Identifikasi Komentar Toksik Dengan BERT

Febrian Adhi Pratama[1], Ade Romadhony[2]

[1,2]Fakultas Informatika, Universitas Telkom, Bandung
[1]apbrian@students.telkomuniversity.ac.id, [2]aderomadhony@telkomuniversity.ac.id

Abstrak

Dalam penelitian ini, penulis menggunakan metode BERT untuk mengidentifikasi komentar toksik. Penulis membandingkan model yang dihasilkan dari dataset *User Generated Content*(UGC) dan *dataset* UGC yang telah dinormalisasikan. Arsitektur BERT yang telah dipre-train ditambah dengan *output layer* untuk klasifikasi teks. Hasil akhir dari penelitian tugas akhir ini menyatakan bahwa menormalisasikan dataset UGC terlebih dahulu tidak diperlukan d alam pelatihan model BERT.

Kata kunci: klasifikasi teks, deep learning, bert, transformers, attention, klasifikasi

Abstract

In this research paper, the author used BERT method to identify toxic comments. In this paper, the author compares BERT models which are trained with a raw User Generated Content(UGC) dataset and a normalized UGC dataset. The pre-trained BERT model is added with an output layer to classify the texts. The end result from this research paper revealed that UGC dataset normalization is not required in BERT model training.

Keywords: text classification, deep learning, bert, transformers, attention, classification

1. Pendahuluan

Latar Belakang

Dalam percakapan antara pengguna internet, terdapat banyak komentar yang berisi makian dan berbagai komentar lain yang tidak layak. [1] menunjukkan bahwa telah banyak remaja yang mengalami *cyber-bullying*. Untuk mengurangi *cyber-bullying*, dibutuhkan sistem yang dapat mengidentifikasi komentar toksik.

Beberapa penelitian telah mendefinisikan komentar toksik. [2] mendefinisikan komentar toksik sebagai komentar tidak sopan dan mengganggu pengguna lain yang mempersulit pengguna lain dalam mengekspresikan diri dan berinteraksi, sehingga membuat pengalaman yang buruk. [3] mengartikan komentar toksik sebagai komentar mengganggu yang berisi makian, secara sengaja mempermalukan, melecehkan secara seksual, mengancam, dan mengganggu secara terus-menerus.

Dataset komentar toksik diambil dari [4] yang mengklasifikasikan komentar toksik menjadi enam label, yaitu *toxic*, *severe toxic*, *obscene*, *threat*, *insult*, dan *identity hate*. Setiap komentar dapat memiliki lebih dari satu label. Jigsaw tidak memberikan definisi masing-masing label, sehingga penulis mendefinisikan masing-masing label berdasarkan kamus bahasa Inggris Cambridge beserta contoh dari dataset *Toxic Comment Classification Challenge* sebagai berikut.

1. *Toxic*

   Label *toxic* mendefinisikan komentar yang tidak menyenangkan. [5]

   Contoh: "*In the meantime someone needs to tell this "admin" to calm the fuck down and stop removing comments mad by editors on their own talk pages. That's a blockable offence, especially when they were told to never post here again. INVOLVED much?*"

2. *Severe toxic*

   Severe berarti menyebabkan sakit luar biasa, kesulitan, dan kerusakan. Penulis mengartikan *toxic comment* sebagai komentar yang menyebabkan rasa sakit hati luar biasa. [6]

   Contoh: "*Paul Tibbit is a fucking-ass little piece of smelly donkey shit that raped SpongeBob rock-hard! I hope Paul Tibbit gets fucking cancer and burns in hell!*"

3. *Obscene*

   Obscene berarti menghina, tidak sopan, atau menjijikkan berdasarkan standar moral yang dapat diterima. [7]

   Contoh: "*With such a heinous murder and rape spree, why isn't there a picture of the thug? His spiteful, bastard face must be seen to have the full effect. Also, it should be mentioned somewhere,*

maybe in a Controversies section, that the Oakland black community and black activist groups actually came to Mixon's defense- calling him a soldier, a hero, and a victim."

