Available online at: https://journal.ikippgriptk.ac.id/index.php/saintek

Jurnal Pendidikan Informatika dan Sains
13(1), 2024, 75-87
E-ISSN: 2407-1536
P-ISSN: 2089-2802

PERBANDINGAN PERFORMA DETEKSI CYBERBULLYING DENGAN TRANSFORMER, DEEP LEARNING, DAN MACHINE LEARNING

Fuad Muftie[1]*, Kamal Muftie Yafi[2], Qinthara Muftie Addina[3]
[1]Fakultas Teknologi Informasi, Universitas Nusa Mandiri, Jakarta, Indonesia
[2]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Indonesia, Jakarta, Indonesia
[3]Fakultas Pendidikan Bahasa dan Sastra Indonesia, Universitas Pendidikan Indonesia, Bandung, Indonesia
*email:14210197@nusamandiri.ac.id

Received: June 30, 2022 Accepted: June 25, 2024 Published: June 30, 2024

Abstrak

Peningkatan aktivitas browsing terutama di situs media sosial mengakibatkan rawannya terjadi *cyberbullying* (perundungan dunia maya). Telah banyak dilakukan penelitian untuk melakukan pendeteksian *cyberbullying*, baik dengan metode machine learning maupun deep learning. Dalam penelitian ini dilakukan perbandingan performa pengklasifikasian data teks apakah termasuk *cyberbullying* atau bukan, dengan menggunakan algoritma Transformer. Kemudian dilakukan perbandingan performa metode transformer dengan metode deep learning lain (RNN, LSTM, dan GRU) serta dengan metode machine learning (Naïve Bayes, *Logistic Regression*, SVM, dan *Decision Tree*). Hasil terbaik untuk model deep learning adalah dataset Youtube dengan model Transformer yang mendapat akurasi 98.49%. Kemudian hasil terbaik model machine learning adalah dataset Youtube dengan model SVM dan menggunakan feature Tf-Idf yang mendapat akurasi 97.82%.
Kata kunci: Transformers, *Sentiment Analysis, Natural Language Processing, Deep Learning*

Abstract

Increased browsing activity, especially on social media sites, makes cyberbullying prone to occur. Many studies have been carried out to detect cyberbullying, both with machine learning and deep learning methods. In this study, a comparison of the performance of classifying text data whether including cyberbullying or not is carried out using the Transformer algorithm. Then a comparison of the performance of the transformer method with other deep learning methods (RNN, LSTM, and GRU) and machine learning methods (Naïve Bayes, Logistic Regression, SVM, and Decision Tree) is carried out. The best result for the deep learning model is the Youtube dataset with the Transformer model which gets 98.49% accuracy. Then the best result of the machine learning model is the Youtube dataset with the SVM model and using the Tf-Idf feature which gets an accuracy of 97.82%.
Keywords: *Transformers, Sentiment Analysis, Natural Language Processing, Deep Learning.*

How to cite (in APA style): Muftie, F., Yafi, K. M., & Addina, Q. M. (2024). Perbandingan performa deteksi cyberbullying dengan transformer, deep learning, dan machine learning. *Jurnal Pendidikan Informatika Dan Sains, 13*(1), 75–87. https://doi.org/10.31571/saintek.v13i1.4002

DOI: 10.31571/saintek.v13i1.4002

PENDAHULUAN

Akhir-akhir ini terjadi peningkatan aktivitas *browsing* yang dilakukan oleh seluruh lapisan masyarakat, baik golongan tua, remaja, anak-anak, dan kaum wanita terutama di situs media sosial atau situs jejaring komunitas yang tentunya kegiatan tersebut rawan akan terjadinya *cyberbullying*.

Cyberbullying telah meningkat secara besar-besaran dikarenakan masuknya teknologi dan terutama media sosial di dunia maya. *Cyberbullying* sering kali dilakukan anonim dan hal ini tidak hanya dapat meningkatkan pengaruh sosial terhadap korban, tetapi juga dapat mengakibatkan depresi dan kecemasan sehingga memperburuk kesehatan mental korban (Jabeen & Treur, 2018). Komentar-komentar yang mengandung kata-kata kasar, sarkasme, ujaran kebencian dan rasisme dapat mempengaruhi psikologis mereka dan khususnya mempengaruhi tumbuh kembang anak-anak yang menjadi korban. *Cyberbullying* dapat dilakukan dengan berbagai macam cara seperti mempermalukan, menguntit, pemaksaan, eksploitasi atau mendominasi korban (Iwendi et al., 2020).

Untuk mengidentifikasi suatu tulisan mengandung bullying dapat digunakan beberapa algoritma yang tergolong dalam bidang Natural Language Processing (NLP) (Caselli et al., 2021). Di dalam bidang NLP, pengklasifikasian teks bertujuan untuk membuat analisa sentimen sehingga mendapatkan prediksi yang baik. Ada beberapa pendekatan Deep Learning untuk text classification diantaranya adalah Stacked Denoising Autoencoder yang digunakan oleh Amazon untuk mereview data sentimen analisis. Recursive Neural Network (RNNs) berdasarkan sebuah pohon syntax yang digunakan untuk hal yang sama. Pendekatan yang terkini adalah mengaplikasikan CovNets pada urutan kata atau urutan karakter. Pertama, CovNets tingkat kata membutuhkan penganalisa morphological terhadap bahasa target. Word embedding yang telah dilatih sebelumnya oleh unsupervised learning meningkatkan akurasi klasifikasi terhadap CovNet tingkat kata. Kemudian, CovNet tingkat karakter tidak memerlukan penganalisa morphologi, penganalisa syntax (Sato et al., 2018).

Di dunia maya, *cyberbullying* terjadi melalui media sosial, yang mana semua orang baik anak-anak, kaum muda, dan orang dewasa mengakses hampir sepanjang waktu. Sebuah penelitian kelompok yang disurvey antara bulan Juli dan Oktober 2016, menghasilkan temuan sebesar 34% siswa SMA mengalami *cyberbullying*. Untuk itu dalam penelitian ini dilakukan pendeteksian *cyberbullying* di dunia maya melalui pendekatan Deep Learning (Iwendi et al., 2020), khususnya menggunakan algoritma Transformer, untuk kemudian dibandingkan dengan model RNN, LSTM, GRU, serta model machine learning.

