Identifikasi *Cyberbullying* pada Media Sosial *Twitter* Menggunakan Metode LSTM dan BiLSTM

Habib Faizal Fadli
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14.5, Sleman, Yogyakarta, Indonesia
17523143@students.uii.ac.id

Ahmad Fathan Hidayatullah
Program Studi Sarjana Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14.5, Sleman, Yogyakarta, Indonesia
Fathan@uii.ac.id

***Abstract*—*Cyberbullying* merupakan masalah yang harus menjadi perhatian penting bagi masyarakat. *Cyberbullying* termasuk kebiasaan buruk yang berdampak mengerikan, mulai dari gangguan psikologis korban, hingga munculnya kasus bunuh diri. Tujuan dari penelitian ini adalah mengidentifikasi konten yang mengandung makna perundungan secara daring (*cyberbullying*) pada media sosial khususnya *Twitter*. Dalam kasus ini, sumber data penelitian ini berasal dari media sosial *Twitter*. Setidaknya ada 6835 data yang telah dikumpulkan. Data tersebut terdiri dari dua jenis cuitan dengan masing-masing cuitan memiliki kecenderungan *cyberbullying* dan *non-cyberbullying*. Tujuan penelitian akan tercapai dengan melakukan beberapa langkah yaitu pengumpulan data, *preprocessing*, klasifikasi, evaluasi, dan diakhiri dengan deteksi konten. Dua algoritma *deep learning* diimplementasikan dalam penelitian ini, yaitu LSTM dan BiLSTM. Kesimpulan yang dapat diambil yaitu kedua algoritma tersebut memiliki performa yang relatif sama. Akurasi dari masing masing algoritma dituliskan sebagai berikut *Long Short Term Memory* 93.77% dan *Bidirectional Long Short Term Memory* 95.24%. Lalu, untuk nilai dari F1-Score dari masing-masing algoritma sebagai berikut *Long Short Term Memory* 92.02% dan *Bidirectional Long Short Term Memory* 93.84%.**

Keywords*—*Twitter, Cyberbullying, Deep Learning, Klasifikasi

I. PENDAHULUAN

Media sosial merupakan tempat dimana penggunanya melakukan interaksi dengan pengguna lain secara daring tanpa mengenal waktu dan tempat. Indonesia sendiri merupakan negara dengan jumlah pengguna media sosial tertinggi di dunia. Berdasarkan Kementrian Komunikasi dan Informatikan (Kemenkominfo) menyatakan 95% dari sekitar 63 juta pengguna internet adalah pengguna media sosial [1]. *Twitter* merupakan aplikasi yang sering digunakan oleh masyarakat Indonesia. *Country Industry Head Twitter* Indonesia mengklaim bahwa Indonesia termasuk negara dengan pertumbuhan pengguna aktif harian *Twitter*-nya paling besar [2].

Penggunaan *Twitter* di Indonesia tidak hanya ditujukan untuk perorangan. Akan tetapi, *Twitter* juga digunakan oleh lembaga-lembaga negara, komunitas, hingga toko *online*. Namun, tidak semua pengguna menggunakan teknologi ini dengan bijak. Tindakan negatif seperti penipuan, penyebaran *hoax*, menyebarkan opini yang cenderung mengandung ujaran kebencian, hingga perundungan secara daring (*cyberbullying*) banyak dilakukan oleh pengguna *Twitter* [3]. Selama tahun 2019, masyarakat telah melaporkan setidaknya 244.738 jumlah konten pornografi. Terdapat juga sejumlah 19.970 konten kategori perjudian. Ada pula konten penipuan sejumlah 18.845, dan konten informasi hoaks, serta konten mengandung SARA, terorisme, radikalisme, pelanggaran HAK, serta kekerasan terhadap anak sejumlah 15.361 [4]. Hal ini tentu menjadi dampak negatif penggunaan media sosial. Dampak-dampak yang dapat ditimbulkan seperti kerusuhan karena menyebarnya berita bohong, kerugian materiil karena penipuan, hingga berbagai kasus yang ditimbulkan karena *cyberbullying* [5].

