JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

SELEKSI FITUR *INFORMATION GAIN* (IG) PADA KLASIFIKASI DATA OPINI SAHAM MENGGUNAKAN METODE NAÏVE BAYES

DEWI FATMARANI SURIANTO[1], KHAERUNNISA NUR FATIMAH SYAHNUR[2]
[1]Program Studi Teknik Komputer, Fakultas Teknik, Universitas Negeri Makassar
[2]Program Studi Manajemen Retail, Institut Teknologi dan Bisnis Kalla
[1]Jl. Dg. Tata Raya, Raya Parang Tambung, Mannuruki, Kec. Tamalate, Kota Makassar, Sulawesi Selatan 90224
[2]Lt 5 & 6 Office Building Nipah Mall, Jl. Urip Sumoharjo, Kota Makassar, Sulawesi Selatan 90231
Email: [1]dewifatmaranis@unm.ac.id, [2]khaerunnisanfsyahnur@kallabs.ac.id

ABSTRAK

Perkembangan media sosial dan *website* sebagai media penyebaran informasi sangat cepat dan mudah diakses. Namun hal tersebut tidak sepenuhnya akurat dan dapat ditanggapi oleh berbagai pihak seperti investor dan calon investor sehingga dapat memicu sentimen investor yang berdampak pada harga saham perusahaan. Sentimen sebagai proses mengidentifikasi opini dalam bentuk teks yang dapat dikelompokkan menjadi sentimen positif atau negatif. Penelitian ini bertujuan untuk menguji penggunaan metode Information Gain dalam proses klasifikasi opini publik atas saham ke dalam kelas sentimen positif atau kelas sentimen negative dengan metode klasifikasi Nave Bayes Classifier. Dari hasil penelitian diperoleh bahwa akurasi klasifikasi yang dihasilkan dengan menggunakan metode seleksi fitur Information Gain meningkat dari 42,86% menjadi 50%.

Kata Kunci: Klasifikasi, *Information Gain, Naïve Bayes*, Opini Saham

I. PENDAHULUAN

Jumlah data saat ini semakin meningkat tiap harinya. Hampir seluruh institusi, organisasi ataupun industri memanfaatkan teknologi untuk menyimpan datanya secara elektronik dan digital. Dengan adanya teknologi, data dapat disimpan di berbagai sumber seperti sosial media, website, portal dan lain-lain (Prakoso et al., 2019). Pesatnya arus informasi publik melalui inovasi teknologi saat ini dapat mempengaruhi

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

pola pikir dan perilaku individu dalam bertindak. Hal ini juga terjadi dalam pasar modal yang melibatkan banyak pihak, salah satunya adalah investor (Xie & Wang, 2017) (Yang, Lin, & Yi, 2017) (Al-Thaqeb, 2018).

Data yang tersebar memiliki beragam bentuk, dimulai dari teks, video, maupun gambar dengan format informasi yang tidak terstruktur dan semi-struktur (Kumar & Bhatia, 2013). Banyaknya data yang tersebar tentu memiliki tantangan tersendiri yaitu menemukan pola atau informasi dari setiap data yang ada, salah satunya adalah dalam bidang *text mining*. Klasifikasi Teks adalah cakupan area dari *text mining*. Klasifikasi Teks bertujuan untuk mengklasifikasikan atau mengkategorisasikan teks yang memiliki karakteristik yang serupa yang selanjutnya dikelompokkan ke dalam kelas yang telah ditentukan sebelumnya. Hasil klasifikasi didasarkan pada fitur-fitur yang terdapat dalam data. Dalam beberapa algoritma, atribut data memegang peran penting. Akan tetapi, tidak semua atribut dapat disimpulkan relevan, karena beberapa diantaranya mungkin tidak terlalu relevan dengan klasifikasi teks yang bahkan dapat membuat dimensi kata menjadi lebih tinggi (Utami & Wahono, 2015).