Penelitian terkait deteksi *cyberbullying* atau sentiment analisis telah banyak dilakukan oleh banyak peneliti. Model, metode, atau algoritma yang diusulkan juga sangat beragam, baik dengan machine learning, maupun deep learning. Penelitian-penelitian ini memberikan konteks yang lebih komprehensif terhadap metodologi yang diadopsi dalam penelitian ini, yaitu perbandingan performa algoritma Transformer dengan metode deep learning dan machine learning lainnya dalam deteksi cyberbullying.

Iwendi et.al (Iwendi et al., 2020) menjelaskan dua aturan untuk ekstraksi fitur yang digunakan untuk mendeteksi komentar negatif dan ofensif yang sering diarahkan kepada korban. Metode yang diusulkan adalah pelabelan data dan komentar yang diterapkan dengan *crowdsourcing* dimana kata-kata kasar dideteksi secara *real time*. Metode yang diterapkan diklaim telah menghasilkan identifikasi yang efisien terhadap deteksi *cyberbullying* melalui analisis empiris *deep learning*. Empat metode *deep learning* yang digunakan yaitu Bidirectional Long Short-Term Memory (BLSTM), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM) dan Recurrent Neural Network (RNN). Langkah-langkah yang dilakukan yaitu melakukan *pre-processing* data terlebih dahulu dengan pembersihan teks, tokenisasi, *stemming*, *lemmatization* dan penghapusan *stopwords*. Setelah dilakukan *pre-processing*, data yang sudah bersih digunakan untuk melakukan prediksi. Hasilnya menunjukkan bahwa BLSTM mencapai akurasi dan skor pengukuran F1 lebih tinggi dibandingkan dengan RNN, LSTM dan GRU.

Zhong et al. (Zhong et al., 2019) melakukan studi untuk mendeteksi adanya insiden *cyberbullying* di jejaring sosial media. Metode yang diusulkan adalah melakukan beberapa varian dari pendekatan *convolutional neural networks* (CNN) yang disesuaikan. Model tersebut akan memproses komentar pengguna jejaring sosial secara independen di front-end layer, model juga akan memperhitungkan kemungkinan pola-pola percakapannya. Output dari front-end layer dari model

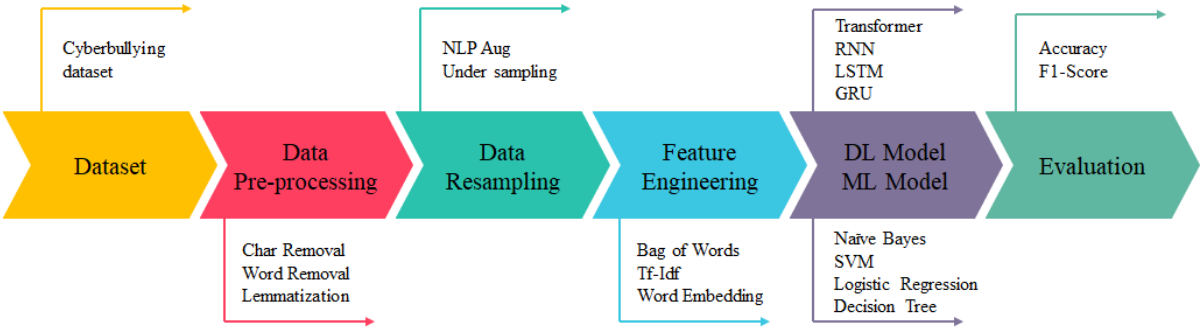
akan di digabungkan dengan layer yang dirancang yang diberi nama max layer / novel sorting layer. Model CNN yang dimodifikasi tersebut berhasil mengungguli baseline CNN dengan nilai akurasi mencapai 84,29% untuk *cyberbullying* dan 83,08% untuk cyberaggression.

Sato et al. (Sato et al., 2018) melakukan eksperimen pada pengklasifikasian teks bahasa Jepang menggunakan ConvNets dan meneliti apa yang diekstrak oleh ConvNets pada character-level dan dari *transfer learning* untuk dimanfaatkan dengan baik. Metode yang digunakan, dengan menggunakan dua jenis representasi masukan, yang pertama adalah simple one-hot representation. Yang kedua adalah distributed representation *(character-level embedding)* untuk menghilangkan proses romanisasi dalam dataset bahasa Jepang. Hasil penelitian yang dilakukan mendapat peningkatan akurasi klasifikasi dengan menggunakan character-level ConvNets kepada corpus bahasa jepang dalam skala besar dibandingkan dengan metode klasifikasi classic-text.

Pendekatan dengan model Transformer juga telah dilakuka pada deteksi ulasan palsu. Kennedy et al. (Kennedy et al., 2020) menggunakan model Transformer pada dua dataset yaitu dataset OpSpam dan Yelp untuk mendeteksi ulasan palsu. Selain dengan transformer, digunakan juga dua pendekatan model yaitu model non-neural (Logistic Regression dan Support Vector Machine), serta model neural network (Feed Forward Neural Network, Convolutional Neural Network dan Long Short-Term Memory). Adapun model transformer yang digunakan adalah model Transformer (BERT). Hasilnya model Transformer dengan dataset OpSpam memperoleh hasil akurasi tertinggi yaitu dengan nilai 90,5%. Penelitian ini menyoroti urgensi untuk mengevaluasi dan membandingkan performa berbagai model machine learning dan deep learning dalam mendeteksi cyberbullying, guna meningkatkan akurasi dan efisiensi deteksi dalam skenario dunia nyata. Penggunaan model Transformer diharapkan mampu memberikan hasil yang lebih baik dibandingkan metode sebelumnya, sehingga dapat berkontribusi pada upaya mitigasi cyberbullying di platform media sosial.

METODE

Metode penelitian yang dilakukan mengadopsi *machine learning pipeline* (Elsafoury et al., 2021) yang merupakan serangkaian langkah-langkah berurutan yang membentuk alur kerja *machine learning*, terdiri dari pengumpulan data, pra-pemrosesan data, pemilihan fitur, pelatihan model, dan evaluasi model. Kemudian untuk mengakomodasi kebutuhan, ditambahkan langkah data resampling, sehingga alur penelitian menjadi seperti digambarkan pada gambar 1.