Cyberbullying atau perundungan secara daring adalah tindakan penyerangan, penghinaan, atau menyakiti orang lain secara sengaja dan berulang-ulang pada sosial media, pesan, atau dengan cara lainnya [5]. *Cyberbullying* di media sosial *Twitter* dilakukan dengan menulis cuitan yang mengandung kata-kata hinaan atau kata-kata kasar, bahkan kata-kata yang menjurus kepada penghinaan terhadap SARA. Pemerintah mengumumkan setidaknya 84% remaja berusia 12 sampai 17 tahun di Indonesia menjadi korban tindakan perundungan (*bullying*) dan kebanyakan kasus *bullying* yang ditemukan merupakan *cyberbullying*[6]. *Cyberbullying* menjadi kekhawatiran publik karena banyak kasus *cyberbullying* sering dikaitkan dengan tindakan bunuh diri. Salah satu organisasi non-profit, *Cyber Bullying Research Center* mengungkapkan bahwa kebiasaan *bullying* baik secara langsung maupun secara daring di kalangan remaja dapat mengakibatkan depresi, tindakan bunuh diri dan percobaan pembunuhan [5]. Oleh karena efek berbahaya yang ditimbulkan *cyberbullying*, tindakan pencegahan atau deteksi perlu dilakukan agar tidak membahayakan korban maupun pelaku.

Penelitian dilakukan dengan melakukan lima tahapan proses. Langkah pertama yaitu pengumpulan data, lalu *preprocessing*, *kemudian* klasifikasi, setelah itu evaluasi, dan terakhir deteksi konten. Metode yang digunakan yaitu *deep learning* dan algoritma yang digunakan yaitu LSTM dan BiLSTM. Penelitian ini penting dilakukan karena belum ditemukan model yang dibuat khusus untuk mendeteksi dan mengklasifikasikan cuitan yang berbahasa indonesia dan bermakna *cyberbullying* dengan metode *deep learning* terutama dengan 2 algortima tersebut. Kedua algoritma tersebut dinilai sangat tepat dalam melakukan deteksi dan klasifikasi dibandingkan algoritma-algortima yang lainnya.

Penelitian ini mengklasifikasikan data untuk mendeteksi *cyberbullying* di media sosial Twitter menjadi dua kelas yaitu *cyberbulying* dan bukan *cyberbullying*. Kelas *cyberbullying* berisi cuitan berupa kata kata yang mengandung unsur *cyberbullying*, sementara kelas bukan *cyberbullying* berisi cuitan berupa kata kata yang tidak mengandung unsur *cyberbullying*. Dengan dikembangkannya penelitian ini, diharapkan akan dapat membantu para orangtua, pemerintah, dan negara untuk melindungi generasi muda dari perundungan secara daring (*cyberbullying*) dan menekan jumlah para pelaku *cyberbullying*.

II. TINJAUAN PUSTAKA

Penelitian mengenai klasifikasi teks sudah banyak dilakukan oleh peneliti-peneliti sebelumnya. Konten kasar yang diklasifikasikan berdasarkan cuitan media sosial *Twitter* dengan kamus bahasa Indonesia juga pernah dilakukan sebelumnya. Hidayatullah, et. al [7] melakukan klasifikasi berdasarkan cuitan menjadi dua kelas dan melakukan perbandingan terhadap performa algoritma NBC dan SVM dalam melakukan klasifikasi. Berikut adalah hasil *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dari masing-masing model yang digunakan dimana NBC mempunyai nilai 0.9834; 0.9912; 0.9762; 0.9836 dan SVM mempunyai nilai 0.9928; 0.9914; 0.9946; 0.9930.

Penelitian tentang deteksi *cyberbullying* pada media sosial *Twitter* berbahasa Indonesia pernah dilakukan sebelumnya. Abdullah dan Hidayatullah [8] melakukan deteksi cuitan pada media sosial *Twitter*. Data yang berupa cuitan tersebut nantinya akan dimasukkan ke dalam kelas *cyberbullying* dan kelas *non-cyberbullying*. Deteksi cuitan tersebut menggunakan metode *Machine* Learning dengan algoritma NBC, *SVM*, *Logistic Regression*, dan *KNN*. Berikut adalah hasil *Accuracy*, *Precision*, *Recall*, dan *F1-Score* dari masing-masing algoritma yang digunakan dimana algoritma NBC mempunyai nilai 0.961; 0.96; 0.96; 0.96, SVM mempunyai nilai 0.994; 0.99; 0.99; 0.99, *Logistic Regression* mempunyai nilai 0.997; 1.00; 1.00; 1.00, dan KNN mempunyai nilai 0.918; 0.93; 0.92; 0.92.

