Terbit online pada laman web jurnal: **https://ejurnalunsam.id/index.php/jicom/**

Jurnal Informatika dan Teknologi Komputer

Vol. 03 No. 02 (2022) 104 - 108 E-ISSN :2774-7115 P-ISSN: 2775-2089

Kajian Penelitian Pembelajaran Mesin Untuk Pemrosesan Bahasa Alami Dalam Kalimat Perundungan Di Media Sosial

Matias Kristian Kelviandy [1*], Ahmad Ihsan [2]

[1]Sistem Komputer, Imu Komputer dan Teknologi Informasi, Universitas Gunadarma

[2]Program Studi Informatika, Universitas Samudra

[*]matias2kk2@gmail.com,

Abstract

This research study discusses machine learning opportunities for natural language processing (NLP) in bullying sentences on social media. This study consists of 4 chapters. The first chapter of the introduction tells the introduction which consists of the background, objectives and research opportunities that can be studied from previous research. The second part of the literature review contains the basic concepts of thought from this research study. The third part contains the concept of research studies that can be done to explore machine learning for natural language processing in bullying sentences on social media, starting from the concept of input and the concept of the process that produces the desired output. The last part is the conclusion of this research study. This research contains opportunities for researchers who want to study natural language processing using machine learning.

Keywords: Natural Language Processing, Machine Learning, Bullying sentences

Abstrak

Kajian Peneliti ini membahas peluang pembelajaran mesin (machine Learning) untuk pemrosesan Bahasa alami (NLP),dalam kalimat perundungan di media sosial. Kajian ini terdiri dari 4 bab. Bab pertama pendahuluan menceritakan pendahuluan yang terdiri dari latar belakang, tujuan dan peluang penelitian yang dapat dikaji dari penelitian sebelumnya. Bagian kedua kajian pustaka berisi dasar konsep pemikiran dari kajian penelitian ini. Bagian ketiga berisi konsep kajian penelitian yang dapat dikerjakan untuk mendalami pembelajaran mesin untuk pemrosesan bahasa alami dalam kalimat perundungan di media sosial, dimulai dari konsep masukkan dan konsep proses yang menghasilkan keluaran yang diingikan. Bagian terakhir kesimpulan dari kajian penelitian ini. Secara garis besar kajian ini berisi peluang bagi para peneliti yang ingin mengkaji pemrosesan bahasa alami menggunakan pembelajaran mesin.

Kata kunci: Pemrosesan Bahasa Alami (NLP), Pembelajaran Mesin (*Machine Learning*), Perundungan (*Bullying*).

1. Pendahuluan

Bahasa Indonesia merupakan bahasa resmi dan nasiona warga negara Indonesia, yang saat ini digunakan dan diucap oleh sekitar lebih dari 250 juta warga Negara Indonesia (BPS, 2017). Fakta tersebut menujukkan bahwa banyaknya jumlah warga disuatu negara, berdampak juga dengan beragamnya bahasa dan makna bahasa walaupun masih dalam konteks bahasa Indonesia, hal tersebut diyakini kurangnya penelitian pemrosesan bahasa alami (NLP, *Natural Language Processing*) (Mirna Adriani dan Ruli Manurung, 2008). Sehingga adanya peluang besar dalam pengembangan penelitian pemrosesan bahasa alami (NLP) berbahasa Indonesia terkait dalam korpus bahasa Indonesia (Nazief,2000), yaitu dengan mengembangkan dalam sebuah sistem, seperti sistem dialog, sistem pencarian informasi, atau sistem pembelajaran mesin (*Machine Learning*).

Penggunaan pembelajaran mesin dalam konteks penelitian bahasa alami, sudah dimulai dari tahun 90-an, dan berkembang di awal tahun 2000-an . saat ini aplikasi penelitian bahasa alami dengan menggunakan teknik pembelajaran mesin mempunyai tantangan dalam akuisisi data yang menjadi pusat perhatian bagi para peneliti dan para komunitas penggunan penelitian pemrosesan bahasa alami (Marquez, 2000).