Gambar 1. Diagram Metode Penelitian

Dataset

Dataset yang digunakan dalam ini diambil dari dataset *cyberbullying* (Elsafoury, 2020), yang merupakan kumpulan dataset dari berbagai sumber terkait dengan pendeteksian *cyberbullying* pada teks. Dataset tersebut bersumber dari berbagai platform media sosial seperti Kaggle, Twitter, Wikipedia Talk, dan YouTube. Data tersebut berisi teks dan diberi label sebagai *cyberbullying* atau bukan. Kategori *cyberbullying* yang ada dalam data tersebut antara lain ujaran kebencian, agresi, penghinaan, dan toksisitas. Namun dalam penelitian ini yang diambil hanya kategori binary 0 dan 1.

Label 0 adalah label untuk teks yang netral atau tidak mengandung *cyberbullying* dan label 1 adalah untuk teks yang mengandung unsur *cyberbullying*.

Dataset yang diperoleh terdiri dari delapan set yaitu *aggression*, *attack*, kaggle, *toxicity*, twitter, twitter-racism, twitter-sexism, dan youtube. Namun beberapa dataset memiliki data teks dan label binary yang sama sehingga hanya lima dataset yang digunakan dalam penelitian ini yaitu *aggression*, kaggle, *toxicity*, twitter, dan youtube dataset.

Tabel 1. Dataset *Cyberbullying*

Dataset	Jumlah Data	Jumlah Label 0	Jumlah Label 1
aggression_parsed_dataset	115.864	101.082	14.782
attack_parsed_dataset	115.864	102.274	13.590
kaggle_parsed_dataset	8.799	5.993	2.806
toxicity_parsed_dataset	159.686	144.324	15.362
twitter_parsed_dataset	16.851	11.501	5.347
twitter_racism_parsed_dataset	13.471	11.501	1.970
twitter_sexism_parsed_dataset	14.881	11.501	3.377
youtube_parsed_dataset	3.464	3.047	417

Data *Pre-processing*

Data Preprocessing yang pertama kali dilakukan adalah mengidentifikasi komposisi dari teks sehingga dapat ditentukan bagian-bagian teks yang perlu dihapus, akan dipertahankan, dan yang perlu dilakukan normalisasi. Proses ini akan memudahkan dalam membuat urutan proses pembersihan dan normalisasi teks agar data berupa teks dapat digunakan dalam proses klasifikasi menggunakan deep learning maupun machine learning.

Data teks awal yang ada dalam dataset masih dapat dikategorikan sebagai data yang tidak berkualitas karena tidak terstruktur, banyak noise, dan banyak kesalahan *grammar* (Birjali et al., 2021). Noise yang dapat teridentifikasi diantaranya masih adanya berbagai karakter yang tidak bermanfaat untuk digunakan dalam *text classification*. Karakter yang harus dibuang antara lain emoji, unicode, tag html, angka, *stopwords*, karakter non-ASCII, karakter asing/*accented character*, hashtag, dan tanggal. Kemudian untuk proses normalisasi teks dilakukan *lemmatization* untuk mengembalikan kata bahasa Inggris ke dalam akar katanya. Selain itu data ganda (*duplicate*) juga dihapus untuk mendapatkan gambaran komposisi yang utuh antara label 0 dan label 1. Setelah proses pembersihan dan normalisasi, atas data teks yang sangat pendek (satu atau dua kata) dilakukan penghapusan.

Setelah dilakukan data preprocessing diperoleh lima dataset yang akan digunakan dalam pengujian model Transformer, deep learning, dan machine learning dengan jumlah data dan jumlah label sebagaimana dimaksud pada Tabel 2.

Tabel 2. Dataset dan Jumlah Data Setelah Data *Pre-processing*

Dataset	Jumlah data	Jumlah label 0	Jumlah label 1
Aggression	113,467	98,971 (87%)	14,496 (13%)
Kaggle	8,532	5,855 (69%)	2,677 (31%)
Toxicity	155,746	140,673 (90%)	15,073 (10%)
Twitter	15,763	10,556 (67%)	5,207 (33%)
Youtube	3,441	3,027 (88%)	414 (12%)

Data *Resampling*

Sebagaimana terlihat pada tabel 2, dataset yang diperoleh merupakan data yang tidak seimbang antara label 0 dan label 1. Seluruh dataset memiliki label 0 sebagai mayoritas dan label 1 sebagai minoritas. Untuk mendapatkan data yang *balance* dilakukan *oversampling* dengan augmentasi data

menggunakan nlpaug (Ma, 2019). Augmentasi data teks dipilih dengan tujuan untuk mendapatkan data yang berkualitas, tidak hanya sekadar menduplikasi data dengan label 1 agar seimbang dengan data dengan 0 sebagaimana dilakukan dengan berbagai teknik *oversampling* (Kovács, 2019).

Proses augmentasi data teks pada dasarnya terinspirasi dari augmentasi pada bidang *computer vision* yang berhasil meningkatkan performa model dengan augmentasi (Wei & Zou, 2019). Proses augmentasi teks yang dilakukan antara lain menggunakan metode *similarity* (mengubah kata dengan kata yang similar berdasarkan data Google News word vector), *back translation* (menerjemahkan teks ke Bahasa Jerman dan diterjemahkan balik ke Bahasa Inggris), *contextual* (mengganti kata tertentu sesuai konteks kalimat), *random deletion* (menghapus sebagian kata dalam teks), dan *spelling mistake* (mengganti kata tertentu dengan kata yang salah eja).

Untuk dataset *aggression* dan *toxicity* setelah dilakukan *oversampling* mengakibatkan proses komputasi yang berat karena jumlah datanya menjadi sangat besar. Sehingga mengakibatkan proses *training* klasifikasi pada model *deep learning* maupun *machine learning* terhenti atau *hang* akibat terbatasnya memori komputer. Selanjutnya untuk kedua dataset tersebut dilakukan undersampling menggunakan RandomUnderSampler dari imbalanced-learn.

