DETEKSI UJARAN KEBENCIAN DENGAN MENGGUNAKAN ALGORITMA CONVOLUTIONAL NEURAL NETWORK PADA GAMBAR

HATESPEECH DETECTION USING CONVOLUTIONAL NEURAL NETWORK ALGORITHM BASED ON IMAGE

Bagas Prakoso Putra[1], Budhi Irawan, S.Si., M.T.[2] Casi Setianingsih, S.T., M.T..[3]

[1,2,3]Prodi S1 Sistem Komputer, Fakultas Teknik Elektro, Universitas Telkom
[1]imbagas@student.telkomuniversity.ac.id, [2]budhiirawan@telkomuniversity.ac.id, [3]setiacasie@telkomuniversity.ac.id

Abstrak

Ujaran kebencian adalah perkataan, perilaku, tindakan yang dilarang karena menimbulkan terjadinya tindak yang memicu kekerasan dan sikap anarkis terhadap individu atau kelompok yang lain. Etika dalam ber-internet perlu ditegaskan mengingat internet merupakan hal yang dianggap kebutuhan pentingmbagi masyarakat jaman sekarang. Tetapi, semakin banyak pihak yang menyalah gunakan internet untuk menyebarluaskan hal-hal yang berkaitan dengan ujaran kebencian, seperti suku bangsa,agama dan ras. karena penyebaran berita yang bersifat ujaran kebencian di internet, menjadi hal yang patut diperhatikan.

Pengembangan sistem untuk mendeteksi ujaran kebencian melalui gambar memang cukup jarang untuk untuk saat ini. Maka dari itu penelitian ini diklasifikasikan untuk mendeteksi adakah unsur ujaran kebencian pada gambar yang nantinya dipilih.

Dalam Tugas Akhir ini, Penulis berharap bisa membuat bagaimana cara mengklasifikasi unsur ujaran kebencian pada sebuah gambar yang dilakukan oleh komputer, yang nantinya komputer bisa mengenali adakah ujaran kebencian pada gambar melalui teks yang ada. Dengan menggunakan metode *Deep Learning* dengan algortima *Convolutional Neural Network (CNN)*. Setelah pembuatan aplikasi ini, diharapkan Komputer dapat mengetahui dan bisa mengklasifikasi adakah ujaran kebencian dengan mendeteksi gambar tersebut.

Kata Kunci: CNN *(Convolutional Neural Network , Deep Learning*

Abstract

Hate speech are a words,actions which is prohibited because it leads to acts that trigger anarchic and violence attidudes toweard other individuals or groups. Ethics in the internet are needed considering that internet is a matter that important use for today's society. However, more parties are miss using the internet to spread such kind a hate speech, such as ethnicity, religion and race.

The development of a system for detecting hate speech through images is quiet rare for now a days. therefore this study is classified to detect whether there is an element of hatred in the image that will be selected in this final project, the author hopes to make how to classify the element of hate speech in an image performed by the machine learning, which later that machine leanring can recognize any kind of hate speech on the image through the existing text. By using Deep Learning method with Convolutional Neural Network (CNN) algorithm. After making this application, machine learning is expected to know and classify any hate speechby detecting some text on an images.

Keywords: *Convolutional Neural Network (CNN), Deep Learning*

1. Pendahuluan

Pada 25 Agustus 2017, Bareskrim polri menyatakan telah menangkap tiga orang dari kelompok Saracen, yang merupakan sindikat penyedia jasa konten kebencian berdasarkan suku, agama, ras dan antar golongan. Mereka mempunyai 2.000 akun media sosial yang kemudian berkembang menjadi 800.000 akun, yang digunakan untuk menyebar konten kebencian.[1]

Salah satunya konten yang dibagikan adalah ujaran kebencian (*hate speech*), yang mengekspresikan rasa benci terhadap sekelompok orang tertentu, konflik agama dan etnis yang berakibat pencemaran nama baik pada yang bersangkutan. Berdasarkan pasal 27 ayat (3) UU ITE "Setiap orang dengan sengaja dan tanpa hak mendistribusikan atau mentransmisikan atau membuat dapat diaksesnya informasi elektronik atau dokumen elektronik yang bermuatan penghinaan dan/atau pencemaran nama baik".