Inti dari permasalahan yang sering muncul dengan teknik pembelajaran mesin adalah adanya makna dari suatu kalimat atau kata yang sering muncul dalam kalimat pemrosesan bahasa alami memiliki nilai ambigu

Diterima Redaksi : 25-08-2022 | Selesai Revisi : 18-09-2022 | Diterbitkan Online : 01-10-2022

yang tinggi atau perbedaan makna yang tinggi dalam kata yang sama, ini dikarenakan pembalajaran mesin dikenal dalam wilayah kecerdasan buatan (AI) yang hanya mampu mengenal klasifikasi objek atau kata dan pengenalan pola kata, hal ini menjadi tantangan konsep pembelajaran mesin ketika mendapat satu pola kata yang sama tapi memili makna yang berbeda dalam konteks pemrosesan kata. (Marquez, 2000).

Metode pembelajaran mesin saat ini sebagian besar bekerja dengan baik jika memiliki representasi yang sesuai dengan perancangan dari manusia, perkembangan metode pembelajaran mesin dalam konteks pemrosesan bahasa alami telah mengalami kemajuan pesat karena sudah mampu mengambil dan mengekstrak informasi dari suatu sumber, seperti dari web atau media sosial yang tentunya memiliki data tidak terstruktur dan memiliki nilai sentiment yang tinggi, hal ini sudah memenuhi tujuan dari pembentukan penelitian bahasa alami yaitu mengembangkan algoritma untuk pemrosesan kata dan menyelesaikan masalah linguistic atau bahasa melalui pendekatan representasi pembelajaran. (socher, 2014). Akan tetapi adanya tantangan dalam pendekatan tersebut yaitu Adanya asumsi bahasa yang diserderhanakan, sering di dalam pemrosesan bahasa alami dengan menggunakan pembelajaran mesin adanya pemaksaan data yang masuk ke dalam format yang sesuai dengan algoritma yang digunakan, misal dalam proses klasifikasi teks, sering mengabaikan susunan kata gramatikal yang mewakili makna teks, sehingga menyebabkan masalah ketika hasil dari algoritma tersebut menghasilkan kalimat yang acak dan tidak berurutan, sehingga makan kalimat tidak dapat dimengerti. . (socher, 2014).

Kajian penelitian dalam jurnal ini membahas peluang pemrosesan bahasa alami dengan menggunakan pembelajaran alami dengan membahas tantangan yang ada dan peluang untuk menyelesaikan tantangan tersebut, sehingga diharapkan adanya model baru yang didapat dalam pembaharuan pembelajaran mesin untuk menyelesaikan tantangan yang pemrosesan bahasa alami. Salah satu metode yang memungkinkan bisa menjawab hal tersebut adalah dengan menggunakan metode *Deep Learning*, hal ini dikarenakan beberapa penelitian besar telah berhasil dilakukan dengan menggunakan metode *Deep Learning* dalam klasifikasi citra (Krizhevsky et, al., 2012), dan pengenalan suara (Hinton et al., 2012), sehingga kemungkinan besar penggunaan metode ini dapat digunakan pada pemrosesan teks dan pemrosesan bahasa alami.(socher 2014)

Pemrosesan bahasa alami dalam konteks nya sangat luas untuk dibahas, banyak tema yang digunakan untuk mencoba menggunakan Pemrosesan bahasa alami dengan pembelajaran mesin, salah satunya bisa diangkat dengan tema sosial yang masih menjadi perhatian di negara Indonesia, yaitu tema perundungan (*bullying*) di dunia maya. Data dari UNICEF Indonesia di tahun 2016, sebanyak 41-50 persen remaja Indonesia dalam rentang usia 13-15 tahun pernah mengalami tindakan perundungan dunia maya, yang mengakibatkan permusuhan, dan efek negatif lainnya. Sehingga ini menjadi tantangan tersendiri cara dan upaya bagaimana mendeteksi percakapan yang mengandung perundungan di dunia maya

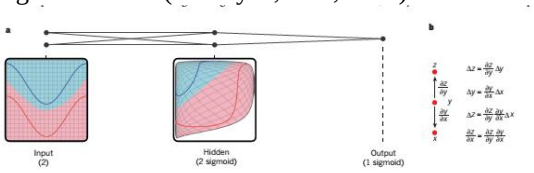
2. Metode Penelitian

Pada bagian ini, merupakan metode penelitian yang digunakan dengan metode *Deep Learning* untuk kalimat perundungan dunia maya

2.1 **Supervised Learning**

Bentuk paling umum dari pembelajaran mesin adalah Algoritma Pembelajaran yang diawasi (*supervised Learning*). Pembelajaran ini membuat sebuah sistem berdasarkan golongan, misalkan ada tiga objek dalam suatu lingkungan, maka objek tersebut akan diberi label sesuai dengan kategori, selama pelatihan algoritma ini diperlihatkan objek yang telah diberi label dan membentuk suatu nilai vector dalam satu kategori, sehingga adanya nilai awal dalam pelatihan ini terhadap sebuah objek dan dijadikan nilai dasar sebagai identifikasi atau parameter pengenal objek.