Feature Engineering

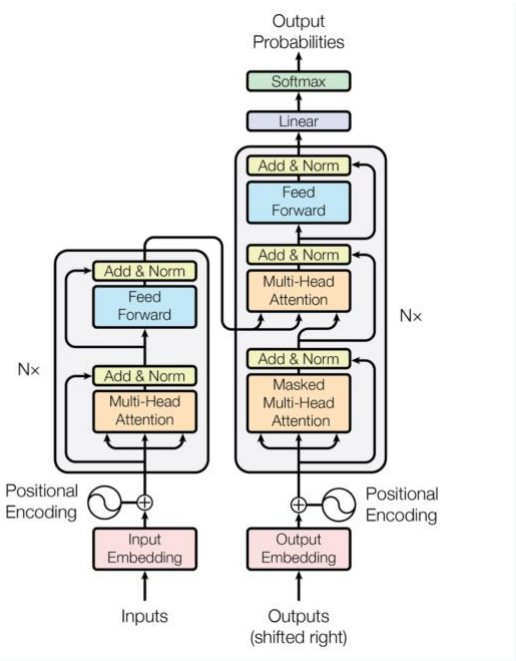
Agar data berupa teks dapat digunakan dalam algoritma *deep learning* maupun *machine learning*, teks harus diubah menjadi sebuah matrix atau vector. Metode yang digunakan dalam penelitian ini adalah *bag of words*, *term frequency-inverse document frequency* (Tf-Idf), dan *word embedding*.

Model *Bag of Words* mewakili kumpulan teks sebagai matriks. Jumlah baris dalam matriks sama dengan jumlah teks, dan jumlah kolom sama dengan jumlah kata dari masing-masing teks (kecuali *stopwords*). Model Tf-Idf pada prinsipnya mirip dengan *Bag of Words*, hanya nilai yang ada dalam matriks berisi nilai sesuai algoritma Tf-Idf. Model *word embedding* menggunakan Word2Vec yang dibuat berdasarkan pada frekuensi kata-kata dalam konteks yang sama (Glazkova, 2020).

Model *Transformer*, *Deep Learning*, dan *Machine Learning*

Pada tahap ini dilakukan pemodelan deep learning menggunakan Transformer (Vaswani et al., 2017). Model Transformers berisi lapisan encoder dan decoder, dimana masing-masing terhubung ke lapisan multi-head dan memberi masukan pada layer-layer yang terhubung secara *fully connected*. Model akan mengingat posisi dan urutan kata dengan bantuan fungsi cosine dan membuat pengkodean posisi tersebut. Di layer encoder dan decoder multi-head menerapkan mekanisme yang disebut *self-attention* yang menjadi masukan dan dibagi menjadi tiga lapisan yang terhubung untuk membuat vektor Query (Q), Key (K) dan Value (V) (Chugh et al., 2021).

Gambar 2 menunjukkan model transformer yang dikembangkan oleh Vaswani et al. (Vaswani et al., 2017).



Gambar 2. Model dan Arsitektur Transformer

Kemudian dibangun model transformer pada lingkungan Jupyter Notebook dengan gambaran model seperti pada Gambar 2. Model terdiri dari layer input dan output, dengan enam hidden layer yaitu layer TokenAndPositionEmbedding, layer TransformerBlock, layer GlobalAveragePooling1D, layer Dropout, layer Dense, dan layer Dropout.

```
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 300)]             0

 token_and_position_embeddin  (None, 300, 128)         5718400
 g (TokenAndPositionEmbeddin
 g)

 transformer_block (Transfor  (None, 300, 128)         140832
 merBlock)

 global_average_pooling1d (G  (None, 128)              0
 lobalAveragePooling1D)

 dropout_2 (Dropout)         (None, 128)               0

 dense_2 (Dense)             (None, 32)                4128

 dropout_3 (Dropout)         (None, 32)                0

 dense_3 (Dense)             (None, 1)                 33

=================================================================
Total params: 5,863,393
Trainable params: 5,863,393
Non-trainable params: 0
```

Gambar 3. Model Transformer yang Dibangun

Sebagai pembanding performa Transformer dilakukan perbandingan dengan model deep learning lain yaitu RNN, LSTM, dan GRU. Untuk itu dibuat model dari deep learning masing-masing. Dari masing-masing model tersebut dilakukan dua pengujian yaitu tanpa pembobotan (*weights*) dan dengan pembobotan menggunakan GloVe *word embedding*. Dengan demikian ada enam model pengujian yaitu model RNN, model RNN + GloVe, model LSTM, model LSTM + GloVe, model GRU, dan model GRU + GloVe. Implementasi GloVe dalam model tersebut berfungsi sebagai *weights* dalam *Embedding Layers* yang merupakan hidden layer pertama pada model deep learning yang dibangun.

Selain membandingkan performa Transformer dengan Deep Learning, dilakukan juga perbandingan menggunakan algoritma *Machine Learning* yaitu SVM, Naïve Bayes, *Logistic Regression*, dan *Decision Tree*. Pemilihan keempat model tersebut didasarkan pada perbedaan jenis supervised learning dari masing-masing model (Mehta & Pandya, 2020), yaitu pendekatan linear menggunakan SVM, pendekatan probabilitas menggunakan Naïve Bayes, pendekatan *decision tree*, dan pendekatan regresi menggunakan *Logistic Regression*.

Sebagai fitur dalam model *machine learning* ini digunakan *Bag of Word* (BoW), TFIdf, dan *Word Embedding*. Masing-masing fitur tersebut akan diaplikasikan pada keempat model *machine learning* untuk mendapatkan performanya.

Evaluation

Metrics yang digunakan untuk mengukur performa Transformer, deep learning, dan machine learning digunakan *accuracy* dan F1 *score* sebagai acuan. Khusus untuk F1 Score, metrics ini dipilih mengingat dataset yang diperoleh merupakan dataset yang *imbalance*. Sehingga F1 Score diharapkan dapat memberikan gambaran performa yang lebih baik daripada accuracy saja. F1 score telah dianggap lebih penting karena menggabungkan *precission* dan *recall* dengan memberikan skor antara 0 dan 1 (Rupapara et al., 2021). Dengan dataset yang imbalance, jika hanya mengacu pada accuracy saja, maka dapat memberikan gambaran yang tidak sebenarnya, karena hanya dengan membuat prediksi atas data mayoritas saja (label 0) akan diperoleh probabilitas yang tinggi untuk mendapatkan nilai akurasi yang tinggi juga.