4. *Threat*

   Threat adalah usulan bahwa sesuatu yang tidak menyenangkan atau kekerasan akan terjadi, terutama jika suatu aksi atau perintah tertentu tidak dilakukan atau diikuti. [8]

   Contoh: "*Please stop. If you continue to ignore our policies by introducing inappropriate pages to Wikipedia, you will be blocked.*"

5. *Insult*

   Insult adalah mengatakan hal yang menghina atau tidak sopan kepada seseorang. [9]

   Contoh: "*Sorry mate, but your messages on James May's talk page prove that you are a sad little prat.*"

6. *Identity hate*

   Hate berarti ketidaksukaan yang sangat besar. [10] *Identity hate* adalah kebencian yang tertuju kepada identitas seseorang.

   Contoh: "*How does it feel to be a negros? Do you find it hard on yourself because your race genetically has average 85 IQ?*"

Salah satu metode klasifikasi adalah BERT. BERT adalah arsitektur *transformer* yang telah dilatih dengan *BookCorpus*(800 juta kata) dan Wikipedia berbahasa Inggris(2.500 juta kata). Dalam [11], terbukti bahwa BERT dapat mencapai performa yang tinggi hanya dengan *fine-tuning* dan *dataset training* yang kecil. Namun, masih belum terbukti pengaruh normalisasi pada *dataset User Generated Content*(UGC) terhadap model BERT. [12]

Dalam penelitian tugas akhir ini, penulis akan melakukan dan mengevaluasi klasifikasi teks toksik terhadap *dataset* komentar toksik dengan menggunakan metode BERT. *Dataset* akan dibagi menjadi dua, yaitu yang langsung diambil dari [4] dan yang telah dinormalisasi, kemudian dibandingkan hasilnya.

Topik dan Batasannya

Berdasarkan latar belakang yang telah dipaparkan, rumusan masalah dari tugas akhir ini adalah bagaimana membangun *classifier* untuk mengidentifikasi komentar toksik yang diambil langsung dari [4] dan yang telah dinormalisasi secara manual, dan bagaimana menganalisis performa dari sistem yang dibangun.

Untuk memastikan tugas akhir terarah, terdapat batasan masalah tugas akhir ini yang meliputi dataset yang diambil langsung dari *Toxic Comment Classification Challenge* dan yang telah dinormalisasi secara manual. *Dataset* memiliki enam label, yaitu *toxic, severe_toxic, obscene, threat, insult,* dan *identity_hate.*

Tujuan

Tujuan penulisan tugas akhir ini adalah membangun *classifier* untuk mengklasifikasi komentar toksik yang diambil secara langsung dan yang telah dinormalisasi secara manual dan menganalisis dan membandingkan performa dari sistem yang dibangun.

Organisasi Tulisan

Organisasi penulisan dimulai dengan pendahuluan yang berisi latar belakang, topik, batasan, dan tujuan penelitian. Selanjutnya adalah studi terkait yang berisi penelitian-penelitian *toxic comment identification* yang telah dilakukan dan metode BERT yang akan digunakan. Dan pada bagian akhir adalah evaluasi dan kesimpulan dari pcnelitian.

2. Studi Terkait

2.1 Toxic Comment Identification

Beberapa penelitian telah dilakukan untuk mengidentifikasi komentar toksik. Di tahun 2018, [13] menyatakan bahwa LSTM memberikan performa yang lebih baik dibandingkan dengan CNN baik dalam hal akurasi maupun waktu eksekusi. Di *epoch* ke-tujuh, CNN menghasilkan akurasi 98,04% dan waktu eksekusi 235 detik, sedangkan LSTM menghasilkan akurasi 98,77% dan waktu eksekusi 198 detik. Hal ini dikarenakan jumlah parameter CNN yang lebih banyak, sedangkan LSTM menggunakan jumlah parameter yang lebih sedikit namun memproses data secara *sequential*.

Dalam [14], dengan menggunakan metode *Random Forest*, *Logistic Regression*, dan *Gradient Boosting*, didapatkan *CV Score* tertinggi bernilai 97,91% dengan metode *Logistic Regression*.

2.2 Text Classification

Masih belum ada penelitian yang munggunakan metode BERT dengan dataset *Toxic Classification Challenge*. Namun, beberapa penelitian klasifikasi teks telah dilakukan menggunakan metode BERT. Guoqin Ma di [15] melakukan penelitian dengan dataset dari CrisisNLP dan CrisisLex. BERT dan BERT+LSTM dalam [15] memperoleh skor macro-f1 terbaik dengan nilai 64%.

