Kajian Literatur: Identifikasi Konten Negatif Pada Twitter Dengan *Deep Learning*

Anggara Chandra Dwinata
Universitas Islam Indonesia
Jl. Kaliurang No.Km. 14,5, Yogyakarta
17523030@students.uii.ac.id

Ahmad Fathan Hidayatullah
Universitas Islam Indonesia
Jl. Kaliurang No.Km. 14,5, Yogyakarta
fathan@uii.ac.id

***Abstrak*—Media sosial telah menjadi media komunikasi antara satu pengguna dengan pengguna lainnya, salah satunya Twitter. Akan tetapi, tidak semua isi dari media sosial mengandung hal yang positif. Terdapat konten negatif yang beredar di sosial media, diantaranya adalah umpatan, pornografi, *cyberbullying*, *hate speech*, dan sebagainya. Telah banyak penelitian yang membahas tentang identifikasi data teks, dan menggunakan berbagai metode untuk mendapatkan hasil luaran yang diinginkan. Penelitian ini merupakan *literature review* untuk membandingkan beberapa penelitian sebelumnya tentang identifikasi teks pada sosial media. Selain itu, pada kajian literatur ini dilakukan perbandingan metode untuk mendeteksi konten negatif pada teks. Hasil *literature review* ini dapat dijadikan referensi pengembangan dalam identifikasi konten negatif pada teks.**

***Keywords*—Media sosial, Twitter, *deep learning*, identifikasi konten negatif**

I. PENDAHULUAN

Twitter adalah sosial media yang banyak digunakan masyarakat Indonesia untuk interaksi sosial jarak jauh. Di Indonesia, Twitter sendiri dimanfaatkan berbagai macam organisasi, instansi, maupun perorangan. Twitter menjadi wadah untuk mengungkapkan ekspresi, pendapat, dan lain sebagainya.

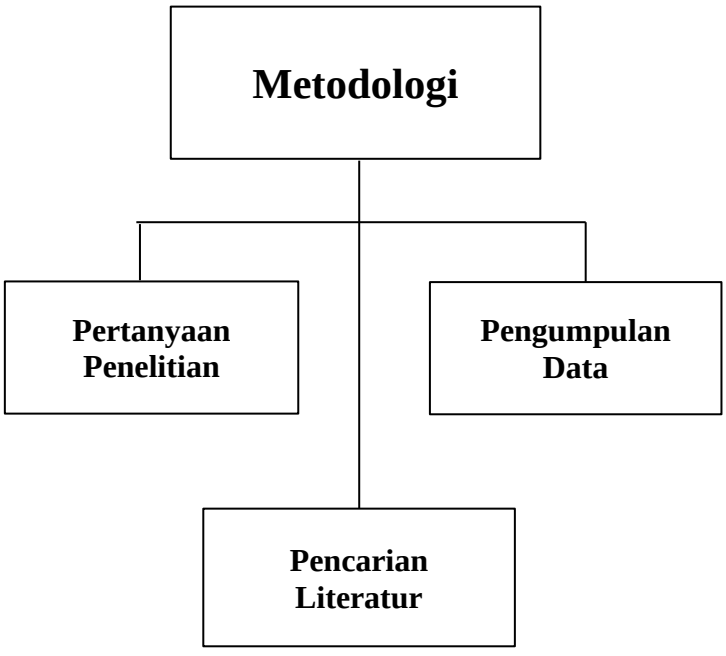
Namun, banyak pengguna Twitter tidak bijaksana untuk memilih kata-kata di tweet-nya. Tidak hanya di Twitter saja, di sosial media lain, seperti Instagram, Facebook, dan lain-lain tidak luput dari konten negatif. Banyak *netizen* yang menulis kata atau kalimat yang mengandung unsur SARA (Suku, Agama, Ras, dan Antargolongan), bahkan mengungkapkan ekspresi melalui bahasa yang kasar dan mengandung konten negatif. Definisi konten negatif adalah informasi yang bermuatan melanggar kesusilaan, penghinaan, perjudian ancaman, dan menyebarkan informasi palsu (*hoax*), serta mengakibatkan kerugian kepada pengguna. Dari permasalahan di atas yang berkaitan konten negatif di sosial media. Penelitian ini berupaya mencari tahu lebih dalam adakah konten negatif selain umpatan dan pornografi.