HASIL DAN PEMBAHASAN

Hasil

Hasil pengujian model Transformer dan model deep learning

Dari pengujian menggunakan berbagai model baik deep learning maupun machine learning diperoleh hasil sebagaimana ditampilkan pada Tabel 2.

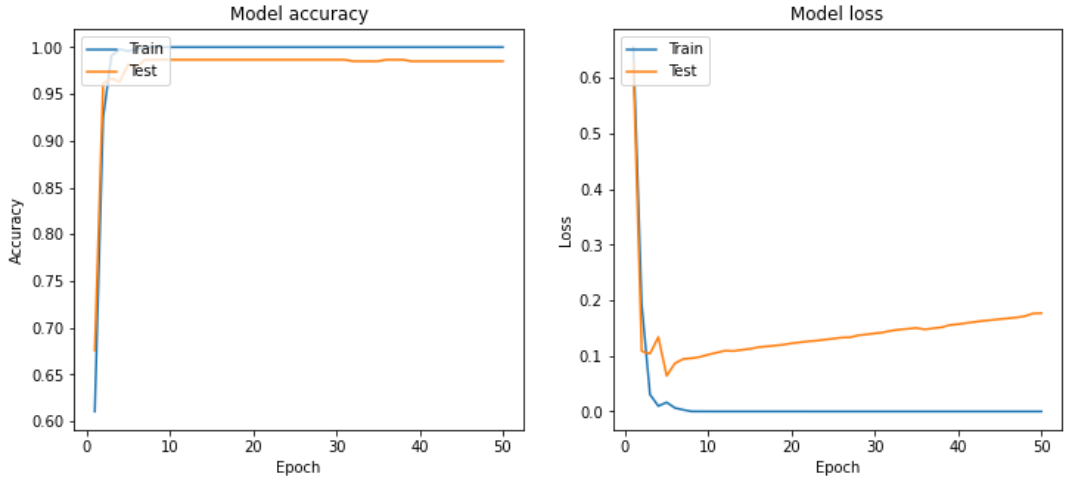
Pada model transformer diperoleh nilai terbaik pada dataset youtube yaitu akurasi sebesar 98.49% dan F1 Score 98.54%. Nilai akurasi dan F1 score yang berdekatan ini menunjukkan model Transformer tersebut tidak overfit. Grafik model akurasi model ini menunjukkan nilai akurasi yang stabil dari epoch 10 sampai dengan epoch 50. Grafik model loss untuk *train* data menunjukkan angka yang stabil dari epoch 10 sampai epoch 50, sedangkan untuk test data menunjukkan angka yang semakin menaik dari epoch 10 sampai epoch 50.

Sementara pada model deep learning dan dataset lainnya diperoleh nilai terbaik sebagai berikut: Dataset Aggression mendapat nilai terbaik pada model RNN + GloVe dengan nilai accuracy 86.83% dan F1 score 86.65%. Dataset Kaggle mendapat nilai terbaik pada model LSTM dengan nilai accuracy 88.20% dan F1 score 87.43%. Dataset Toxicity mendapat nilai terbaik pada model GRU + GloVe dengan nilai accuracy 90.48% dan F1 score 90.13%. Dataset Twiter mendapat nilai terbaik accuracy 84.38% pada model GRU + GloVe sementara F1 score 84.32% pada model LSTM + GloVe.

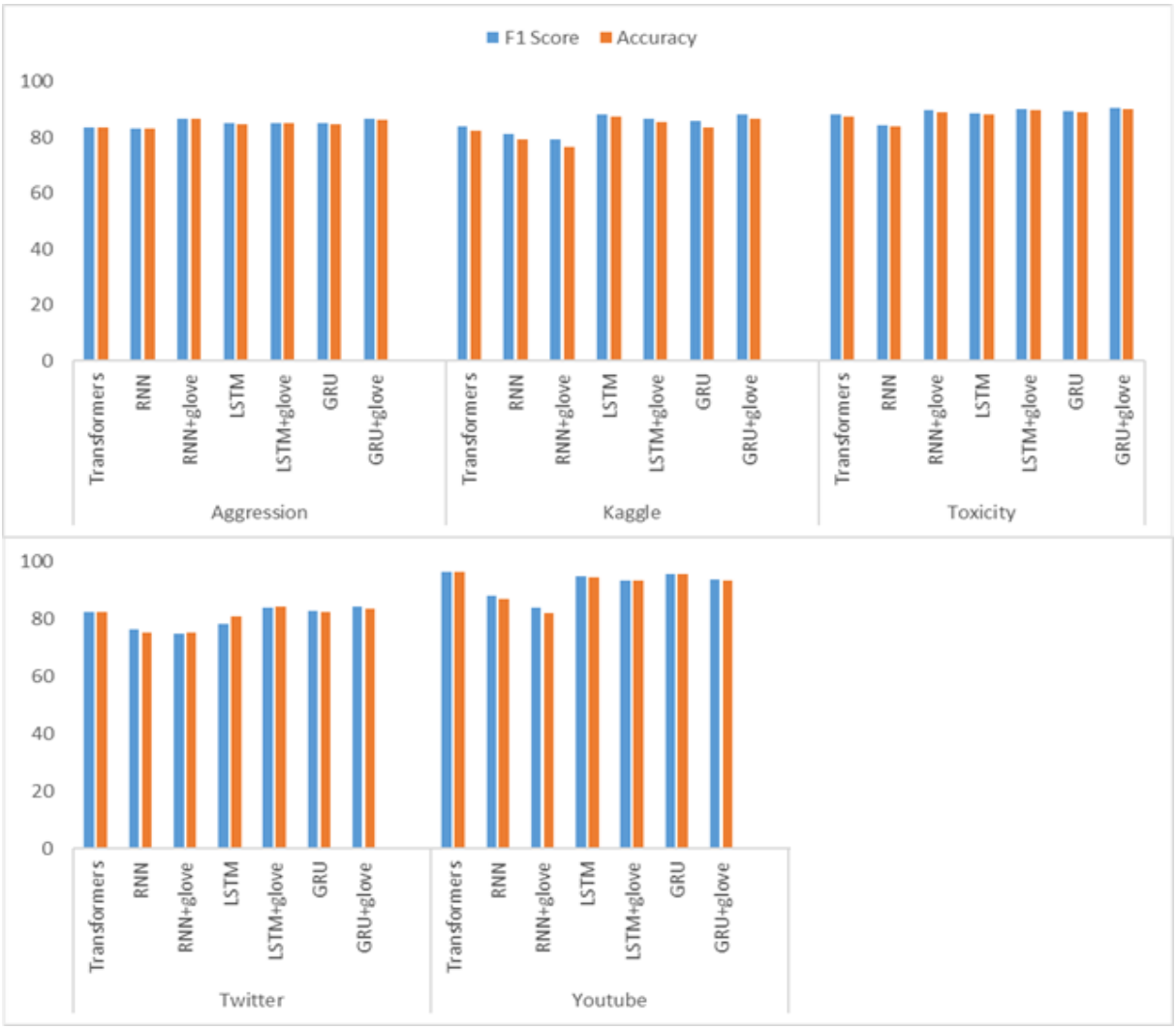
Tabel 3. F1 Score dan Akurasi Deep Learning