Oleh karena itu, segala atribut atau fitur perlu diekstrak terlebih dahulu. Hal ini dapat diatasi dengan melakukan seleksi atribut/fitur. Seleksi fitur merupakan suatu pendekatan untuk mereduksi dimensi atau matriks data dengan mempertimbangkan informasi-informasi penting yang perlu diproses sehingga dapat meningkatkan kinerja klasifikasi (Maulida, Suyatno, & Rahmania Hatta, 2016). Information Gain merupakan salah satu metode dalam penyeleksian fitur yang kurang relevan dalam suatu dokumen. Metode tersebut digunakan untuk mengurutkan kata-kata penting dari proses penyeleksian fitur. Hasil dari metode Information Gain adalah kata, fitur atau term yang dianggap penting pada data yang digunakan (Maulida et al., 2016).

Telah banyak penelitian yang dilakukan khususnya pada klasifikasi teks. Salah satunya adalah studi oleh Fitrhi (Jumeilah, 2018). Hasil penelitian diperoleh rata-rata akurasi 85%, rata-rata nilai presisi adalah 79%, dan rata-rata recall yang diperoleh adalah 67% (Jumeilah, 2018). Tahun 2018, Novan dkk (Dimas Pratama, Sari, & Adikara, 2018)., melakukan penelitian klasifikasi sentimen dengan menggunakan

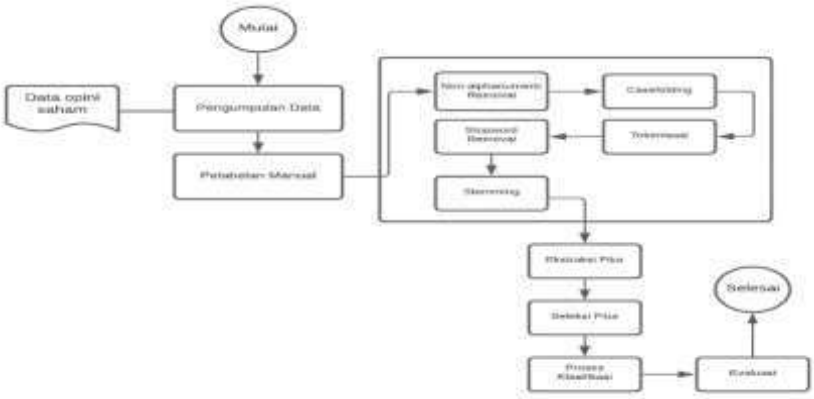
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

metode Naïve Bayes dengan seleksi fitur Chi Square. Hasil studi menunjukkan bahwa penggunaan *feature selection* tidak memiliki pengaruh yang signifikan terhadap akurasi klasifikasi yang diperoleh (Dimas Pratama, Sari, & Adikara, 2018).

Namun, hasil yang berbeda diperoleh oleh Arif, dkk (Negara, Muhardi, & Putri, 2020). Pada tahun 2020, Arif dkk., melakukan penelitian klasifikasi opini masyarakat terhadap layanan maskapai penerbangan (Negara et al., 2020). Berdasarkan hasil pengujian yang didapatkan, diperoleh bahwa implementasi metode Information Gain berhasil meningkatkan akurasi klasifikasi hingga 0,054% dibandingkan tanpa penggunaan fitur seleksi (Negara et al., 2020). Metode Information Gain juga diaplikasikan dalam proses klasifikasi teks yang dilakukan oleh Lila Dini Utami dan Romi Satria (Utami & Wahono, 2015).

Berdasarkan fenomena dan hasil penelitian terdahulu, dapat dikatakan penggunaan metode penyeleksian fitur berhasil meningkatkan akurasi pada proses klasifikasi dengan mengurangi banyak *term* yang tidak mewakili dokumen. Pada penelitian ini, penulis menggunakan metode Naïve Bayes serta metode *Information Gain* sebagai kombinasi metode untuk meningkatkan performansi klasifikasi opini saham.