Dalam hal ini, melalui *literature review* ini akan mencari metode-metode dari *machine learning* dan *deep learning* yang digunakan peneliti-peneliti sebelumnya untuk identifikasi konten negatif.

Oleh karena itu, tujuan dari kajian literatur ini adalah mencari penerapan identifikasi konten negatif menggunakan *machine learning* atau *deep learning*, dan melihat tren terkini penerapan *deep learning* untuk identifikasi teks. Peneliti berharap dapat memberikan kontribusi terhadap peneliti lain dalam menentukan model untuk identifikasi konten negatif pada teks berbahasa Indonesia, dan mengetahui tren di bidang identifikasi teks.

II. METODOLOGI

Tujuan dari tinjauan literatur adalah mendapatkan landasan teori yang mendukung pemecahan atau solusi masalah yang diteliti. Teori yang sudah diperoleh, merupakan langkah awal peneliti agar lebih memahami masalah yang diteliti dengan tepat berdasarkan kerangka berpikir ilmiah. *Literature review* dikembangkan sebagai pendekatan untuk mengidentifikasi dan meninjau identifikasi konten negatif. Tinjauan literatur dilakukan dengan menggunakan standar tematik, eksplisit, dan ketat, yang bertujuan tidak hanya meringkas penelitian terkini tentang topik terkait, tetapi juga melibatkan sebuah elemen analitik yang kritis.



Gambar 1. Bagan metodologi

Bab metodologi sendiri terdiri dari pertanyaan penelitian, pencarian literatur, kriteria literatur, dan pengumpulan data. Semua yang disebutkan diatas, akan dijelaskan secara rinci di bab ini.

A. Pertanyaan Penelitian

Mengidentifikasi pertanyaan penelitian merupakan langkah pertama dari tinjauan sistematis. Pada langkah ini harus ringkas dan jelas. Dalam konteks penelitian ini, pertanyaan penelitian dapat dikemukakan sebagai berikut:

1) Penelitian tentang identifikasi konten negatif atau kasar apa yang terbaru? Siapa yang menerbitkan dan kapan?
2) Jenis konten negatif apa saja yang beredar di sosial media?
3) Metode apa yang biasanya digunakan untuk melakukan klasifikasi konten negatif?
4) Sumber data apa saja yang berisi konten-konten negatif?

B. Pencarian Literatur

Peneliti dalam mencari literatur menggunakan Google Scholar. Pencarian literatur dirancang sekitaran tahun 2016 sampai 2020.

Menurut pertanyaan penelitian, pencarian dilakukan dengan kata kunci, yaitu: identifikasi teks, klasifikasi teks, identifikasi konten negatif di sosial media. Alhasil, studi yang diselidiki menghasilkan 12 jurnal penelitian. Jurnal penelitian yang terpilih akan diilustrasikan pada tabel 1.

Dalam hal kriteria literatur, peneliti menentukan kriteria pencarian, yaitu berdasarkan metode dan kasus. Untuk metode, peneliti mencari dengan kata kunci "*deep learning*" dan *"machine learning"*. Sedangkan untuk kasus, kata kunci yang dicari adalah identifikasi atau klasifikasi konten negatif pada sosial media. Penelitian ini berfokus pada perbandingan *deep learning*.

Peneliti mencari makalah-makalah yang membahas tentang penggunaan *deep learning* dan *machine learning*, untuk identifikasi konten negatif pada Twitter. Pada bab hasil dan pembahasan, akan dijabarkan hasil luaran dari tiap penelitian sebelumnya.

C. Pengumpulan Data

Data yang dikumpulkan dari setiap penelitian untuk melakukan *review* identifikasi konten negatif diidentifikasi sebagai berikut:

1) Jurnal dan referensi yang lengkap
2) Penulis jurnal dan institusi mereka
3) Judul penelitian, tahun terbit, dan penerbit
4) Dataset, *domain*, dan sumber
5) Pendekatan terhadap identifikasi konten negatif
6) Proses dari *preprocessing*
7) Pemilihan fitur dan proses pembuatan

Di sini, peneliti ingin menganalisis terhadap 12 penelitian yang terbit dari bulan Juni 2016 - Desember 2020. Hasil analisis akan dibahas pada bab hasil dan pembahasan pada tabel 1.