Dengan tersedianya fitur unggah gambar pada media sosial (*Facebook, Twitter dan Instagram*), beberapa oknum telah memanfaatkan fitur tersebut dengan tidak bertanggung jawab dalam penyebaran ujaran kebencian diberbagai media sosial, karena *image* dinilai sebagai media paling menarik dan mudah diterima oleh masyarakat.

Solusi dari banyaknya penyebaran ujaran kebencian di publik yaitu dengan membuat sistem deteksi ujaran kebencian berbasis *image to text*. Pada sistem tersebut menggunakan *input* berupa *image* yang mengandung teks dimana nantinya akan dikonversikan menjadi teks, kemudian teks tersebut akan dianalisa, apakah teks tersebut mengandung ujaran kebencian atau tidak dengan menggunakan metode *Deep Learning* dengan algoritma *Convolutional Neural Network. Deep learning* dikenal sangat bagus pada *vision,sentiment analysis* dll. *Sentiment analysis*, sudah banyak digunakan dalam riset menggunakan model *deep learning*, karena memiliki koneksi dan parameter yang jauh lebih sedikit dan lebih mudah untuk dilatih, *Convolutional Neural Network* adalah model dalam yang sangat kuat dalam memahami kontak gambar.[2]

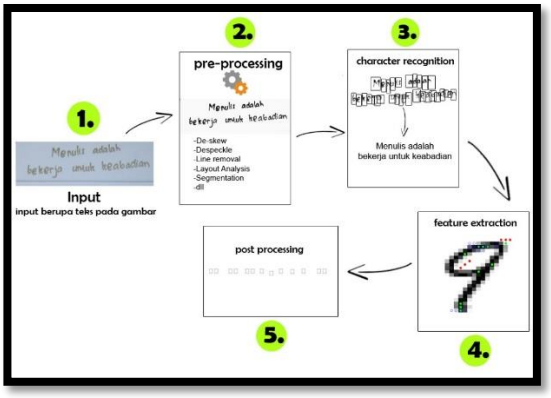
2. Dasar Teori

2.1 Ujaran Kebencian

Ujaran kebencian adalah tindakan komunikasi yang dilakukan oleh suatu individu atau kelompok dalam bentuk provokasi, hasutan, ataupun hinaan kepada individu atau kelompok yang lain dalam hal berbagai aspek seperti ras, warna kulit, gender, cacat, orientasi seksual, kewarganegaraan, agama dan lain-lain.
Dalam arti hukum Ujaran Kebencian *(Hate Speech)* adalah perkataan, perilaku, tulisan, ataupun pertunjukan yang dilarang karena dapat memicu terjadinya tindakan kekerasan dan sikap prasangka entah dari pihak pelaku pernyataan tersebut ataupun korban dari tindakan tersebut. *Website* yang menggunakan atau menerapkan Ujaran Kebencian *(Hate Speech)* ini disebut *(Hate Site)*. Kebanyakan dari situs ini menggunakan Forum Internet dan Berita untuk mempertegas suatu sudut pandang tertentu.