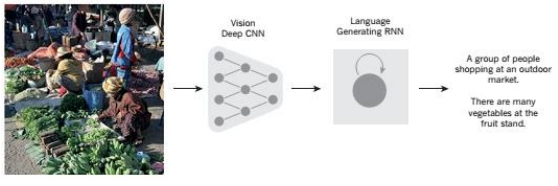
Parameter ini sering disebut bobot yang merupakan angka yang khas sebagai pembeda label objek satu dengan objek lainnya, sehingga jumlah bobot yang ada sama dengan jumlah objek yang ada didalam suatu lingkungan, objek ini pada saat proses pengujian sudah bisa diatur oleh pembelajaran mesin model pembelajaran yang diawasi ini. (lecun yan, dkk., 2015)



Gambar 1. Pembelajaran terawasi(lecun yan, dkk., 2015)

2,2 Metode Convolutional neural networks

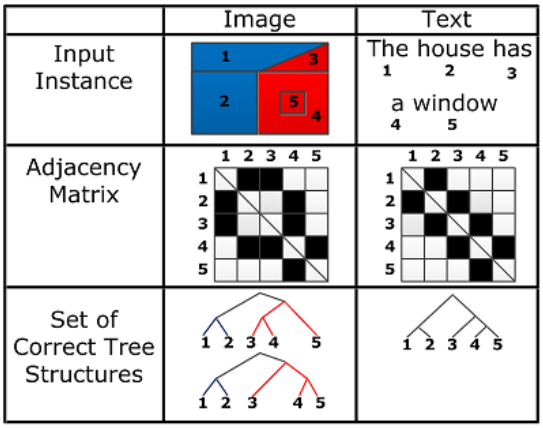
Convolutional Neural Networks (CNN), merupakan salah satu aplikasi pembelajaran mesin, yang menggunakan proses jaringan syaraf tiruan sebagai analisis data, dalam contoh kasusnya CNN digunakan untuk menganalisa suatu visual citra. CNN terinsipari oleh proses biologis dimana pola pemikiran dibuat sebagaimana neuron pada jaringan syaraf bekerja, dengan konsep variasi *perceptron Multilayer* yang dirancang sebagai pengambil keputusan dalam analisa keseluruhan suatu visual citra dan kemungkinan terbesar dapat digunakan dengan data lainnya seperti suara dan tulisan/teks (LeCun, Yann, 2015)



Gambar 2. Konsep CNN (lecun yan, dkk., 2015)

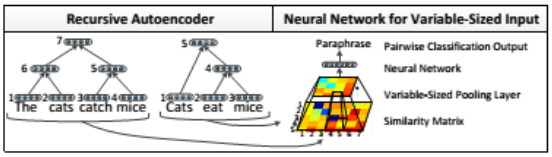
2.3 Recursive Deep Learning (Recursive Neural Network (RNN))

Metode ini diajukan oleh Richard Socher Tahun 2014, mengembangkan model baru dari pembelajaran terawasi (*supervised Learning*) dengan tidak diawasi yang digabungkankan menjadi model baru yang diperkenalkan. Metode ini menggeneralisasi ide dari suatu metode pembelajaran yang diaplikasikan dalam pembelajaran mesin ke dalam struktur hirarkis, model ini diaplikasikan untuk mendeteksi paraphrase, analisis sentiment, klasifikasi relasi kata dalam penyelesaian berbasis pengetahuan.



Gambar 3. Model RNN (Socher, 2014)

Pada gambar 3 dijelaskan bahwa masukkan dalam model RNN ini bisa dalam dua bentuk yaitu berupa gambar (Image) dan teks (Text), kedua masukkan ini dimasukkan kedalam kedekatan pola dalam bentuk matriks, setiap masukkan atau teks dijadikan dalam bentuk angka metriks untuk mewakili indeks dalam suatu kata, matriks ini dapat dijadikan aturan kebenaran dari pohon struktur yang akan dipelajari oleh mesin untuk mengenal pola uji dari model parafrase.