Peneliti	Judul	Dataset	Metode
Trihapsari, E., Pembimbing, D., Magister, P., Telematika-cio, B. K., Elektro, J.T., & Industri, F. T. (2016)	Klasifikasi Cyber Bullying pada Media Sosial Twitter Dengan Menggunakan Algoritma Naïve Bayes	Twitter	Naïve Bayes
Purnamasari, N. M. G. D., Fauzi, M. A., Indriarti, & Dewi, L. S. (2018)	Identifikasi Tweet Cyberbullying pada Aplikasi Twitter Menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur	Twitter	Support Vector Machine (SVM), Information Gain (IG)
Putra, A. K. B. A., Fauzi, M. A., Setiawan, B. D., & Setiawan, E. (2018)	Identifikasi Ujaran Kebencian Pada Facebook Dengan Metode Ensemble Feature Dan Support Vector Machine	Facebook	Ensemble Feature, Support Vector Machine (SVM)
Hidayatullah, A. F., Aulia, A., Yusuf, F., Juwairi, K. P., Nayoan, R. A. N. (2019)	Identifikasi Konten Kasar pada Tweet Bahasa Indonesia	Twitter	Support Vector Machine (SVM), Naïve Bayes
Zamil. (2019)	Klasifikasi Kalimat Ofensif Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes Classifier	Twitter	Naïve Bayes
Hidayatullah, A. F., Hakim, A. M., & Sembada, A. A. (2019)	Adult Content Classification on Indonesian Tweets using LSTM Neural Network	Twitter	LSTM Neural Network
Chrismanto, A. R., & Lukito, Y. (2017)	Identifikasi Komentar Spam Pada Instagram	Instagram	Support Vector Machine (SVM), Naïve Bayes
Stiawan, M. M., & Hidayat, R. (2019)	Pengembangan Sistem Identifikasi Fakta Dan Tidak Fakta Berita di Media Informasi Berbahasa Indonesia	Media Informasi	Support Vector Machine (SVM)
Abdulloh, N. (2020)	Deteksi Cyberbullying pada Cuitan Media Sosial Twitter	Twitter	K-Nearest Neighbor (KNN), Support Vector Machine (SVM)

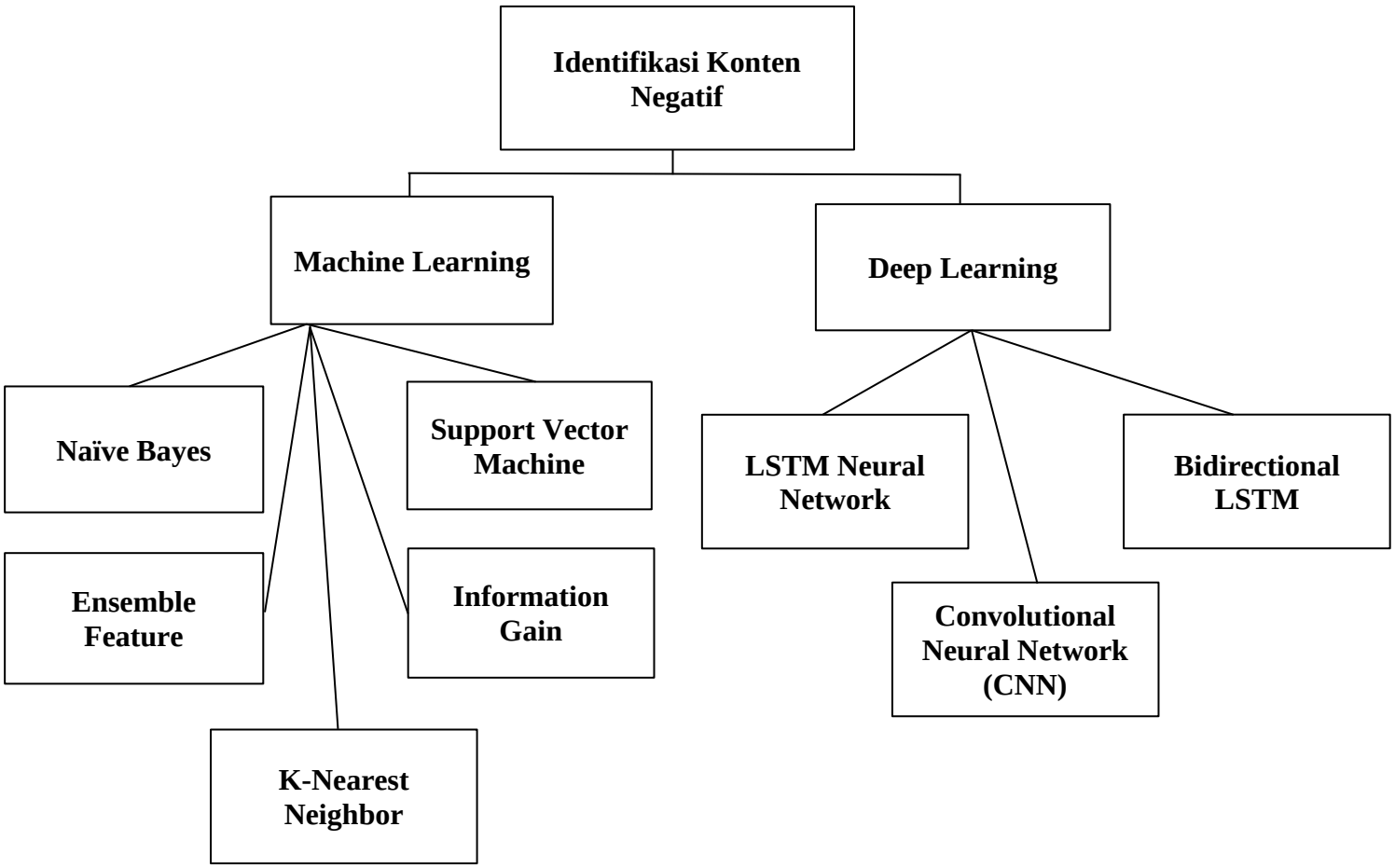
Ardiada, D., Sudarma, M., & Giriantari, D. (2019)	Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour	Twitter	Support Vector Machine (SVM), K-Nearest Neighbor (KNN)
Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. (2020)	Malicious Text Identification: Deep Learning from Public Comments and Emails	Komentar Sosial Media & Email	Long Short-Term Memory (LSTM)
Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J. (2018)	Deep Learning for Detecting Inappropriate Content In Text	Kolom Komentar pada Sosial Media	Bidirectional LSTM, Convolution Neural Network (CNN)