II. METODE PENELITIAN



Gambar 1. Rangkaian Penelitian

Tahap penelitian meliputi proses mengumpulkan data, membersihkan data, mengimplementasikan metode seleksi fitur, mengklasifikasikan data hingga uji

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

performansi pada model yang telah dibuat. Berikut adalah rangkaian proses penelitian yang dilakukan:

A. Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data dari media sosial twitter dan stockbit terkait saham LQ45 yang terpengaruh dari peristiwa pemilihan umum (PEMILU) presiden Indonesia tahun 2019. Pemilihan topik peristiwa PEMILU berdasar pada hasil Google Trends yang menunjukkan bahwa peristiwa PEMILU merupakan peristiwa yang mencapai angka 100% selama 1 minggu dan merupakan topik dengan penelusuran tertinggi di Internet. Jumlah data yang digunakan dalam studi ini adalah 350 untuk setiap kelas, baik positif maupun negatif. Jumlah data latih yang diterapkan adalah 80% dari total data, sedangkan data uji yang digunakan sebesar 20%. Berikut adalah distribusi dari dataset yang digunakan:

Tabel 1 Distribusi Dataset

No	Kelas (Polaritas)	#Data Uji	#Data Latih
1	Positif	70	280
2	Negatif	70	280

B. Pelabelan Manual

Selanjutnya, data opini yang telah dikumpulkan dari twitter dan media sosial stockbit akan diberikan label polaritas sentiment masing-masing. Proses *Manual-labeling* dilakukan untuk menentukan polaritas dari opini tersebut, apakah masuk ke dalam kelas positif atau kelas negatif. Proses hand-labeling dilakukan oleh para ahli di bidang saham.

C. Pra-pemrosesan

Setelah seluruh cuitan dan data *stockbit* terkait saham LQ45 telah dikumpulkan dan diberi label, tahap selanjutnya adalah tahap pre-processing. Prepemrosesan adalah salah satu fase penting pada proses penambangan. Proses ini dilakukan dengan mengubah data mentah atau dikenal dengan *rawdata* menjadi informasi yang lebih

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

bersih dan mengurangi *noise* dalam teks untuk dapat digunakan pada tahapan selanjutnya. Teks akan direpresentasikan ke dalam sejumlah fitur.

D. Fitur Ekstraksi

TF-IDF atau yang merupakan singkatan dari *Term Frequency – Inverse Document Frequency*, hal ini dilakukan untuk menganalisa seberapa penting suatu kata dalam sebuah dokumen atau dataset. Metode TF-IDF membantu proses klasifikasi untuk merepresentasikan dan mengekstrak fitur-fitur yang terdapat dalam suatu dokumen.

E. Seleksi Fitur

Feature Selection atau dapat disebut dengan pemilihan fitur merupakan langkah penting klasifikasi teks dengan membentuk ruang vektor guna memperbaiki skalabilitas, efisiensi, serta akurasi pada proses pengelompokkan teks. Ide utama dari Feature Selection adalah memilih fitur-fitur dari sumber dokumen yang ingin diproses (Korde, 2012). Tujuan utama pada studi ini adalah penggunaan algoritma *Information Gain*. *Information Gain* bekerja dengan memilih fitur-fitur yang memiliki bobot tertinggi sesuai dengan jumlah fitur yang diinginkan. Information Gain melibatkan entropi untuk menemukan term terbaik. Jika nilai information gain dari suatu term semakin besar, maka fitur tersebut dianggap semakin signifikan dan semakin penting pada dokumen teks (Negara et al., 2020).

F. Klasifikasi

Setelah melakukan proses pembersihan data melalui tahap pra-pemrosesan dan ekstraksi fitur hingga pemilihan fitur, maka tahap berikutnya adalah proses klasifikasi. Pada tahap ini, algoritma klasifikasi yang dimanfaatkan adalah algoritma Naïve Bayes. Naïve Bayes bekerja melalui proses memprediksi berbasis probabilitas berdasar pada penerapan aturan Bayes yakni perhitungan probabilitas di masing-masing kelas dengan asumsi yaitu tidak adanya ketergantungan antar satu kelas dengan kelas yang lain (independen). Konsep dasar teori *bayes* yaitu peluang bersyarat P(X|C), dimana X merupakan *posterior,* dan C disebut dengan *prior* (Negara et al., 2020).