	Aggression		Kaggle		Toxicity		Twitter		Youtube	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Transformers	83.72	83.73	83.83	82.55	88.41	87.62	82.41	82.58	**98.54**	**98.49**
RNN	83.08	83.05	81.32	79.36	84.43	83.92	76.41	75.29	94.12	93.95
RNN+ GloVe	**86.83**	**86.65**	79.47	76.83	89.90	89.20	74.95	75.40	83.90	82.52
LSTM	85.05	84.79	**88.20**	**87.43**	88.62	88.21	78.49	80.99	94.06	93.78
LSTM+GloVe	85.34	85.29	86.84	85.55	90.19	89.96	84.18	**84.32**	95.51	95.29
GRU	85.19	84.94	85.90	83.58	89.46	89.24	82.79	82.52	95.99	95.80
GRU+GloVe	86.66	86.29	88.12	86.77	**90.48**	**90.13**	**84.38**	83.63	95.35	95.13

RNN = Recurrent Neural Networks; LSTM = Long Short Term Memory; GRU = Gated Recurrent Unit
GloVe = Global Vectors for Word Representation



Gambar 4. Grafik *Model Accuracy* dan *Model Loss* pada Transformer untuk dataset Youtube



Gambar 5. Grafik Hasil Pengujian Transformer dan Deep Learning

Hasil pengujian model machine learning

Hasil pengujian dengan machine learning ditampilkan pada Tabel 4. Model machine learning terbaik untuk masing-masing dataset. Dataset Aggression menghasilkan akurasi terbaik sebesar 87.18% dan F1 score sebesar 86.56% dengan menggunakan SVM classifier dan menggunakan feature TfIdf. Dataset Kaggle menghasilkan akurasi terbaik sebesar 86.72% dan F1 score sebesar 87.15% dengan menggunakan SVM classifier dan menggunakan feature TfIdf. Dataset Toxicity menghasilkan akurasi terbaik sebesar 89.66% dan F1 score sebesar 89.55% dengan menggunakan Logistic Regression classifier dan menggunakan feature Bag of Word. Dataset Twitter menghasilkan akurasi terbaik sebesar 85.30% dan F1 score sebesar 85.10% dengan menggunakan SVM classifier dan menggunakan feature TfIdf. Dataset Youtube menghasilkan akurasi terbaik sebesar 97.82% dan F1 score sebesar 97.78% dengan menggunakan SVM classifier dan menggunakan feature TfIdf.

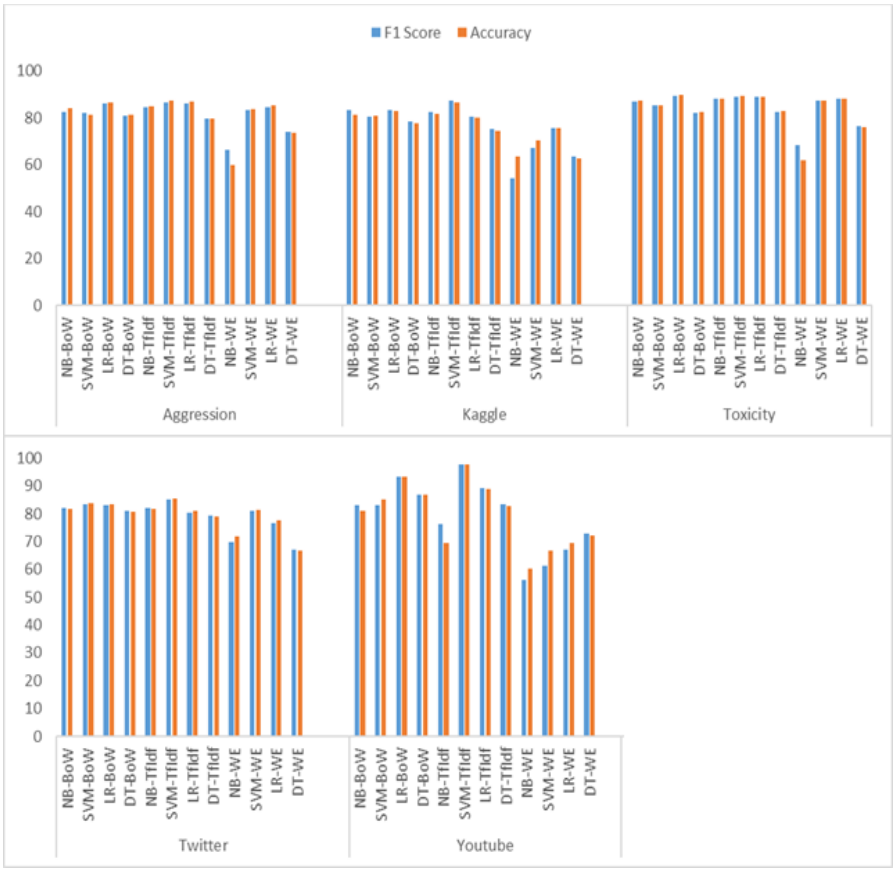
Tabel 4. F1 Score dan Akurasi Machine Learning