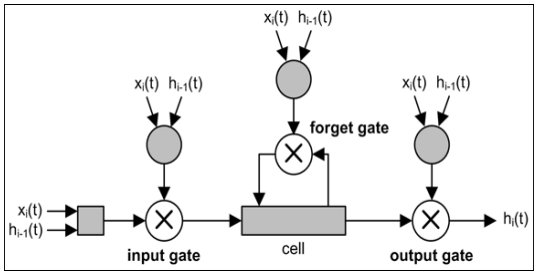
Penelitian tentang deteksi *cyberbullying* pada media sosial selain *Twitter* juga sudah pernah dilakukan sebelumnya. Hoseinmardi, et. al [9] melakukan penelitian tentang klasifikasi *cyberbullying* menggunakan algoritma NBC dan SVM pada media sosial Instagram berdasarkan foto dan komentar. Berikut adalah hasil *Accuracy*, *Precision*, dan *Recall* dari masing-masing algortima yang digunakan dimana algoritma SVM mempunyai nilai 0.74; 0.74; 0.78 dan SVM mempunyai nilai 0.87; 0.88; 0.87.

Ada pula penelitian tentang deteksi *cyberbullying* yang digabungkan dengan metode pendekatan psikologi yang pernah dilakukan sebelumnya. Balakhrishnan, et. al [10] melakukan penelitian ini menggunakan model *Big Five* and *Triad* dan algoritma *Random Forest*. Berikut adalah hasil *Precision*, *Recall*, dan *F-Measure* dari algoritma yang digunakan dimana algoritma *Random Forest* mempunyai nilai 0.960; 0.952; 0.929.

Penelitian tentang *bullying* yang mengambil data lebih dari satu media sosial juga pernah dilakukan sebelumnya. Agrawal dan Awekar [11] melakukan penelitian tentang *cyberbullying* di tiga media sosial yaitu *Formspring*, *Twitter*, dan *Wikipedia* dengan membagi 4 kelas *bullying* yaitu *bully*, *racism*, *sexism*, dan *attack* kemudian menguji data dengan mengklasifikasikannya dengan algortima CNN, LSTM, BiLSTM, dan BiLSTM *with attention*. Akan tetapi, pada penelitian tersebut masih menggunakan korpus bahasa inggris. Berikut adalah hasil *Precision*, *Recall*, dan *F1-Score* dari masing-masing algoritma yang digunakan dimana CNN mempunyai nilai 0.93; 0.90; 0.91, LSTM mempunyai nilai 0.91; 0.85; 0.88, BiLSTM mempunyai nilai 0.91; 0.81; 0.86, dan BiLSTM *with attention* mempunyai nilai 0.90; 0.91; 0.91.

III. *LONG SHORT TERM MEMORY*

LSTM adalah modifikasi dari *Recurrent Neural Network* (RNN) dengan adanya penambahan *memory cell* yang digunakan untuk menyimpan informasi dengan jangka waktu yang panjang, serta LSTM juga dapat menangani masalah *vanishing gradient* yang terdapat pada RNN saat memproses data sekuensial yang panjang dengan menggunakan satu set gerbang yang digunakan untuk mengontrol informasi yang masuk ke memori [12].

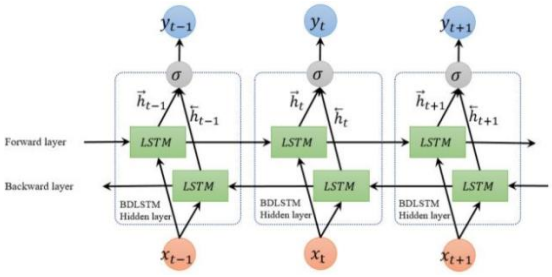


Gambar 1. Arsitektur LSTM

Gambar 1 menunjukkan arsitektur dari LSTM. *Cell state* merupakan tempat untuk menyimpan informasi yang diberikan dari satu langkah waktu ke langkah waktu berikutnya. *Gate units* berperan dalam memproses informasi yang dibutuhkan dan dibuang. *Gate units* terdiri dari *input gate*, *forget gate*, dan *output gate*. *Input gate* merupakan *gate* yang berfungsi untuk memutuskan nilai *input* yang akan diteruskan pada *cell state* untuk diperbaharui. *Forget gate* marupakan *gate* yang memutuskan informasi mana yang perlu dibuang dari *cell state*. *Output gate* merupakan *gate* yang memutuskan *output* yang akan dihasilkan. Pada Gambar 1 menunjukkan alur *gate units* dalam mengontrol konektivitas LSTM.