2.2 Optical Character Recognition

OCR (Optical Character Recognition) adalah salah satu metode paling sering dipakai untuk melakukan konversi data dari Image to Text. OCR (Optical Character Recognition) adalah modul yang berfungsi untuk melakukan scan gambar yang menghasilkan image dan dijadikan text, aplikasi ini juga bisa menjadi support aplikasi tambahan untuk scanner. Dengan adanya OCR, Image yang bertulisan tangan, tulisan mesin ketik atau computer text, dapat dimanipulasi. Text yang discan dengan OCR dapat dicari baris per baris dan setiap text dapat dimanipulasi, diganti, atau diberikan barcode.[3]

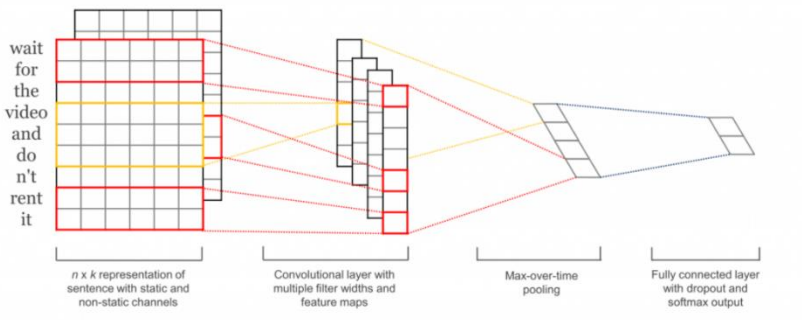


Gambar 2.1 Proses OCR

Pada Gambar 2.1 OCR dijelaskan bahwa memeliki beberapa tahap sebelum hasil dari gambar yang telah di training menjadi sebuat teks yang utuh. Tahap-tahap yang dimiliki OCR: Pertama Input: Input berupa teks yang terdapat pada gambar, umumnya gambar yang digunakan adalah gambar yang memiliki pixel yang bagus / resolusi yang tidak pecah sehingga memudahkan sistem OCR untuk membacanya. Kedua Pre-Processing: Tahap untuk meningkatkan kualitas hasil. didalam pre-processing memiliki bermacam-macam teknik: *De-skew, Despeckle, Line removal, Layout analysis, segmentation dan normalize*. Ketiga Character *recognition:* menggunakan *Matrix matching* yaitu membandingkan gambar ke mesin yang tersimpan berdasarkan piksel per piksel. Feature extraction: Menguraiakan glyphs menjadi "fitur" seperti garis, loop tertutup, arah garis dan persimpangan garis. Fitur ekstraksi mengurai dimensi representasi dan membuat proses recognition menjadi efisien. Keempat *Post-Processing*: Proses yang dilakukan pada tahap ini adalah proses koreksi ejaan sesuai dengan bahasa yang digunakan.

2.3 Convolutional Neural Network pada NLP

CNN pada biasanya banyak diaplikasikan pada pengenalan gambar, namun berbeda pada NLP input yang digunakan adalah kalimat atau dokumen yang di representasikan sebagai matriks. setiap baris dari matriksnya sesuai degan satu token, biasanya kata, tetapi bisa juga karakter.dimana setiap barisnya adalah vektor yang merepresentasikan sebuah kata. Biasanya, vektor ini adalah *word embeddings (low-dimensional representasition)* seperti word2vec tetapi mereka juga bisa menjadi *one-hot vektor* yang mengindeks kata menjadi kosakata.[4] Misalkan 1 kalimat mengandung 10 kata lalu menggunakan 100 dimensi embedding maka kita memiliki matriks 10x100 sebagai input. Berikut contoh visualisasi CNN pada NLP.

