*
4	*Please suspend work and bills for* 14 *days let our*
5	*Yes , Dr said stay home* 14 *days*
6	*Stop listening to the evil people...they will always try..*
7	*These knuckle heads can't even keep out streets,publ*
8	*It is great time to prepare . Stay home & read*
9	*I've been thinking that this is what we should*
10	*You need to pay us to stay home, period*

Dari Tabel 1 kita dapat melihat bahwa data *tweets* kotor yang terkumpul, dimana masih banyak data yang kurang sempurna. Tentu saja data tersebut belum dapat digunakan untuk diuji.

2. Preprocessing

Setelah memperoleh data *tweets* kotor dari hasil *scrapping twitter*, maka tahapan selanjutnya adalah *preprocessing*. Melalui tahapan ini *tweets* kotor diolah melalui berbagai macam proses seperti : *Regex Removal, Remove URL, Remove Annotation, Remove Duplicate Tweets, Extract Sentiment, Transform Case*. Tahapan selanjutnya adalah proses *preprocessing*, target dari proses ini sendiri akan menghasilkan *tweets* bersih. Berikut adalah hasil dari 10 *tweets* yang sudah dilakukan tahap *preprocessing.*

Tabel 2. Hasil *Preprocessing*

No	*Tweets*
1	*were extreme distancing because no one knows*
2	*interesting the advertisement under your post*
3	*following directives the said to stay*
4	*please suspend work and bills for* 14 *days let our*
5	*yes dr said stay home* 14 *days*
6	*stop listening to the evil peoplethey will always try*
7	*these knuckle heads cant even keep out streetpubl*
8	*if it makes me a stay home mom im down*
9	*it is great time to prepare stay home read*
10	*stay home cuzao*

Dari Tabel 2 setelah melalui tahap *preprocessing* dapat dilihat perubahan data *tweets*, data sudah menjadi lebih baik, sudah tidak lagi terdapat simbol, tanda baca, dll.

Berikut adalah contoh hasil dari *tweets* yang sudah dilakukan pelabelan menggunakan *vader*, yang dapat dilihat di Tabel 3.

Tabel 3. Hasil *Vader*

No	*Tweets*	Nilai	Status
1	*were extreme distancing because no one knows*	-0,025	*Negative*
2	*interesting the advertisement under your post*	2,384	*Positive*
3	*following directives the said to stay*	-0,948	*Negative*
4	*please suspend work and bills for* 14 *days let our*	-0,307	*Negative*
5	*yes dr said stay home* 14 *days*	0,435	*Positive*
6	*stop listening to the evil peoplethey will always try*	-0,692	*Negative*
7	*these knuckle heads cant even keep out streetpubl*	0,179	*Positive*
8	*I pharma screwed us*	-0,564	*Negative*
9	*it is great time to prepare stay home read*	0,794	*Positive*
10	*you need to pay us to stay home period*	0,461	*Positive*

Seperti yang diketahui bersama berdasarkan dari Tabel 3 diatas, diperoleh nilai dari tiap *tweets* yang seluruh prosesnya menggunakan bantuan dari *vader*. Jika sudah mendapatkan data tersebut maka data sudah dapat digunakan untuk proses selanjutnya yaitu proses pengujian data.

3. Pengujian Data

Pada tahapan ini, data tweets *hasil dari preprocessing* akan diolah dengan menggunakan tiga algoritma yang terdiri dari *Naive Bayes (NB), Support Vector Machine (SVM),* dan *k-Nearest Neighbor (k-NN).*

a. *Naive Bayes*

Berikut ini adalah hasil pengolahan data yang sudah diproses menggunakan algortima *naive bayes*, yang dapat dilihat pada Tabel 4.

Tabel 4. Hasil dari *NB + SMOTE*

	True Negative	True Positive	Class precision
Pred. Negative	466	268	63,49%
Pred. Positive	123	321	72,30%
Class Recall	79,12%	54.50%	

Sedangkan untuk hasil dari akurasi menggunakan *naive bayes + SMOTE* sebesar 66,81%

b. Support Vector Machine

Berikut ini adalah hasil pengolahan data yang sudah diproses menggunakan algortima *support vector machine*, yang dapat dilihat pada Tabel 5.