IV. *BIDIRECTIONAL LONG SHORT TERM MEMORY*

BiLSTM adalah perkembangan dari model LSTM dimana terdapat dua lapisan yang prosesnya saling berkebalikan arah, model ini sangat baik untuk mengenali pola dalam kalimat karena setiap kata dalam dokumen diproses secara sekuensial, karena cuitan dapat dipahami bila pembelajaran secara berurut setiap kata. Lapisan dibawahnya bergerak maju (*forward*), yaitu memahami dan memproses dari kata pertama menuju kata terakhir sedangkan lapisan diatasnya bergerak mundur (*backward*), yaitu memahami dan memproses dari kata terakhir menuju kata pertama. Dengan adanya lapisan dua arah yang saling berlawanan ini maka model dapat memahami dan mengambil perspektif dari kata terdahulu dan kata terdepan, sehingga proses pembelajaran akan semakin dalam yang berdampak pada model akan lebih memahami konteks pada cuitan tersebut.



Gambar 2. Arsitektur BiLSTM

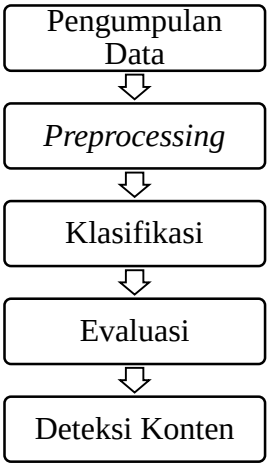
Gambar 2 menunjukkan arsitektur BiLSTM. Setiap *hidden unit* keluaran *unit* pada lapisan bawah dan atas di

digabungkan membentuk nilai fitur kata tersebut dengan ukuran lebih panjang daripada menggunakan LSTM biasa. Karena lebih panjang nilai fitur, maka informasi yang akan di proses pada tahap selanjutnya yaitu *feed forward neural* akan mengklasifikasikan dengan lebih akurat.

BiLSTM akan sangat bermanfaat dalam hal pelabelan sekuensial apabila memiliki akses terhadap kedua informasi dari sebelum dan sesudahnya. Namun *hidden state* pada LSTM hanya mengambil informasi dari sebelumnya (masa lalu), sedangkan untuk informasi yang ada setelahnya tidak diketahui. Permasalahan tersebut dapat dipecahkan dengan menggunakan BiLSTM [13]. BiLSTM pada dasarnya terdiri dari dua LSTM, *forward* LSTM dan *backward* LSTM. Gabungan *forward* LSTM dan *backward* LSTM tersebut akan menangkap informasi dari kedua arah.

V. METODOLOGI

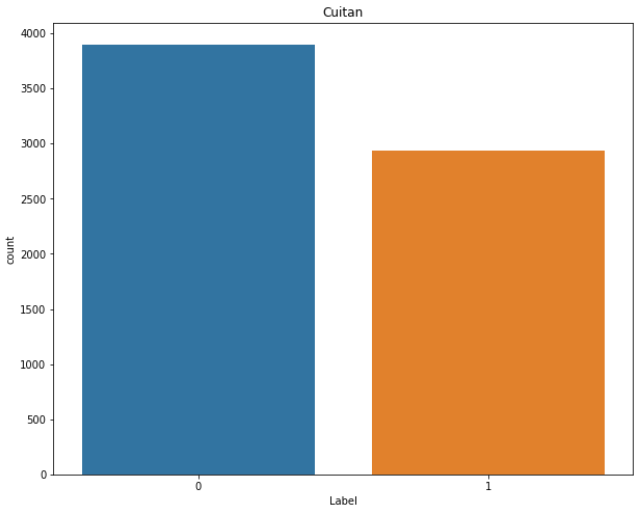
Bagian ini berisi pokok bahasan yang menjelaskan tentang langkah-langkah yang digunakan dalam penelitian ini. Pertama yaitu pengumpulan data, lalu *preprocessing*, kemudian klasifikasi, setelah itu evaluasi, dan terakhir deteksi konten. Gambar 3 memberikan ilustrasi tentang alur yang dilakukan pada penelitian ini.