Denis Gordeev dan Olga Lykova di [16] melakukan penelitian *hate speech detection* menggunakan BERT dengan dataset TRAC-2 yang terdiri dari tiga bahasa, yaitu bahasa Bengali, Inggris, dan India. Ada dua tugas dalam [16]. Tugas pertama adalah melakukan klasifikasi dengan label *hate speech*, *offensive*, dan tidak keduanya. Tugas kedua adalah melakukan klasifikasi dengan label *gendered* dan *non-gendered* untuk menentukan apakah teks *sexist* atau tidak. Model tugas kedua dengan bahasa Bengali dan Inggris memperoleh *F1-score* terbaik dengan skor 93% dan 87%.

Chi Sun, Xipeng Qiu , Yige Xu dan Xuanjing Huang di [17] menunjukkan bahwa variasi BERT mengalahkan performa berbagai metode klasifikasi teks termasuk LSTM dan CNN dengan banyak dataset berbeda.

BERT juga sangat berpengaruh dengan baik sebagai ekstrasi fitur. Di [18], dilakukan penelitian *hate speech classification* dengan lima metode, lima representasi fitur, dan gabungan dari empat dataset. Metode XGBoost dengan representasi fitur BERT mendapatkan skor terbaik dengan *F1-score* 92% dan ROC-AUC 99%.

2.3 BERT

BERT(*Bidirectional Encoder Representations from Transformers*) adalah metode perkembangan dari *Transformer* yang dikeluarkan pada tahun 2019. Sesuai dengan namanya, BERT hanya melakukan *encode* dan menghasilkan sebuah model bahasa. Sehingga tidak seperti Transformer, BERT hanya memerlukan *encoder*. BERT memiliki kelebihan dibandingkan metode-metode lain yang di antaranya sebagai berikut.

1. BERT memanfaatkan *encoder*, prinsip *attention*, dan membaca teks secara keseluruhan sebagai input, bukan urutan sekuensial. Dengan karakteristik BERT ini, BERT dapat mengerti hubungan kontekstual setiap token dengan baik. Cara BERT membaca *input* dengan *positional encoding* juga membuat BERT dapat membaca teks yang panjang lebih baik dari RNN. [11] [19]
2. BERT memanfaatkan arsitektur *Transformer* dan menggunakan lebih sedikit *parameter* dari model CNN, sehingga mampu mencapai performa yang lebih tinggi dengan waktu yang lebih singkat. [20]
3. Tidak seperti CNN, BERT baik digunakan untuk dataset berukuran kecil karena telah di*pre-train* sehingga hanya memerlukan fine-tuning. [11]
4. Dibandingkan dengan RNN yang memiliki kompleksitas O(n) dan CNN yang memiliki kompleksitas waktu $O(\log_k(n))$, *self-attention* pada BERT hanya memiliki kompleksitas waktu O(1).
5. Tidak seperti OpenAI GPT [21] dan ELMo [22], setiap *encoder* di arsitektur BERT memiliki akses ke setiap token sehingga model dapat mengerti arti token dari dua arah secara bersamaan.

Berikut blok-blok pembangun *BERT*.

1. *Input*

   BERT menggunakan WordPiece embedding yang mengandung 30.000 suku kata. [23] Di awal dari setiap sequence terdapat token [CLS]. Untuk membedakan kalimat di setiap sequence, di setiap akhir kalimat ditempatkan token [SEP]. Kemudian ditambahkan embedding di setiap token untuk membedakan apakah token termasuk di kalimat yang mana yang disebut *segment embedding*. Hasil dari penjumlahan ini akhirnya akan ditambahkan dengan *position embedding* yang memiliki jumlah dimensi sama dengan token *embedding*. *Position embedding* dihitung dengan rumus sebagai berikut.

$$\vec{p}_{2k} = sin\left(\frac{1}{10000^{2k/d}} \cdot t\right) \quad (1)$$

$$\vec{p}_{2k+1} = cos\left(\frac{1}{10000^{2k/d}} \cdot t\right) \quad (2)$$

   Setiap embedding memiliki ukuran dimensi 768.