Tabel 1. Penelitian sebelumnya terkait dengan identifikasi konten negatif

Tabel 1 menunjukkan penelitian sebelumnya berdasarkan tahun, judul, dataset yang digunakan dan metode yang digunakan.

III. HASIL DAN PEMBAHASAN

Pada bab hasil dan pembahasan, akan memaparkan hasil dari tiap temuan penelitian sebelumnya mengenai identifikasi konten negatif. Hasil yang dibahas adalah *output* dan temuan yang telah ditemukan dari masing-masing penelitian.



Gambar 2. Pendekatan dan metode identifikasi konten negatif berdasarkan tabel 1

Penelitian yang dilakukan oleh Trihapsari, et al. [5] melakukan klasifikasi *cyber bullying* pada Twitter menggunakan metode *Naïve Bayes*, menghasilkan 87,67% dengan menggunakan pembobotan TF-IDF dengan minimal *term frequency* = 3. Purnamasari, et al. [2] melakukan penelitian tentang identifikasi *tweet cyber bullying* di Twitter, dengan SVM dan Information Gain (IG), menyimpulkan bahwa SVM mendapatkan hasil terbaik berdasarkan pengujian *iterMax*, parameter *lambda*, konstanta *gamma*, *epsilon* dan *complexity*.

Putra, et al. [3] dengan penelitian identifikasi ujaran kebencian pada Facebook dengan *Ensemble Feature* dan *Support Vector Machine*, mendapat parameter yang optimal sebesar 0,5 untuk *lambda*, 0,001 untuk *learning rate* dan 0,0001 untuk *epsilon*. Hidayatullah, et al. [1] meneliti identifikasi konten kasar pada Tweet Bahasa Indonesia, menghasilkan akurasi dari masing-masing model tersebut adalah 0,9834% dan 0,9982%. Kesimpulan dari penelitian tersebut adalah, *Support Vector Machine* dengan *linear kernel* lebih unggul secara keseluruhan dibandingkan dengan model *Naïve Bayes*.

Pada penelitian yang dilakukan Zamil. [4], menghasilkan akurasi tertinggi sebesar 84,00% dengan model *bigram*, kemudian dengan nilai *precision* pada model *trigram* sebesar 97,33% dan nilai *recall* tertinggi yaitu pada model *bigram* sebesar 81,48%.