	Aggression		Kaggle		Toxicity		Twitter		Youtube	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
NB-BoW	82.51	83.94	83.21	81.32	86.92	87.30	82.02	81.71	83.22	80.92
SVM-BoW	82.10	81.30	80.43	80.99	85.34	85.12	83.41	83.84	82.96	85.13
LR-BoW	86.26	86.70	83.39	83.01	**89.55**	**89.66**	82.92	83.26	93.39	93.28
DT-BoW	81.00	81.27	78.40	77.71	82.20	82.48	80.89	80.70	86.88	86.72
NB-TfIdf	84.61	84.90	82.47	81.75	88.02	88.16	82.01	81.55	76.28	69.58
SVM-TfIdf	**86.56**	**87.18**	**87.15**	**86.72**	89.04	89.25	**85.10**	**85.30**	**97.78**	**97.82**
LR-TfIdf	86.25	86.79	80.35	80.24	88.78	89.00	80.26	80.94	89.03	88.74
DT-TfIdf	79.61	79.72	75.35	74.33	82.67	82.81	79.35	78.96	83.53	82.77
NB-WE	66.38	59.79	54.42	63.44	68.20	61.90	69.71	71.78	56.22	60.34
SVM-WE	83.25	83.90	67.05	70.34	87.24	87.30	80.93	81.26	61.43	66.55
LR-WE	84.70	85.17	75.73	75.79	88.12	88.17	76.72	77.46	67.09	69.58
DT-WE	73.81	73.68	63.71	62.74	76.38	76.19	67.09	66.84	72.89	72.18

NB = Naïve Bayes; SVM = Support Vector Machine; LR = Logistic Regression; DT = Decision Tree
BoW = Bag of Word; TfIdf = Term Frequency-Inverse Document Frequency; WE = Word Embedding

Beberapa keterbatasan yang ada dalam penelitian ini dapat menghasilkan bias pada hasil yang diperoleh pada setiap model. Beberapa keterbatasan diantaranya besarnya dimensi input pada setiap model, distribusi jumlah kata pada setiap dataset, dan belum adanya evaluasi atas hasil klasifikasi yang tidak sesuai.

Hasil pengujian dengan model transformer, deep learning, dan machine learning untuk masing-masing dataset masih menggunakan model yang sama dengan dimensi input sebesar 300. Model yang dibangun belum mempertimbangkan distribusi jumlah kata pada setiap dataset yang tersedia. Distribusi jumlah kata dalam setiap dataset setelah *pre-processing* dapat dilihat pada Tabel 5.



Gambar 6. Grafik Hasil Pengujian Machine Learning

Tabel 5. Distribusi Jumlah Kata Dataset

Dataset	count (row)	mean	std	min	q1	med	q3	max
Aggression	113.467	36,25	71,97	1	8	18	38	2,500
Kaggle	8.532	15,24	30,03	1	4	8	16	1,357
Toxicity	155.746	35,03	53,99	1	9	18	39	1,250
Twitter	15.763	7,67	3,49	1	5	8	10	22
Youtube	3.441	111,45	121,91	1	30	78	154	2,303

Distribusi jumlah kata dataset aggression, kaggle, toxicity, dan youtube adalah distribusi tidak normal, melainkan distribusi miring positif / miring kanan / positively skewed distribution. Jumlah kata max dalam tabel merupakan data outlier mengingat terpaut jauh dengan median dan kuartil 3. Sementara untuk distribusi jumlah kata pada dataset twitter mendekati distribusi normal.

Sebagai akibat dari penetapan jumlah dimensi input sebesar 300 adalah banyaknya matrix dari data text tersebut hanya berisi angka 0. Hal ini dapat mempengaruhi proses *training* model. Untuk penelitian berikutnya distribusi jumlah kata dalam dataset perlu dipertimbangkan sehingga dimensi input matrix pada model dapat ditentukan dengan nilai yang optimal.

Pembahasan

Model Transformer memberikan performa yang sangat baik pada dataset YouTube, mengungguli model deep learning dan machine learning lainnya. Hal ini sejalan dengan penelitian sebelumnya yang menunjukkan bahwa model Transformer, khususnya BERT, dapat memberikan performa superior dalam berbagai tugas NLP karena kemampuannya untuk memahami konteks

secara lebih baik melalui mekanisme self-attention (Vaswani et al., 2017; Shaikh et al., 2023; Dadvar & Eckert., 2020). Penggunaan model Transformer diharapkan mampu memberikan hasil yang lebih baik dibandingkan metode sebelumnya, sehingga dapat berkontribusi pada upaya mitigasi cyberbullying di platform media sosial.

Namun, hasil ini juga mengindikasikan bahwa tidak ada satu model yang selalu terbaik untuk semua jenis dataset. Misalnya, pada dataset Aggression dan Twitter, model GRU+GloVe memberikan performa yang lebih baik dibandingkan model lainnya. Ini menunjukkan bahwa pemilihan model harus disesuaikan dengan karakteristik spesifik dari dataset yang digunakan. Penelitian oleh Iwendi et al. (2020) juga menunjukkan bahwa model yang berbeda memiliki keunggulan pada dataset yang berbeda, menggarisbawahi pentingnya pendekatan yang disesuaikan.

Beberapa keterbatasan dalam penelitian ini mencakup besarnya dimensi input pada setiap model dan distribusi jumlah kata yang tidak merata pada setiap dataset. Distribusi kata yang tidak normal, terutama pada dataset seperti Aggression dan YouTube, dapat mempengaruhi performa model. Selain itu, belum adanya evaluasi atas hasil klasifikasi yang tidak sesuai atau tidak terdeteksi dengan baik oleh model juga menjadi keterbatasan yang perlu diperhatikan. Distribusi yang tidak normal dapat menyebabkan bias dalam pelatihan model, seperti yang diidentifikasi dalam penelitian oleh Kennedy et al. (2020) yang menunjukkan bahwa distribusi data yang tidak seimbang dapat menurunkan performa model. Selain itu pengujian belum melakukan evaluasi atas data teks yang belum sesuai diklasifikasikan, atau belum terdeteksi dengan baik oleh model sebagai bullying atau bukan bullying. Beberapa hal yang patut dilakukan investigasi lebih lanjut salah satunya adalah adanya teks yang bukan berbahasa Inggris atau bercampur antara bahasa Inggris dengan bahasa lain dalam satu teks. Salah satu upaya yang telah diupayakan adalah melakukan deteksi bahasa dengan menggunakan modul Python Spacy language detector dan Googletrans, namun belum memberikan hasil yang akurat sehingga tidak dilanjutkan.