Gambar 1. Representasi *input*

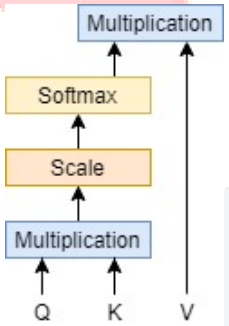
2. *Encoder*

Dalam arsitektur yang akan digunakan, akan digunakan dua belas layer *encoder*. *Encoder* tersusun dari beberapa bagian di antaranya sebagai berikut. [24]

a. *Self-attention*

Self-attention dihitung dengan rumus *scaled dot product attention* sebagai berikut.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d^k}}\right)V \quad (3)$$



Gambar 2. Diagram persamaan *scaled dot product*

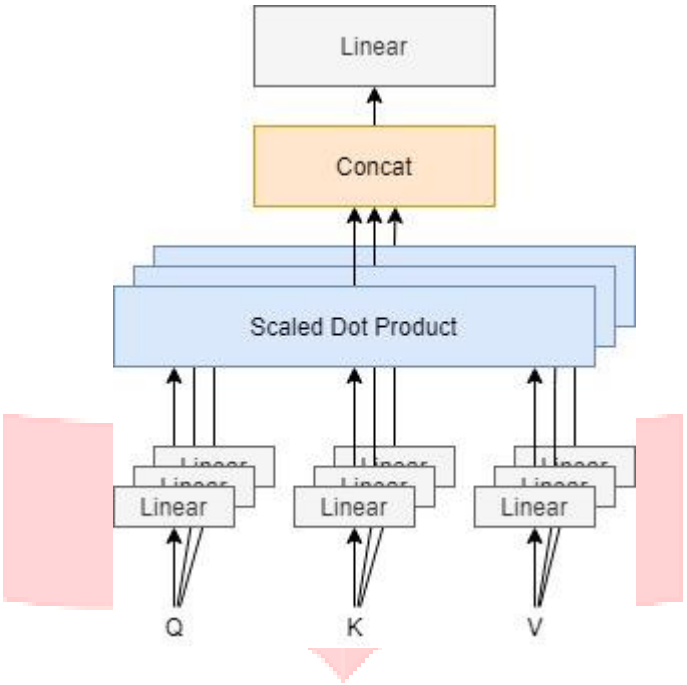
b. *Multihead-attention*

Multihead-attention adalah tumpukan *self-attention* yang memiliki rumus sebagai berikut.

$$Multihead(Q,K,V) = Concat(head_1,\dots,head_h)W^o \quad (4)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (5)$$

Jumlah *head* yang akan digunakan adalah 12.



Gambar 3. Diagram *multihead-attention*

Model BERT telah di*pre-train* dengan dua tugas berikut sehingga dalam penggunaan BERT selanjutnya hanya diperlukan untuk melakukan fine-tuning.

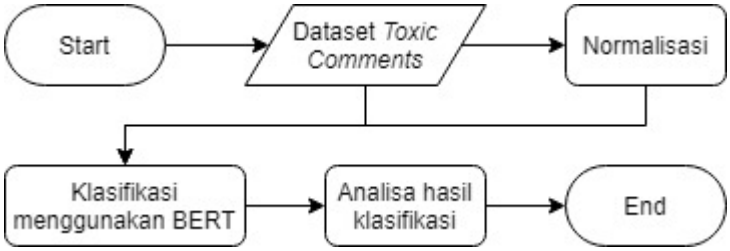
1. *Masked Language Modelling*
   Secara acak, 15% dari semua token input digantikan dengan token [MASK] untuk melatih representasi dua arah secara mendalam. Karena pada saat *fine-tuning* tidak akan ada token [MASK], token yang telah terpilih secara acak akan diganti dengan token [MASK] dengan probabilitas 80%, token acak dengan probabilitas 10%, atau tidak diganti dengan probabilitas 10%. Kemudian, token yang asli akan diprediksi dengan menggunakan *cross entropy loss.*

2. *Next Sentence Prediction* (NSP)
   Dalam tugas seperti Question answering, model perlu mengerti hubungan dari setiap kalimat. Untuk ini, dilakukan pre-training dengan NSP. Dengan 50% probabilitas akan kalimat kedua adalah terusan dari kalimat pertama dengan label IsNext, dan NotNext untuk sebaliknya.

BERT dipre-train menggunakan data BookCorpus sebanyak 800 juta kata dan Wikipedia berbahasa Inggris sebanyak 2,5 miliar kata. Dengan d ata pre-training yang banyak dan tugas pre-training yang membuat model BERT mengerti setiap kata dengan pemahaman yang dalam, hanya diperlukan fine-tuning untuk menggunakan BERT dalam berbagai tugas.

3. Sistem yang Dibangun

Training dan testing akan dilakukan untuk masing-masing dataset UGC yang telah dinormalisasikan dan yang tidak. Berikut adalah diagram alir dari sistem yang akan dibangun.