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

G. Pengujian Performansi

Pengujian performa dilakukan sebagai langkah terakhir dalam eksperimen ini. Proses ini dilakukan menggunakan metode confusion matrix yang bertujuan untuk mengevaluasi kinerja dari sebuah model yang diusulkan. Pada penelitian ini, terdapat beberapa skenario pengujian yang dilakukan, namun seluruh skenario tersebut diuji menggunakan confusion matrix yang mencakup akurasi, presisi, dan recall.

III. HASIL DAN PEMBAHASAN

A. Dataset

Media sosial twitter dan stockbit terkait saham LQ45 yang terpengaruh dari peristiwa pemilihan umum (PEMILU) presiden Indonesia tahun 2019 merupakan sumber dataset yang digunakan pada studi ini. Data kemudian dilabeli secara manual ke dalam dua kelas yakni sebagai opini positif dan opini negatif. Berikut ini merupakan contoh dataset yang digunakan:

Tabel 2 Contoh Dataset

No	Data	Dilabelkan sebagai
1	$BBTN Elliot Wave teori kah? Kalau iya yasudah beli saja. Asal jangan ditutup 2.520	Negatif
2	$PTPP sudah profit banyak sejak sinyal panah hijau muncul, naikkan terus penjualan diatas harga modal setiap hari, ambil banyak profit anda dan ikuti terus kenaikannya	Positif

B. Performansi Klasifikasi Berdasarkan Penggunaan Seleksi Fitur

Pada eksperimen ini, digunakan metode Naïve Bayes sebagai metode klasifikasi untuk mengukur kinerja klasifikasi menggunakan metode pemilihan fitur yaitu Information Gain. Seleksi fitur dilakukan setelah tahap pra-pemrosesan dan proses ekstraksi fitur TF-IDF. Dari proses TF-IDF diperoleh fitur sejumah 939 fitur.

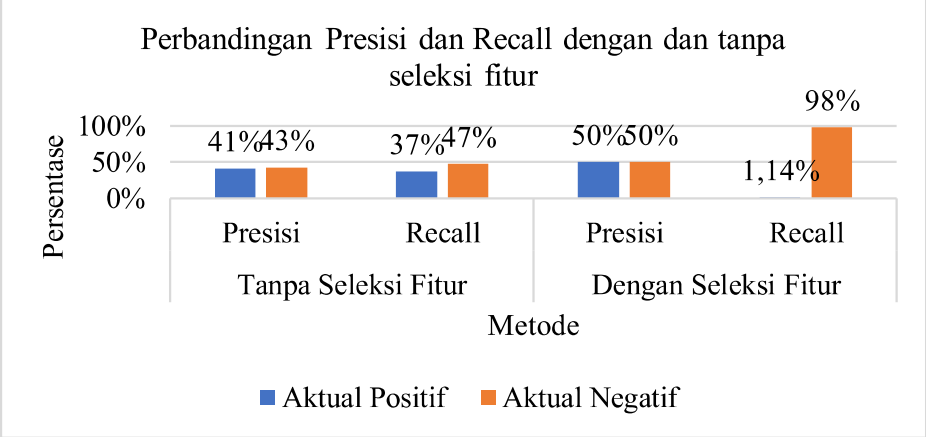
JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Skenario pengujian ini dilakukan untuk membandingkan performa klasifikasi dengan dan tidak melakukan proses seleksi fitur. Metode pemilihan fitur yang digunakan adalah Information Gain dimana jumlah *term* yang diseleksi adalah 10 fitur. Dari proses eksperimen, diperoleh hasil akurasi mencapai 50% dengan peningkatan 7.14%. Berikut adalah tabel dan gambar grafik perbandingan performa model berdasarkan akurasi:

Tabel 3 Performa Model Berdasarkan Akurasi

Metode	Akurasi
Tanpa seleksi fitur	42.86%
Dengan seleksi fitur	**50%**

Pada tabel diatas, dapat kita deskripsikan bahwa jika dibandingkan dengan kinerja klasifikasi tanpa memanfaatkan pemilihan fitur, performa klasifikasi dengan memanfaatkan metode *Information Gain* berhasil meningkatkan akurasi sebesar 7,14% yaitu dari 42,86% menjadi 50%. Gambar dibawah ini merupakan gambar perbandingan Presisi dan Recall dengan dan tanpa seleksi fitur.



Gambar 2 Perbandingan Presisi dan Recall

Dari Gambar 2, dapat dideskripsikan bahwa implementasi seleksi fitur meningkatkan nilai presisi untuk data opini positif maupun data negatif, yakni untuk data opini positif mengalami peningkatan dari 41% ke 50%, dan untuk data opini negatif dari 43% ke 50%. Namun, terdapat perbedaan nilai recall yang dihasilkan.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Untuk data opini positif, implementasi seleksi fitur tidak berhasil meningkatkan nilai recall. Namun, untuk data opini negatif, implementasi seleksi fitur berhasil meningkatkan nilai recall hingga 51% yakni dari 47% menjadi 98%. Dari hasil tersebut, dapat dianalisis bahwa model yang dihasilkan dengan menggunakan seleksi fitur memiliki ketepatan klasifikasi data opini negatif yang baik, yakni hingga 98%, namun model belum mampu mengklasifikasikan data opini positif dengan tepat yang ditunjukkan dengan nilai recall 1,14%. Hal ini terjadi karena fitur-fitur hasil seleksi oleh sistem yang digunakan pada proses klasifikasi tidak cukup efektif. Berikut adalah 5 fitur dengan bobot tertinggi yang digunakan dalam proses klasifikasi menggunakan seleksi fitur yang ditampilkan pada Tabel 4.

Tabel 4. Lima Fitur dengan Bobot Tertinggi

No	Fitur	Bobot
1	Wave	0.067953
2	Buang	0.067099
3	Akurat	0.060204
4	Solusi	0.059136
5	Pantau	0.056667

Observasi lebih lanjut dilakukan dengan membandingkan 5 fitur yang diseleksi. Observasi diawali dengan melakukan tahap pra-pemrosesan pada data-data yang dilabeli positif maupun negatif. Dari hasil pra-pemrosesan data yang dilabeli positif, diperoleh 547 *term,* sedangkan dari hasil pra-pemrosesan pada data yang dilabeli negatif, diperoleh 668 *term*. Selanjutnya, melakukan pemetaan pada contoh 5 fitur yang diseleksi. Berikut adalah hasil observasi yang diperoleh:

Tabel 5. Hasil Observasi Fitur

No	Fitur	Termasuk pada term Dataset yang dilabeli Positif	Termasuk pada term Dataset yang dilabeli Negatif
1	Wave	Tidak	Ya

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

No	Fitur	Termasuk pada term Dataset yang dilabeli Positif	Termasuk pada term Dataset yang dilabeli Negatif
2	Buang	Ya	Ya
3	Akurat	Tidak	Ya
4	Solusi	Ya	Ya
5	Pantau	Ya	Ya

Dari Tabel 5, dapat dijelaskan bahwa dari 5 fitur yang diseleksi, hanya terdapat 3 fitur yang juga terdapat pada *term* hasil pemrosesan dataset yang dilabeli positif, sedangkan jika dibandingkan pada dataset negatif, seluruh fitur tersebut berada pada daftar term hasil pra-pemrosesan dataset yang dilabeli negatif. Hal ini dapat dianalisis bahwa fitur-fitur yang diseleksi tidak cukup efektif untuk mengklasifikasikan data yang dilabeli sebagai opini positif sehingga memiliki nilai *True Positive* yang cukup rendah dan berdampak pada nilai Recall pada proses klasifikasi data positif yang turun secara drastis seperti yang tergambar pada Gambar 3, namun cukup mampu mengklasifikasikan dataset yang dilabeli sebagai opini negatif.