Gambar 3. Alur Penelitian

A. *Pengumpulan Data*

Twitter adalah tempat pengambilan data untuk penelitian ini. Langkah pengumpulan data dilakukan dengan memanfaatkan Twitter API. Sebelum menggunakan *Twitter* API, penulis harus mendaftar mendaftarkan diri sebagai *developer* untuk mendapatkan izin akses berupa *Consumer Key*, *Consumer Secret*, *Access Token*, dan *Access Secret*. Data yang akan diambil dari *Twitter* berupa data cuitan berbahasa Indonesia. Cuitan yang terkumpul untuk membangun model pada penelitian ini adalah 6835 cuitan seperti yang terlihat pada Gambar 4 dengan cacah masing masing yaitu 3900 cuitan *non-cyberbullying* dan 2935 cuitan *cyberbullying*.



Gambar 4. Cacah Data

Cuitan *non-cyberbullying* diambil menggunakan kata kunci yang positif. Sedangkan cuitan *cyberbullying* diambil dengan menggunakan kata kunci yang merupakan kata-kata yang bermakna perundungan.

B. *Preprocessing*

Preprocessing merupakan langkah dimana membersihkan data mentah berupa cuitan-cuitan sehingga menjadi data yang baik dan terstruktur. Langkah *preprocessing* yang digunakan merujuk pada langkah-langkah *preprocessing* yang dilakukan Hidayatullah dan Ma'arif [14]. Berikut langkah–langkah *preprocessing* yang dilakukan antara lain:

- Menghilangkan *link* atau URL.
- Menghilangkan karakter NON-ASCII.
- Menghilangkan angka, simbol dan tanda baca.
- Menghilangkan *hashtag*, *username*, dan *retweet.*
- Merubah huruf ke dalam bentuk *lowercase.*
- Menghilangkan *stopwords.*
- Merubah Bahasa informal ke Bahasa formal.
- Menghilangkan kata yang terdiri dari satu huruf.
- Menghilangkan digit.

Tabel 1 merupakan contoh dari hasil *preprocessing* cuitan.

Tabel 1. *Preprocessing*

Sebelum	Sesudah
@pengguna 1 @pengguna 2 dihh itu artis sok sokan ngartis	itu artis sok ngartis
@pengguna 3 heloow yg sombong gaya gapernah ngerasain ditampol spokat	yang sombong gaya gapernah ngerasain ditampol spokat

C. Klasifikasi

Langkah ini akan menjelaskan tentang proses klasifikasi cuitan dengan menggunakan beberapa algoritma *deep learning* yaitu LSTM dan BiLSTM. Data yang berupa teks

diklasifikasikan ke dalam kelas-kelas *cyberbullying* menggunakan algoritma tersebut karena algoritma tersebut terbukti karena memiliki akurasi yang sangat baik dalam klasifikasi [11]. Klasifikasi pada penelitian ini data cuitan akan dibagi menurut 2 kelas yaitu kelas *cyberbullying* dan kelas *non-cyberbullying*. Data yang berupa cuitan nantinya akan diklasifikasikan menggunakan algoritma yang sudah ditentukan yaitu LSTM dan BiLSTM.

D. Evaluasi

Confusion matrix merupakan matriks yang menyimpan informasi untuk mengetahui performa dari model yang digunakan dan digunakan sebagai acuan dari performa klasifikasi dari algoritma yang digunakan pada tahap evaluasi [7].

Tabel 2. *Confusion Matrix*

		Predicted Values	
		Positive (0)	Negative (1)
Actual Values	Positive (0)	TP	FP
	Negative (1)	FN	TN

Confusion Matrix adalah sumber informasi apakah model yang digunakan berkinerja baik atau tidak. Hal ini bisa dilihat dari nilai yang dari variabel TP (*True Positive*) dan variabel TN (*True Negative*) menunjukkan total prediksi benar yang dibuat oleh model. Sedangkan nilai variabel FP (*False Positive*) dan variabel FN (*False Negative*) menunjukkan total prediksi salah yang dibuat oleh model. Penghitungan kinerja model dapat dilakukan dengan menghitung nilai *accuracy*, *precision*, *recall*, dan *F1-Score* berdasarkan rumus yang terlihat pada persamaan (1), 2), (3) dan (4).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = \frac{2 \times Precission \times Recall}{Precission+Recall} \quad (4)$$

keterangan :

TP = Jumlah data kelas positif (0) diprediksi benar sebagai kelas positif (0)