Gambar 4. Diagram alir sistem

3.1 Dataset

Dataset yang digunakan berasal dari komentar Wikipedia yang disusun oleh Jigsaw dan tersedia di Kaggle, yang berisi 159.571 baris data. [4] Setiap baris dari dataset memiliki satu komentar dan dapat memiliki lebih dari satu label. Dataset yang akan digunakan dibagi menjadi dua, yaitu yang diambil secara langsung dan yang telah dinormalisasikan. Berikut jumlah baris berdasarkan label pada dataset UGC mentah.

Label	Jumlah Baris
Toxic	15294
Severe toxic	1595
Obscene	8449
Threat	478
Insult	7877
Identity hate	1405

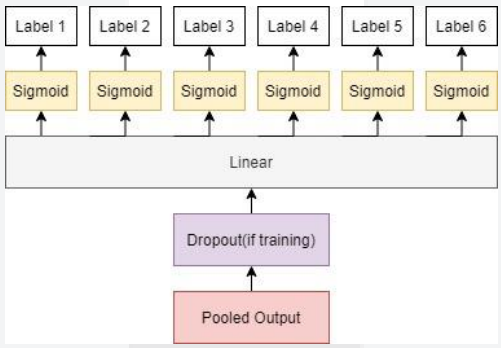
Tabel 1. Jumlah komentar dataset berdasarkan label

3.2 Normalisasi

Normalisasi dilakukan dengan membuat setiap huruf menjadi *lower case*, menghilangkan karakter spesial, menormalisasikan setiap komentar dengan menstandarkan setiap kata-kata yang tidak standar, dan menghilangkan white space. Daftar kata-kata tidak standar didapatkan dari [25]. Kemudian, baris dengan komentar yang kosong dibuang dari dataset.

3.3 Klasifikasi

Model yang akan digunakan adalah BERT Base yang memiliki 12 *hidden layer* berukuran 768 dan 12 *attention head*. Kemudian output layer berikut akan dilakukan untuk melakukan klasifikasi.



Gambar 5. Diagram output layer

Training dan validasi akan dilakukan untuk masing-masing dataset UGC yang telah dinormalisasikan dan yang tidak untuk melakukan fine-tuning. Kemudian, model akan dianalisa dengan data *testing* dan dibandingkan.

4. Evaluasi

4.1 Skenario Pengujian

Berikut adalah dua skenario yang digunakan dalam penelitian tugas akhir ini.

1. Pelatihan model menggunakan dataset UGC yang mentah

   Skenario pertama dibuat untuk dianalisa dan dibandingkan dengan hasil dari skenario lain.
2. Pelatihan model menggunakan dataset UGC yang telah dinormalisasikan tanpa penghapusan *stop words* dan tanpa lematisasi

   BERT telah di-*pretrain* menggunakan dataset BookCorpus dan Wikipedia berbahasa Inggris yang besar, penulis menduga bahwa tidak diperlukan penghapusan *stop word* dan lematisasi.
3. Pelatihan model menggunakan dataset UGC yang telah dinormalisasikan dengan penghapusan *stop words* dan tanpa lematisasi

BERT telah di-pretrain dengan dataset BookCorpus dan Wikipedia berbahasa Inggris yang besar tanpa penghapusan *stop word*, sehingga penulis menduga bahwa penghapusan *stop word* tidak diperlukan, dan bahkan dapat mengurangi performa model BERT. Untuk membuktikan hipotesis ini, maka penulis melakukan pelatihan model menggunakan dataset UGC yang telah dinormalisasikan dengan penghapusan *stop words* dan tanpa lematisasi.

4. Pelatihan model menggunakan dataset UGC yang telah dinormalisasikan tanpa penghapusan *stop word* dan lematisasi

   BERT telah di-pretrain dengan dataset BookCorpus dan Wikipedia berbahasa Inggris yang besar tanpa lematisasi, dan telah menggunakan WordPiece *embedding*, sehingga penulis menduga bahwa lematisasi tidak akan efektif dan bahkan dapat mengurangi performa BERT. Untuk membuktikan hipotesis penulis, maka dilakukan skenario pelatihan model menggunakan dataset UGC yang telah dinormalisasikan tanpa penghapusan *stop word* dan lematisasi.