IV. KESIMPULAN

Berdasarkan data dan hasil dari analisis serta pengujian yang dilakukan terhadap implementasi seleksi fitur pada dataset opini saham dengan menggunakan metode Naïve Bayes dan seleksi fitur Information Gain, maka disimpulkan bahwa proses seleksi fitur menggunakan Information Gain berhasil memperbaiki akurasi klasifikasi yakni mengalami peningkatan dari 42,86% menjadi 50%. Lebih lanjut, seleksi fitur Information Gain berhasil meningkatkan presisi pada data opini positif dan negatif yakni mencapai hingga 50% serta nilai recall pada data opini negatif hingga 98%. Namun, Information Gain belum mampu meningkatkan recall pada data opini positif.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 8, Nomor. 1, April 2023
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

DAFTAR PUSTAKA

Al-Thaqeb, S. A. (2018). Do international markets overreact? Event study: International market reaction to U.S. local news events. *Research in International Business and Finance*, *44*(July 2017), 369–385. https://doi.org/10.1016/j.ribaf.2017.07.106

Dimas Pratama, N., Sari, Y. A., & Adikara, P. P. (2018). *Analisis Sentimen Pada Review Konsumen Menggunakan Metode Naive Bayes Dengan Seleksi Fitur Chi Square Untuk Rekomendasi Lokasi Makanan Tradisional* (Vol. 2). Retrieved from http://j-ptiik.ub.ac.id

Jumeilah, F. S. (2018). Klasifikasi Opini Masyarakat Terhadap Jasa Ekspedisi JNE dengan Naïve Bayes. *JURNAL SISTEM INFORMASI BISNIS*, *8*(1), 92. https://doi.org/10.21456/vol8iss1pp92-98

Korde, V. (2012). Text Classification and Classifiers: A Survey. *International Journal of Artificial Intelligence & Applications*, *3*(2), 85–99. https://doi.org/10.5121/ijaia.2012.3208

Kumar, L., & Bhatia, P. K. (2013). Text Mining: Concepts, Process, and Applications. *Journal of Global Research in Computer Science*, *4*(3), 36–39. Retrieved from https://www.researchgate.net/publication/260341572

Maulida, I., Suyatno, A., & Rahmania Hatta, H. (2016). Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain. *JSM STMIK Mikroskil*, *17*, 249–258.

Negara, A. B. P., Muhardi, H., & Putri, I. M. (2020). Analisis Sentimen Maskapai Penerbangan Menggunakan Metode Naive Bayes dan Seleksi Fitur Information Gain. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 599–606. https://doi.org/10.25126/jtiik.202071947

Prakoso, B. S., Rosiyadi, D., Aridarma, D., Utama, H. S., Fauzi, F., & Qhomar, M. A. N. (2019). Optimalisasi Klasifikasi Berita Menggunakan Feature Information Gain Untuk Algoritma Naive Bayes Terhubung Random Forest. *Jurnal Pilar Nusa Mandiri*, *15*(2), 211–218. https://doi.org/10.33480/pilar.v15i2.684

Utami, L. D., & Wahono, R. S. (2015). Integrasi Metode Information Gain Untuk Seleksi Fitur dan Adaboost Untuk Mengurangi Bias Pada Analisis Sentimen Review Restoran Menggunakan Algoritma Naïve Bayes. *Journal of Intelligent Systems*, *1*(2).

Xie, C., & Wang, Y. (2017). Does Online Investor Sentiment Affect the Asset Price Movement? Evidence from the Chinese Stock Market. *Mathematical Problems in Engineering*, *2017*. https://doi.org/10.1155/2017/2407086

Yang, W., Lin, D., & Yi, Z. (2017). Impacts of the mass media effect on investor sentiment. *Finance Research Letters*, *22*, 1–4. https://doi.org/10.1016/j.frl.2017.05.001