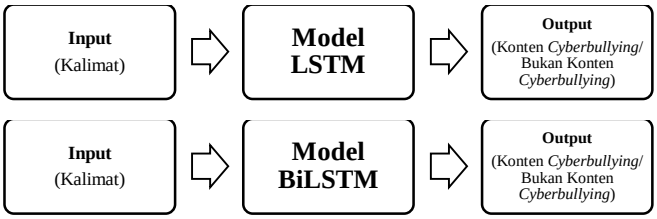
FN = Jumlah data kelas positif (0) diprediksi salah sebagai kelas negatif (1)

TN = Jumlah data kelas negatif (1) diprediksi benar sebagai kelas negatif (1)

FP = Jumlah data kelas negatif (1) diprediksi salah sebagai kelas positif (0)

E. Deteksi Konten

Deteksi konten merupakan tahapan akhir untuk mendeteksi kalimat apakah merupakan konten *cyberbullying* atau bukan. Tahapan ini menggunakan model dengan nilai *accuracy* terbaik dari skenario yang telah dibuat sebelumnya. **Error! Reference source not found.**5 menunjukkan proses deteksi konten.



Gambar 5. Deteksi Konten

VI. HASIL DAN PEMBAHASAN

A. Skenario

Semua model akan menggunakan *embeding layer* dengan ukuran 32 dimensi. Semua layer dengan 32 *units* dan sebuah *dense layer* dimana fungsi aktivasinya menggunakan *Sigmoid* dan fungsi optimasinya menggunakan *Adam*. Semua model akan dijalankan menggunakan 20 *epoch* dan 32 *batch size*.

Hasil dari model akan dibandingkan satu sama lain untuk melihat model manakah yang terbaik dalam klasifikasi dengan melihat beberapa parameter yaitu *Accuracy, Pression, Recall, dan F1-Scor*e. Berdasarkan hasil cacah *dataset* menunjukkan ketidakseimbangan yaitu 43% banding 57% sehingga perlukan beberapa parameter di atas untuk melihat model manakah yang memiliki performa terbaik.

B. Performa Model

Data yang sudah dibersihkan pada langkah *preprocessing* selanjutnya dibagi menjadi dua yaitu data *testing* dan data *training* dengan cacah masing-masing 0.2 data *testing* sejumlah 1367 data dan 0.8 data *training* sejumlah 5468 data. Berikut disajikan tabel hasil dari masing-masing *confusion matrix* algoritma *deep learning.*

Tabel 3. *Confusion Matrix* LSTM

LSTM			
Actual Values	*Predicted Value*		
		0	**1**
	0	**743**	**54**
	1	**46**	**524**

Tabel 4. *Confusion Matrix* BiLSTM

BiLSTM			
Actual Values	*Predicted Value*		
		0	**1**
	0	**771**	**26**
	1	**50**	**520**

Berdasarkan *confussion matrix* yang sudah didapat selanjutnya nilai dari *accuracy, precision, recall,* dan *F1-Score* dapat diketahui. Tabel 5 merupakan nilai dari masing-masing variabel tersebut.

Tabel 5. *Accuracy, Precision, Recall,* dan *F1-Score*

Metode	*Accuracy*	*Precision*	*Recall*	*F1-Score*
LSTM	**93.77**	**91.59**	**92.45**	**92.02**
BiLSTM	**95.24**	**94.29**	**93.40**	**93.84**

Berdasarkan tabel 5, algoritma BiLSTM terbukti lebih unggul dibanding dengan LSTM baik dari segi dari *accuracy, precision, recall,* maupun *F1-Score.*

C. Deteksi Konten

Setelah menguji performa dari model tersebut langkah selanjutnya yaitu menyimpan model dan berikutnya melakukan deteksi kalimat apakah merupakan konten *cyberbullying* atau bukan *cyberbullying*. Gambar 6 merupakan contoh deteksi kalimat.

Gambar 6. Deteksi Kalimat

Berdasarkan Gambar 6 menunjukkan bahwa model sudah bisa melakukan deteksi kalimat mana yang merupakan konten *cyberbullying* atau bukan.

VII. KESIMPULAN

Berdasarkan hasil yang telah dicapai, penelitian ini telah berhasil melakukan identifikasi cuitan bermakna *cyberbullying* pada media sosial *Twitter* dengan melakukan klasifikasi antara dua kelas cuitan yang tersedia pada *dataset.*

Hasil evaluasi dari masing-masing algoritma *deep learning* menempatkan *Bidirectional Long Short Term Memory* sebagai algoritma terbaik dalam mengklasifikasi data cuitan dibanding dengan *Long Short Term Memory*. Hal ini dibuktikan dengan nilai dari *accuracy*, *precision*, *recall*, dan *F1-Score* dari BiLSTM yaitu 95.24; 94.29; 93.40; 93.84.