Semua model kemudian dianalisa dan dibandingkan antara performa model dengan dataset UGC mentah dan dataset UGC dinormalisasi. Evaluasi dilakukan berdasarkan skor *precision, recall* dan *f1=score.* [26]

4.1 Analisis dan Hasil Pengujian

Berikut adalah hasil rata-rata dari data *testing* yang didapatkan dari lima kali pelatihan model untuk masing-masing skenario.

	Toxic		*Severe Toxic*		*Obscene*		*Threat*		*Insult*		*Identity hate*	
Skenario 1	14.075	340	15.746	62	14.921	173	15.893	15	14.934	208	15.759	55
	227	1.316	80	70	129	735	30	20	175	642	64	80
Skenario 2	14.084	14.084	15.744	58	14.907	181	15.883	19	14.907	228	15.756	52
	14.084	14.084	83	67	122	742	27	23	166	651	66	78
Skenario 3	14.032	335	15.694	64	14.866	179	15.844	15	14.852	240	15.716	48
	269	1.272	79	71	126	737	36	13	197	620	71	73
Skenario 4	14.078	331	15.744	58	14.903	186	15.884	18	14.907	228	15.757	51
	233	1.311	83	68	120	744	28	22	163	654	67	77

Tabel 2. *Confusion matrix* hasil pengujian

	Precision	***Recall***	***F1-score***
Toxic	79,5576%	85,3014%	82,3260%
Severe toxic	52,6888%	46,0000%	49,1090%
Obscene	80,9662%	84,7454%	82,8073%
Threat	60,1817%	42,4000%	49,5828%
Insult	75,8078%	78,3599%	77,0575%
Identity hate	58,8636%	55,0000%	56,8202%
Rata-rata	68,0110%	65,3011%	66,2838%

Tabel 3. Hasil analisa model dengan dataset UGC mentah

	Precision	***Recall***	***F1-score***
Toxic	80,0866%	84,9514%	82,4455%
Severe toxic	53,1555%	44,4000%	48,3766%
Obscene	80,2308%	85,7407%	82,8893%
Threat	55,4819%	46,0000%	50,2222%
Insult	74,1410%	79,7797%	76,8527%
Identity hate	59,8574%	53,7500%	56,6162%
Rata-rata	67,1588%	65,7703%	66,2337%

Tabel 4. Hasil analisa model dengan dataset UGC yang dinormalisasikan tanpa penghapusan *stop word* dan lematisasi

	Precision	***Recall***	***F1-score***

Toxic	79,3481%	82,4140%	80,8471%
Severe toxic	52,2249%	47,0667%	49,4992%
Obscene	80,5769%	85,2375%	82,8394%
Threat	47,3546%	27,7551%	34,9223%
Insult	72,2397%	75,8824%	74,0123%
Identity hate	60,2930%	50,9722%	55,2111%
Rata-rata	65,3395%	61,5547%	62,8886%

Tabel 5. Hasil analisa model dengan dataset UGC yang dinormalisasikan dengan penghapusan *stop word* dan tanpa lematisasi

	Precision	***Recall***	***F1-score***
Toxic	79,7273%	85,2236%	82,3835%
Severe toxic	53,3594%	45,3333%	49,0115%
Obscene	79,6687%	85,7639%	82,5980%
Threat	58,1021%	44,8000%	50,3722%
Insult	74,7433%	78,4578%	76,5550%
Identity hate	59,7043%	53,4722%	56,4133%
Rata-rata	67,5508%	65,5085%	66,2222%

Tabel 6. Hasil analisa model dengan dataset UGC yang dinormalisasikan dengan lematisasi dan tanpa penghapusan *stop word*

Dari hasil pengujian, dapat terlihat bahwa normalisasi dataset UGC dapat mengurangi skor *precision* dan meningkatkan skor *recall*, namun mengurangi *F1-score*. Peningkatan paling signifikan dapat terlihat pada label *threat* dengan peningkatan *F1-score* dan skor *recall* tertinggi, dan *identity hate* dengan skor *precision* tertinggi. Penulis menyimpulkan bahwa semakin kecil ukuran dataset, maka akan semakin tinggi skor *recall*, *precision*, dan *F1-Score* yang didapat. Namun, dari hasil evaluasi terlihat bahwa normalisasi hanya mengurangi performa model BERT terhadap dataset UGC secara keseluruhan. Hal ini disebabkan karena BERT telah di-*pretrain* dengan *corpus* yang sangat besar.