Hidayatullah, et al. [6] melakukan penelitian tentang klasifikasi konten dewasa pada *tweet* berbahasa Indonesia menggunakan *LSTM Neural Network*, peneliti membangun empat model *LSTM* dengan skenario yang berbeda di setiap model *LSTM*. Setelah itu, membandingkan performa *LSTM* dengan metode *machine learning* yang tradisional, meliputi *Naïve Bayes*, *Logistic Regression* (LR) dan *Support Vector Machine* (SVM). Menurut penelitian tersebut, model yang terbaik didapatkan dari 2 lapisan *LSTM* dengan akurasi 98,38%. Pencapaian tersebut lebih tinggi 0,06% daripada akurasi menggunakan SVC model. Hasil dari l*oss value* menurun dari 12,88% menjadi 5,08% dan hasil akurasi meningkat dari 97,89% menjadi 98,39%.

Chrismanto dan Lukito [7] melakukan identifikasi komentar spam pada Instagram, menyimpulkan bahwa metode *Support Vector Machine* memiliki kinerja yang baik dibandingkan dengan *Naïve Bayes*, namun tidak terlalu signifikan peningkatannya. Tingkat akurasi antara *Naïve Bayes* dan *Support Vector Machine* berkisar antara 70-79% dimana kemampuan deteksi keduanya termasuk kategori yang baik. Akurasi klasifikasi dengan *Naïve Bayes* adalah 74,31% untuk skenario I (data tidak seimbang) dan sebesar 77,25% untuk skenario II (data seimbang). Terjadi peningkatan sebesar 2,94 untuk data yang seimbang.

Akurasi untuk klasifikasi dengan *Support Vector Machine* sebesar 78,49% untuk skenario I (data tidak seimbang) dan sebesar 75,78% untuk skenario II (data seimbang). Terjadi penurunan sebesar 2,71% untuk data seimbang. Untuk tahapan *preprocessing*, peneliti melakukan *setting encoding* teks ke *encoding unicode* (UTF-8), tokenisasi, *case folding*, *stop words removal*, *stemming*, dan konversi simbol-simbol, serta *emoticon*.

Stiawan dan Hidayat [8] dengan penelitian pengembangan sistem identifikasi fakta dan tidak fakta pada media informasi, menghasilkan pengujian dengan nilai precision dan recall diatas 87%. Setelah itu peneliti mengeluarkan hasil dari sistem mereka dengan probabilitas tidak fakta (hoaks): 0,9196343010150771, dan nilai fakta: 0,80365698984923.

Penelitian yang dilakukan oleh Abdulloh [9] meneliti tentang deteksi *cyberbullying* pada Twitter dengan *Linear SVM* sebagai algoritma terbaik dengan nilai *accuracy*, *precision*, *recall*, dan *F1-Score* paling tinggi dengan nilai masing-masing 0,997; 1,00; 1,00; 1,00. Ardiada, et al. [10] melakukan text mining pada Twitter untuk mendeteksi emosi pengguna dengan *Support Vector Machine* dan *K-Nearest Neighbour*, menyimpulkan kedua metode tersebut mengalami peningkatan nilai presisi, *recall*, *accuracy* dan kesesuaian yang signifikan. Hal tersebut terlihat ketika peneliti mendapatkan hasil dengan nilai rata-rata precision sebesar 0,4564, nilai recall 0,502 dan pada nilai accuracy sebesar 0,8104 dalam melakukan klasifikasi emosi.

Pada penelitian yang dilakukan oleh Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. [11] meneliti tentang identifikasi teks yang berbahaya pada komentar publik di sosial media dan email menggunakan *Long Short-Term Memory* (LSTM). Dari penelitian tersebut, menunjukkan model jaringan syaraf lebih unggul daripada model jaringan syaraf sederhana dan model LSTM menunjukkan hasil performa tertinggi diantara yang di laporkan pada literatur sebelumnya yang telah disebutkan pada penelitian ini.

Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J [12] melakukan deteksi konten yang tidak sopan berbentuk teks. Dari penelitian tersebut, peneliti menggabungkan dua metode *deep learning* yang disebut *Convolutional Bi-Directional LSTM (C-BiLSTM)* yang merupakan gabungan dari Convolutional Neural Network (CNN) dan Bidirectional LSTM. C-BiLSTM terbukti daripada arsitektur deep learning individu lainnya seperti CNN, LSTM, dan Bi-LSTM. Untuk menyaring percakapan yang tidak sopan, peneliti menggunakan model LSTM dan BLSTM. Mengevaluasi dengan model LSTM dan BLSTM pada data percakapan di dunia nyata mengungkapkan bahwa LSTM dan BLSTM mengungguli dari fitur berbasis *pattern* dan *hand-crafted* lainnya.