Namun, *Long Short Term Memory* juga algoritma yang cukup baik dalam mengklasifikasi data teks. Ini dibuktikan dengan nilai dari *accuracy*, *precision*, *recall*, dan *F1-Score* yang selisih sedikit nilainya dengan *Bidirectional Long Short Term Memory*.

Penelitian ini masih sebatas model dalam klasifikasi cuitan yang bermakna *cyberbullying* pada media sosial. Penulis berharap penelitian ini bisa dikembangkan dengan melakukan menambah *dataset*, menambah data *training*, menambah fitur kelas tidak hanya *cyberbullying* dan *non-cyberbullying* saja tapi hasil kelas *cyberbullying* bisa diekstrak menjadi kelas yang lebih spesifik lagi seperti *cyberbullying* berupa *bully, racism, sexism,* dan *attack*. Terakhir menambah algoritma untuk dibandingkan untuk melihat algoritma mana yang lebih baik.

DAFTAR PUSTAKA

[1] Kominfo, "Kominfo : Pengguna Internet di Indonesia 63 Juta Orang," *kominfo.go.id*, 2013. https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker (accessed Mar. 26, 2020).

[2] B. Clinten, "Pengguna Aktif Harian Twitter Indonesia Diklaim Terbanyak," *kompas.com*, 2019. https://tekno.kompas.com/read/2019/10/30/16062477/pengguna-aktif-harian-twitter-indonesia-diklaim-terbanyak (accessed Mar. 26, 2020).

[3] F. K. Bohang, "Disebut sebagai Penyebar Konten Negatif, Ini Kata Twitter," *kompas.com*, 2017. https://tekno.kompas.com/read/2017/12/06/17122657/disebut-sebagai-penyebar-konten-negatif-ini-kata-twitter (accessed Jul. 15, 2020).

[4] Redaksi WE Online, "Kacau, Konten Porno Paling Banyak Berasal dari Twitter! Kemenkominfo Ambil Langkah Apa Nih?," *wartaekonomi.co.id*, 2020. https://www.wartaekonomi.co.id/read271389/kacau-konten-porno-paling-banyak-berasal-dari-twitter-kemenkominfo-ambil-langkah-apa-nih (accessed Jul. 15, 2020).

[5] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Arch. Suicide Res.*, vol. 14, no. 3, pp. 206–221, 2010, doi: 10.1080/13811118.2010.494133.

[6] B. A. Laksana, "Mensos: 84% Anak Usia 12-17 Tahun Mengalami Bullying," *news.detik.com*, 2017. https://news.detik.com/berita/d-3568407/mensos-84-anak-usia-12-17-tahun-mengalami-bullying (accessed Jun. 24, 2020).

[7] A. F. Hidayatullah, A. A. Yusuf, K. P. Juwairi, and R. A. Nayoan, "Identifikasi Konten Kasar pada Tweet Bahasa Indonesia," *J. Linguist. Komputasional*, vol. 2, no. 1, pp. 1–5, 2019, doi: 10.26418/jlk.v2i1.15.

[8] N. Abdulloh and A. F. Hidayatullah, "Deteksi Cyberbullying pada Cuitan Media Sosial Twitter," *automata*, vol. Vol 1, no. 1, pp. 1–5, 2019.

[9] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network Homa," *arXiv Prepr. arXiv1503.03909*, 2015, doi: 10.1007/978-3-319-27433-1_4.

[10] V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "Cyberbullying detection on twitter using Big Five and Dark Triad features," *Pers. Individ. Dif.*, vol. 141, pp. 252–257, 2019, doi: 10.1016/j.paid.2019.01.024.

[11] S. Agrawal and A. Awekar, "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms," *Eur. Conf. Inf. Retr.*, pp. 141–153, 2018.

[12] N. K. Manaswi, *Deep Learning with Applications Using Python*. 2018.

[13] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016, vol. 2, doi: 10.18653/v1/p16-1101.

[14] A. F. Hidayatullah and M. R. Ma'arif, "Pre-processing Tasks in Indonesian Twitter Messages," *J. Phys. Conf. Ser.*, vol. 801, pp. 1–6, 2017.