Pada tabel performa data UGC yang dinormalisasikan dengan penghapusan *stop word*, dapat terlihat bahwa terjadi penurunan performa secara keseluruhan. Lematisasi juga mengurangi performa model secara keseluruhan. Dengan metode BERT, lematisasi tidak diperlukan karena tokenisasi telah dilakukan dengan menggunakan *wordpiece*.

5. Kesimpulan

Berdasarkan penjelasan hasil pengujian, dapat disimpulkan bahwa normalisasi dataset UGC tidak diperlukan sebelum melakukan pelatihan model BERT. Model BERT baik digunakan untuk dataset UGC berukuran kecil maupun besar. Model BERT juga baik digunakan untuk dataset yang bersumber dari Wikipedia berbahasa Inggris seperti dataset [4].

Daftar Pustaka

[1] O'Dea, B. & Campbell, A., 2012. Online Social Networking and the Experience of Cyber-Bullying. *Studies in health technology and informatics*, Volume 181, p. 213.

[2] Almerekhi, H., Jansen, B. J., Kwak, H. & Salminen, J., 2019. Detecting Toxicity Triggers in Online Discussions. Dalam: HT '19: Proceedings of the 30th ACM Conference on Hypertext and Social Media. New York: Association for Computing Machinery, pp. 291-292.Ex Machina: Personal Attacks Seen at Scale,

[3] Ex Machina: Personal Attacks Seen at Scale,

[4] Jigsaw, 2018. *Toxic Comment Classification Challenge.* [Online]
Available at: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
[Diakses 5 February 2020].

[5] Cambridge University Press, t.thn. *Toxic.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/toxic
[Diakses 24 Juli 2020].

[6] Cambridge University Press, t.thn. *Severe.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/severe
[Diakses 24 July 2020].
[7] Cambridge University Press, t.thn. *Obscene.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/obscene
[Diakses 24 July 2020].
[8] Cambridge University Press, t.thn. *Threat.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/threat
[Diakses 24 July 2020].
[9] Cambridge University Press, t.thn. *Insult.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/insult
[Diakses 24 July 2020].
[10] Cambridge University Press, t.thn. *Hate.* [Online]
Available at: https://dictionary.cambridge.org/dictionary/english/hate
[Diakses 24 July 2020].
[11] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Dalam: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Minneapolis: Association for Computational Linguistics, p. 4171–4186.
[12] Muller, B., Sagot, B. & Seddah, D., 2019. Enhancing BERT for Lexical Normalization. Dalam: *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019).* Hongkong: Association for Computational Linguistics, pp. 297-306.
[13] Sharma, R. & Patel, M., 2018. Toxic Comment Classification Using Neural Networks and Machine Learning. *International Advanced Research Journal in Science*, 5(9), p. 51.
[14] Ravi, P., Batta, H. N., S, G. & Yaseen, S., 2019. Toxic Comment Classification. *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 3(4), p. 27.
[15] Ma, G., 2019. Tweets Classification with BERT in the Field of Disaster Management.
[16] Gordeev, D. & Lykova, O., 2020. BERT of all trades, master of some. Dalam: *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying.* Marseille: European Language Resources Association (ELRA), pp. 93-98.
[17] Sun, C., Qiu, X. & Huang, X., 2020. How to Fine-Tune BERT for Text Classification?. *arXiv preprint arXiv:1905.05583*, Issue 3.
[18] Salminen, J. et al., 2020. Developing an online hate classifier for multiple social media platforms. Dalam: *Human-centric Computing and Information Sciences.* s.l.:s.n., p. 10.
[19] Sherstinsky, A., 2020. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, Issue 404.
[20] Khan, A., Sohail, A., Zahoora, U. & Qureshi , A. S., 2020. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review.*
[21] Radford, A. et al., 2018. Language Models are Unsupervised Multitask Learners.
[22] Peters, M. et al., 2018. Deep Contextualized Word Representations. Dalam: *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies.* New Orleans: Association for Computational Linguistics, pp. 2227-2237.
[23] Wu, Y. et al., 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144.*
[24] Vaswani, A. et al., 2017. Attention is All you Need. Dalam: *Advances in Neural Information Processing Systems.* s.l.:Curran Associates, Inc., pp. 5998-6008.
[25] Zafar, 2018. *Toxic Data Preprocessing.* [Online]
Available at: https://www.kaggle.com/fizzbuzz/toxic-data-preprocessing
[26] Powers, D. & A., 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1), pp. 38, 41.