Berdasarkan penemuan penelitian-penelitian sebelumnya, pendekatan untuk identifikasi konten negatif dibagi menjadi dua, yaitu *Machine Learning*, dan *Deep Learning*. Berdasarkan tabel 1, untuk pendekatan *Machine Learning*, beberapa penelitian menggunakan metode *Naïve Bayes*, *Support Vector Machine (SVM)*, *Ensemble Feature*, *Information Gain*, *K-Nearest Neighbour (KNN)*, *Long Short-Term Memory (LSTM)*, *Bidirectional LSTM*, *Convolution Neural Network (CNN)*. Sedangkan untuk *Deep Learning*, menggunakan metode *LSTM Neural Network*. Dari macam-macam metode yang telah disebutkan di atas, akan digambarkan pada gambar 2.

IV. KESIMPULAN

Dalam kajian literatur ini, artikel penelitian tentang identifikasi atau klasifikasi teks dikaji secara sistematis. Dari 12 penelitian yang sudah dilakukan sebelumnya, di dominasi dengan pendekatan *machine learning*. Dikarenakan pendekatan *machine* learning menjadi tren untuk kasus identifikasi konten negatif pada sosial media daripada

pendekatan dengan *deep learning*. Untuk penggunaan *deep learning* hanya terdapat 2 penelitian, yaitu dengan menggunakan metode *Long Short-Term Memory (LSTM)*, *Bidirectional LSTM*, dan *Convolutional Neural Network (CNN)*.

REFERENSI

[1] Abdulloh, N. (2020). Deteksi Cyberbullying pada Cuitan Media Sosial Twitter.

[2] Ardiada, D., Sudarma, M., & Giriantari, D. (2019). Text Mining pada Sosial Media untuk Mendeteksi Emosi Pengguna Menggunakan Metode Support Vector Machine dan K-Nearest Neighbour. 18(1), 55–60.

[3] Baccouche, A., Ahmed, S., Sierra-Sosa, D., & Elmaghraby, A. (2020). Malicious text identification: Deep learning from public comments and emails. Information (Switzerland), 11(6). https://doi.org/10.3390/info11060312.

[4] Chrismanto, A. R., & Lukito, Y. (2017). Identifikasi Komentar Spam Pada Instagram. 8(3), 219–231.

[5] Hidayatullah, A. F., Aulia, A., Yusuf, F., Juwairi, K. P., Nayoan, R. A. N. (2019). Identifikasi Konten Kasar pada Tweet Bahasa Indonesia. 2(1), 1-5.

[6] Hidayatullah, A. F., Hakim, A. M., & Sembada, A. A. (2019). Adult Content Classification on Indonesian Tweets using LSTM Neural Network. 235–240.

[7] Purnamasari, N. M. G. D., Fauzi, M. A., Indriarti, & Dewi, L. S. (2018). Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya, 2(11), 5326-5332.

[8] Putra, A. K. B. A., Fauzi, M. A., Setiawan, B. D., & Setiawan, E. (2018). Identifikasi Ujaran Kebencian Pada Facebook Dengan Metode Ensemble Feature Dan Support Vector Machine. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 2(12).

[9] Stiawan, M. M., & Hidayat, R. (2019). Pengembangan Sistem Identifikasi Fakta Dan Tidak Fakta Berita di Media Informasi Berbahasa Indonesia. November, 34–39.

[10] Trihapsari, E., Pembimbing, D., Magister, P., Telematika-cio, B. K., Elektro, J.T., & Industri, F. T. (2016). KLASIFIKASI CYBER BULLYING PADA MEDIA SOSIAL TWITTER DENGAN MENGGUNAKAN CYBER BULLYING CLASSIFICATION ON TWITTER SOCIAL MEDIA USING NAÏVE BAYES ALGORITHM.

[11] Yenala, H., Jhanwar, A., Chinnakotla, M. K., & Goyal, J. (2018). Deep learning for detecting inappropriate content in text. International Journal of Data Science and Analytics, 6(4), 273–286. https://doi.org/10.1007/s41060-017-0088-4.

[12] Zamil. (2019). Klasifikasi Kalimat Ofensif Pada Media Sosial Twitter Menggunakan Metode Naïve Bayes Classifier Mhd. Desember.