Tabel 5. Hasil dari *SVM + SMOTE*

	True Negative	True Positive	Class precision
Pred. Negative	452	82	82,18%
Pred. Positive	137	491	78,18%
Class Recall	76,74%	83,36%	

Sedangkan untuk hasil dari akurasi menggunakan *support vector machine + SMOTE* sebesar 80,05%

c. k-Nearest Neighbor

Berikut ini adalah hasil pengolahan data yang sudah diproses menggunakan algortima *k-Nearest Neighbor*, yang dapat dilihat pada Tabel 6.

Tabel 6. Hasil dari *k-NN + SMOTE*

	True Negative	True Positive	Class precision
Pred. Negative	589	572	50,73%
Pred. Positive	0	17	100%
Class Recall	100%	2,89%	

Sedangkan untuk hasil dari akurasi menggunakan *k-Nearest Neighbor + SMOTE* sebesar 51,45%

Berikut adalah hasil rekapitulasi dari ketiga algoritma, setelah proses pengujian data dilakukan dan didapati hasil akurasi yang berbeda-beda. Data tersebut dapat dilihat pada Tabel 7.

Tabel 7. Hasil Rekapitulasi Akurasi

	Akurasi
NB + SMOTE	66,81%
SVM + SMOTE	80,05%
k-NN + SMOTE	51,45%

KESIMPULAN

Setelah melalui proses panjang dari pengumpulan data sampai dengan pengujian data. Ketiga algoritma yang digunakan mendapatkan hasil akurasi yang berbeda-beda. Hasil terbaik didapatkan dengan menggunakan *support vector machine* mengungguli kedua algoritma yang lain dengan mendapatkan akurasi sebesar 80,05%. Sangat besar kemungkinan hasil dari akurasi dapat meningkat apabila dilakukan beberapa optimalisasi ataupun dengan metode yang berbeda terkait dengan pengujian data diatas.

REFERENSI

Faradhillah, N. Y. A., Kusumawardani, R. P., Hafidz, I., Informasi, J. S., & Informasi, F. T. (2016). *Eksperimen Sistem Klasifikasi Analisa Sentimen Twitter pada Akun Resmi Pemerintahan Kota Surabaya Berbasis Pembelajaran Mesin*.

Hashimi, H., Hafez, A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior*, *51*, 729–733. https://doi.org/10.1016/j.chb.2014.10.062

Kim, S.-M., & Hovy, E. (2004). *Determining the sentiment of opinions*. 1367-es. https://doi.org/10.3115/1220355.1220555

Liu, B. (2012). Opinion spam detection. In *Sentiment Analysis and Opinion Mining* (Issue April). https://doi.org/10.1142/9789813100459_0007

Lu, H., Stratton, C. W., & Tang, Y. W. (2020). Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology*, *92*(4), 401–402. https://doi.org/10.1002/jmv.25678

Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. *Journal of Autoimmunity*, *February*, 102433. https://doi.org/10.1016/j.jaut.2020.102433

Saad, S. E., & Yang, J. (2019). Twitter Sentiment Analysis Based on Ordinal Regression. *IEEE Access*, *7*, 163677–163685. https://doi.org/10.1109/ACCESS.2019.2952127

Sudiantoro, A. V., Zuliarso, E., Studi, P., Informatika, T., Informasi, F. T., Stikubank, U., & Mining, T. (2018). *Analisis Sentimen Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier*. 398–401.

Trupthi, M., Pabboju, S., & Narasimha, G. (2017). Sentiment analysis on twitter using streaming API. *Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017*, 915–919. https://doi.org/10.1109/IACC.2017.0186

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. *International Conference on Information and Knowledge Management, Proceedings*, 625–631. https://doi.org/10.1145/1099554.1099714

PROFIL PENULIS

Ikhwanul Hakim, B.Ict
Mahasiswa Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, STMIK Nusa Mandiri. Jl. Kramat Raya No.18, RT.5/RW.7, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450.

Arifin Nugroho, S.Kom
Mahasiswa Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, STMIK Nusa Mandiri. Jl. Kramat Raya No.18, RT.5/RW.7, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450.

Sulaeman Hadi Sukmana, M.Kom
Dosen STMIK Nusa Mandiri. Jl. Kramat Raya No.18, RT.5/RW.7, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450.

Dr. Windu Gata, M.Kom
Dosen STMIK Nusa Mandiri. Jl. Kramat Raya No.18, RT.5/RW.7, Kwitang, Kec. Senen, Kota Jakarta Pusat, Daerah Khusus Ibukota Jakarta 10450.