Untuk penelitian berikutnya, disarankan untuk mempertimbangkan distribusi jumlah kata dalam dataset sehingga dimensi input matrix pada model dapat ditentukan dengan nilai yang optimal. Selain itu, perlu dilakukan investigasi lebih lanjut mengenai teks yang tidak berbahasa Inggris atau campuran bahasa dalam satu teks untuk meningkatkan akurasi deteksi. Deteksi bahasa dengan alat yang lebih canggih dapat meningkatkan kualitas pra-pemrosesan data dan hasil akhir model, sebagaimana diusulkan oleh Sato et al. (2018).

Penelitian ini memberikan wawasan berharga tentang penggunaan model Transformer dan model deep learning lainnya dalam mendeteksi cyberbullying, serta pentingnya pemilihan model yang sesuai dengan karakteristik dataset untuk meningkatkan akurasi dan efisiensi deteksi. Penelitian lebih lanjut dengan mempertimbangkan variabilitas data dan optimalisasi parameter dapat menghasilkan model yang lebih andal dan efektif dalam aplikasi nyata.

SIMPULAN

Dalam Natural Language Processing (NLP) kualitas data teks sangat menentukan kualitas model dan prediksi. Dataset yang berkualitas rendah harus dilakukan preprocessing dengan hati-hati agar diperoleh teks yang berkualitas. Model yang sama baik transformer, deep learning, maupun machine learning akan memberikan hasil performa yang berbeda ketika diberikan dataset yang berbeda. Dalam percobaan terhadap lima dataset pada model deep learning diperoleh performa tertinggi pada dataset Youtube dengan model Transformer yaitu akurasi sebesar 98.49%. Sementara pada model machine learning juga didapat performa tertinggi pada dataset Youtube dengan model SVM dan menggunakan feature Tf-Idf yaitu akurasi sebesar 97.82%. Sebagai saran untuk pengembangan penelitian berikutnya adalah perlunya mempertimbangkan distribusi jumlah kata dalam dataset untuk menentukan parameter input dimensi pada model. Kemudian penelitian perlu menggunakan pendeteksi bahasa dalam hal ini bahasa Inggris agar dataset yang digunakan menjadi seragam berbahasa Inggris dan lebih berkualitas lagi datasetnya.

REFERENSI

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134. https://doi.org/https://doi.org/10.1016/j.knosys.2021.107134

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for abusive language detection in English. *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, 17–25. https://doi.org/10.18653/v1/2021.woah-1.3

Chugh, D., Anjum, A., & Katarya, R. (2021). *Automated news summarization using transformers*.

Dadvar, M., & Eckert, K. (2020). Cyberbullying detection in social networks using deep learning based models. In *Big Data Analytics and Knowledge Discovery: 22nd International Conference, DaWaK 2020, Bratislava, Slovakia, September 14–17, 2020, Proceedings 22* (pp. 245-255). Springer International Publishing.

Elsafoury, F. (2020). *Cyberbullying datasets*. Mendeley Data. https://doi.org/10.17632/jf4pzyvnpj.1

Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the Timeline meets the pipeline: A survey on automated cyberbullying detection. *IEEE Access*, *9*, 103541–103563. https://doi.org/10.1109/ACCESS.2021.3098979

Glazkova, A. (2020). A Comparison of synthetic oversampling methods for multi-class text classification. *CoRR*, *abs/2008.0*. https://arxiv.org/abs/2008.04636

Iwendi, C., Srivastava, G., Khan, S., & Maddikunta, P. K. R. (2020). Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems*. https://doi.org/10.1007/s00530-020-00701-5

Jabeen, F., & Treur, J. (2018). *Computational analysis of bullying behavior in the social media era BT - Computational Collective Intelligence* (N. T. Nguyen, E. Pimenidis, Z. Khan, & B. Trawiński (eds.); pp. 192–205). Springer International Publishing.

Kennedy, S., Walsh, N., Sloka, K., Foster, J., & Mccarren, A. (2020). *Fact or factitious? contextualized opinion spam detection*.

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, *83*, 105662. https://doi.org/https://doi.org/10.1016/j.asoc.2019.105662

Ma, E. (2019). *NLP Augmentation*. https://github.com/makcedward/nlpaug

Mehta, P., & Pandya, D. S. (2020). A Review on sentiment analysis methodologies, Practices And Applications. *International Journal of Scientific \& Technology Research*, *9*, 601–609.

Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC Model. *IEEE Access*, *9*, 78621–78634. https://doi.org/10.1109/ACCESS.2021.3083638

Shaikh, A. R., Alhoori, H., & Sun, M. (2023). YouTube and science: models for research impact. *Scientometrics*, *128*(2), 933-955.

Sato, M., Orihara, R., Sei, Y., Tahara, Y., & Ohsuga, A. (2018). *Text classification and transfer learning based on character-level deep convolutional neural networks BT - agents and artificial intelligence* (J. van den Herik, A. P. Rocha, & J. Filipe (eds.); pp. 62–81). Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, *abs/1901.1*, 6382–6388. https://doi.org/10.18653/v1/d19-1670

Zhong, H., Miller, D. J., & Squicciarini, A. (2019). Flexible inference for cyberbully incident

detection. In U. Brefeld, A. Marascu, F. Pinelli, E. Curry, B. MacNamee, N. Hurley, E. Daly, & M. Berlingerio (Eds.), *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2018* (pp. 356-371 BT-Machine Learning and Knowledge Disco). Springer Verlag. https://doi.org/10.1007/978-3-030-10997-4_22