Gambar4. Model Parafrase RNN (Socher, 2014)

Pada gambar 4 merupakan gambaran umum dari model parafrase dari RNN, *Recursive Autoencoder* merupakan fitur unik dari dari suaru kata yang disajikan dalam simpul pohon, adanya jarak di setiap simpul merupakan nilai kemiripan dalam bentuk matriks yang akan dijadikan masukkan pada arsitektur pembelajaran mesin Jaringan syaraf tiruan, lalu hasil akhir dapat membandingkan kalimat setiap variabel dan mengklasifikase apakah terdapat parafrase yang sesuai atau tidak (socher, 2014).

3. Hasil dan Pembahasan

Pada bagian ini, hasil dan pembahasan akan diuraikan konsep kajian penelitian pembelajaran mesin untuk pemrosesan bahasa alami untuk mengetahui kalimat yang mengandung perundungan atau tidak.

Konteks konsep kajian penelitian ini dihubungkan bagaimana pembelajaran mesin mengetahui konteks kalimat yang dijadikan sebagai masukkan mengandung makna perundungan atau tidak, oleh karena itu perlu satu konsep matang untuk mewujudkan hal tersebut. Berikut bagian-bagian konsep kajian penelitian

3.1 Konsep Perancangan Masukkan

Konsep perancangan masukkan digunakan sebagai konteks awal bagaimana pembelajaran mesin dapat mengkategori kalimat perundungan dari suatu kalimat yang ada, Menurut Nancy Willard, penulis Cyber-Safe Kids dan Cyber-Savvy (2007), kalimat perundungan siber (*Cyber Bullying*) adalah merupakan kalimat sengaja yang dilakukan oleh pihak yang ingin menyembunyikan identitasnya, untuk berulang kali melakukan penghinaan, mengancam , melecehkan atau mempermalukan orang lain melalui penggunaan alat-alat media digital seperti internet, telepon seluler, Tablet, laptop dalam media teknologi informasi seperti e-mail, sosial media, forum dan lain lain.

Salah satu kategori perundungan siber menurut Nancy Willard adalah *Flamming*, yaitu penggunaan bahasa kasar atau fulgar melalui teknologi informasi media sosial. Kategori lainnya adalah *Outing and Trickery*, yaitu kegiatan mengungkapkan informasi yang memalukan atau mengungapkan rahasia untuk menjatuhkan salah satu pihak dengan cara berbagi foto dengan tulisan melalu media online.

Dari kedua kategori tersebut maka konsep perancangan masukkan terbagi menjadi dua, yaitu masukkan teks dan masukkan gambar yang didalamnya terdapat tulisan.



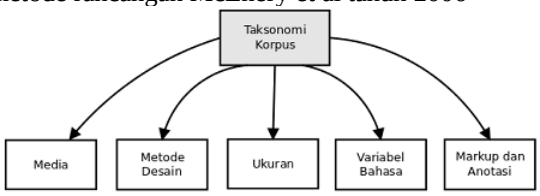
Gambar 5 Konsep Perancangan Masukkan

Pada gambar 5 dijelaskan konsep masukkan diambil dari sosial media, lalu masukkan dibagi menjadi dua yaitu teks dan gambar, masukkan melalui teks akan langsung diproses melalui konsep pemrosesan bahasa alami (NLP), kemudian hasilnya akan dimasukkan dalam

pembelajaran mesin (*Machine Learning*) untuk menentukan konteks teks yang diambil dari media sosial cenderung perundungan atau tidak. Masukkan kedua berupa gambar yang didalamnya berada tulisan, tulisan yang berada di dalam gambar di segmentasi atau dipisahkan dari gambar, sehingga masukkan yang diperlukan hanya teks untuk diproses selanjutnya melalui konsep pemrosesan bahasa alami (NLP), dan pembelajaran mesin (*Machine Learning*)

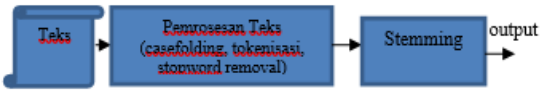
3.2 **Konsep Pemrosesan Bahasa Alami (NLP)**

Dalam konsep pemrosesan bahasa alami, tahap pertama yang dilakukan adalah pembentukan korpus, dalam kajian penelitian ini korpus yang digunakan adalah korpus linguistic, yaitu proses pengolahan himpunan kata yang dapat mempresentasikan faktor khusus atau tertentu dan korpus harus mewaliji dari generalisasi variasi bahasa (Leech, 1991). Pada kajian penelitian ini menggunakan taksonomi korpus yang digunakan metode rancangan McEnery et al tahun 2006



Gambar 6. Taksonomi Korpus (McEnery et al., 2006)

Gambar 6 menujukkan taksonomi korpus yang harus mencakup beberapa aspek, yaitu media, metode desain, ukuran, variable bahasa, markup dan anotasi, sehingga kajian penelitian ini merujuk aspek taksonomi tersebut.
Untuk langkah selanjutnya dalam konsep pemrosesan Bahasa Alami adalah pemrosesan teks (*text processing*). pemrosesan teks digunakan untuk mengambil informasi yang ada dalam satu kalimat dan menentukan bagian kata yang digunakan untuk penelitian, dan bagian kata yang tidak digunakan, sehingga langkah ini merupakan langkah penting bagi pemrosesan bahasa alami (Kannan S., 2015). Menurut Kannan S, langkah pemrosesan teks pada gambar 3

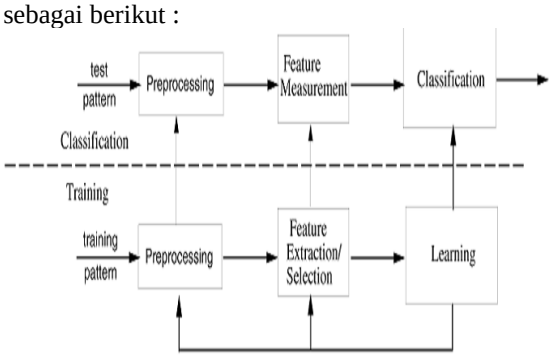


gambar 7 Pemrosesan Teks alami (Kannan S., 2015).

Pada gambar 3 merupakan pemrosan teks yang akan digunakan pada kajian penelitian ini, yaitu meratakan semua kata dengan huruf kecil (*casefolding*), membagi kata dalam setiap kalimat menjadi satu kata yang unik (tokenisasi), menghapus kata yang tidak perlu seperti kata sambung , dan kata lain yang tidak digunakan dalam penelitian (*stopword removal*), kemudian hasil dari pemrosesan teks menjadi masukkan *stemming* yaitu menghapus kata berimbuhan seperti (ber, be, -an, dan lain lain).

3.3Konsep Segmentasi Gambar dari Teks

Pada bagian ini dijelaskan konsep dari masukkan gambar yang terdapat teks, untuk dianalisa makna nya mengandung konteks perundungan atau tidak. Masalah segmentasi sebenarnya merupakan bagian dari pengolahan cita, dimana adanya tantangan besar bagi penglihatan komputer untuk melihat informasi yang terdapat pada gambar. Dalam kajian penelitian ini komputer harus melihat apakah terdapat teks dalam suatu gambar, jika ada maka teks harus di segmentasi atau diambil dari gambar tersebut untuk selanjutnya di proses melalui bahasa pemrosesan alami. Metode segmentasi gambar sangat banyak telah dilakukan oleh beberapa peneliti sebelumnya, diantaranya *Statistical pattern recognition Method.*
Statistical pattern recognition Method merupakan metode segmentasi pengenalan pola statistic yang digunakan untuk merancang sejumlah pola penanda dalam satu gambar, pola diwaliki oleh serangkaian fitur yang dipandang sebagai sebuah vector dimensi, dalam arti setiap informasi yang ada di dalam suatu gambar dijasikan sebuah dimensi berbentuk angka dan disajikan dalam vector. Metode ini digunakan dalam kajian penelitian karena beberapa penelitian telah berhasil mengapilkasikannya dalam sistem komersial (Subba, 2011), dan keunggulan dalam melakukan segmentasi, serta digunakan untuk hubungan atau sebab akibat antara masukkan dan keluaran (ashraf., 2011). Adapun rancangan dari *Statistical pattern recognition Method* sebagai berikut :



gambar 8 *Statistical pattern recognition Method* (Subba, 2011)

pada gambar 4 dijelaskan rangkaian metode *statistical pattern recognition*, ketika masukkan masuk berupa pola (*pattern*) dalam hal ini gambar dilakukan proses pemrosesan awal (*preprocessing*), untuk memperbaiki citra, menormalkan pola, membagi pola yang tampak dan serupa, dan mengurangi gangguan gambar (*noise*), langkah selanjutnya adalah *extraction* atau ekstraksi pada model latihan (*training*), hal ini untuk melakukan seleksi atau menemukan fitur yang sesuai untuk mewaliki pola masukkan dalam hal ini teks, dan

membuang informasi yang tidak perlu seperti layar gambar, kemudian proses *learning* dilakukan untuk melatih pola tersebut dalam representasi matriks, sebagai satu nilai unik dari suatu pola . Dalam pengujian pola (*test pattern*) terdapat fungsi *Feature measurement* digunakan untuk pengukuran pola yang akan diuji oleh proses klasifikasi dan dibandingkan dengan pola teks hasil pelatihan (*training*).

3.4 Konsep *Machine Learning*

Pada bagian konsep ini *machine learning* atau pembelajaran mesin yang digunakan pada kajian penelitian ini adalah menggunakan *Recursive Deep Learning*, dengan konsep yang sudah tertera pada gambar 3 dan gambar 4.

Dengan menggunakan *Recursive Deep Learning* secara baik dapat melakukan klasifikasi dalam pemrosesan kata semantik dengan pola hitung yang baik, sehingga berpotensi untuk pemrosesan kalimat dan anotasi (socher., 2014)

4. Kesimpulan

Berdasarkan hasil dari kajian penelitian ini adalah adanya potensi besar dalam mendeteksi kalimat perundungan media sosial, hal-hal yang dapat dilakukan adalah sebagai berikut :

1. Mengambil sampel kalimat dan gambar yang cenderung mempunyai makna perundungan dari media social
2. Memproses kalimat dan gambar (yang terdapat kalimat) dengan menggunakan pemrosesan bahasa alami
3. Khusus dari masukkan berupa gambar yang terdapat kalimat, dilakukan proses segmentasi atau pemisahan kalimat dari gambar, dan dilanjutkan proses pemrosesan bahasa alami
4. Hasil dari proses pemrosesan bahasa alami menjadi masukkan pembelajaran mesin (*machine learning*) dengan menggunakan *recursive deep learning* untuk mengambil keputusan makna dari kalimat yang dihasilkan terdapat perundungan atau tidak
5. Dalam proses nya kajian penelitian ini dilakukan dua proses yaitu proses pelatihan untuk mengambil pola kalimat perundungan dalam database, dan proses pengujian yaitu proses menguji kalimat dan diklasifikasi apakah kalimat tersebut mengandung makna perundungan atau tidak dengan membandingkan dengan pola yang ada dalam database

Daftar Rujukan

[1] YuhefiA. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet classication with deep convolutional neural networks. In NIPS

[2] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag., 29(6):82{97}

[3] Lagmin, Gomez.(2016. Bullying Prevention –UNICEF Report Reveals Bullying Is a Worldwide Problem. Diakses pada 2 Juli 2017 dari : http://www.thegomezfirm.com/bullying-prevention-unicef/

[4] Kannan S, Gurusamy V.2015. Preprocessing Techniques for Text Mining.

[5] Leech, G. 1991. The state of the art in corpus linguistics.In English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik. Longman.

[6] Marquez, L.2000.Machine Learning And Natural Language Processing.Universitat Politecnica de Catalunya. UPC

[7] McEnery, T., Xiao, R., and Tono, Y. 2006. Corpusbased language studies. Routledge London.

[8] Mirna Adriani and Ruli Manurung. 2008. "A survey of bahasa Indonesia NLP research conducted at the University of Indonesia" .Jakarta : Faculty of Computer Science, University of Indonesia

[9] Nazief, B. 2000. Development of computational linguistics research: A challenge for indonesia. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 1–2. Association for Computational Linguistic

[10] Socher,R 2014. Recursive Deep Learning For Natural Language Processing And Computer Vision. A Dissertation Submitted To The Department Of Computer Science And The Committee On Graduate Studies Of Stanford University.

[11] Subba M.2011. Comparative Analysis of Pattern Recognition Methods: An Overview. Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166 Vol. 2 No. 3 Jun-Jul 2011

[12] Yann LeCun, Yoshua Bengio & Geoffrey Hinton.2015. Deep Learning :Review.

[13] Willard, Nancy E.2007. Cyberbullying and Cyberthreats : Responding challenge of Online Aggression, Threats and Distress. United States, Research Press