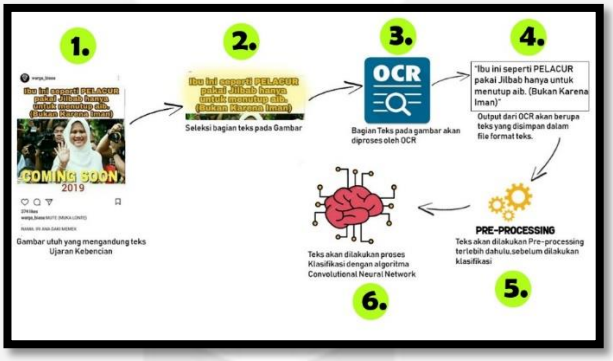


Gambar 2.2 Visualisasi CNN pada NLP

3. Pembahasan

3.1 Gambaran Umum Sistem

Dalam Penelitian ini akan diterapkan salah satu teknik pengklasifikasian data mining dengan menggunakan algoritma *Convolutional Neural Network.* Algoritma ini berguna untuk membentuk pola yang akan mempresentasikan pengklasifikasian. Study case yang dibahas dalam tugas akhir ini adalah mengindentifikasi kalimat ujaran kebencian pada gambar dengan mengambil data dari banyak akun Twitter. yang sudah dikategorikan menjadi Kata Sara, Sifat, Hewan, Politik Adapun input yang diberikan adalah gambar yang mengandung teks , dan output yang akan dicari adalah dari hasil gambar yang mengandung teks tersebut apakah termasuk ujaran kebencian atau bukan.

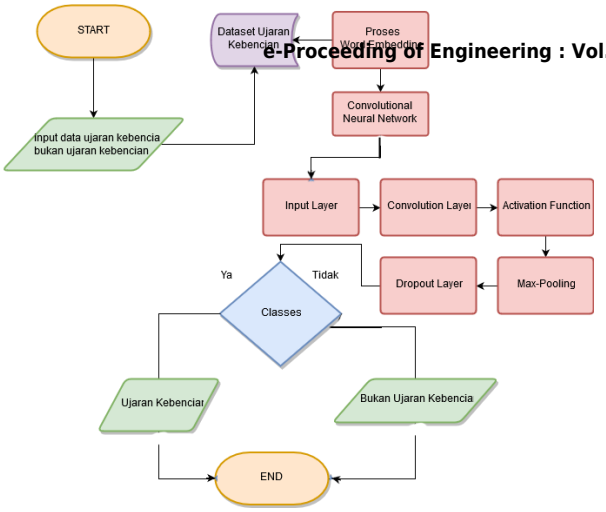


Gambar 3,1 Gambaran umum sistem simulator mengemudi

Pada **Gambar 3.1** diperlihatkan alur pengerjaan sistem :Pertama Gambar utuh yang mengandung teks Ujaran Kebencian di Upload pada sistem sebagai input awal.Kedua Gambar dilakukan seleksi pada bagian teks, untuk mempermudah proses deteksi teks.Ketiga Gambar teks tersebut akan dilakukan proses OCR, untuk mengolah gambar menjadi teks yang utuh.Keempat Output dari OCR akan menghasilkan berupa teks yang akan disimpan dalam bentuk file format teks (txt,csv).Kelima Teks akan dilakukan Pre-Processing terlebih dahulu (Case Folding, Stemming, Stopword Cleaning, Pos Tagging) untuk memmpermudah proses klasifikasi nantinya. Keenam Hasil Teks akhir dilakukan proses Klasifikasi dengan metode Convolutional Neural Network.[5]

3.2 Perancangan Sistem

Data yang akan digunakan dalam tugas akhir ini adalah data yang diambil dari berbagai akun yang dianggap menyebarkan *tweet* ujaran kebencian. Dan data tersebut sudah di verifikasi oleh **Balai Bahasa Provinsi Jawa Barat** total data sebanyak **1253** data Berikut contoh data latih yang akan diolah oleh system.
Dari banyak tweet yang diambil maka diberikan label untuk dikenali oleh sistem nantinya dimana Pada **Tabel 3.1** kolom Label digunakan untuk *scoring* pada data untuk data yang mengandung ujaran kebencian diberikan nilai (1) dan untuk data yang tidak mengandung ujaran kebencian diberikan nilai (0), sedangkan kolom Tweet adalah isi dari tweet. Sebanyak 685 data (Ujaran Kebencian) dan 610 data (Bukan Ujaran Kebencian)



Gambar 3.2 Flowchart Convolutional Neural Network

Proses pengumpulan data Ujaran Kebencian dan Bukan Ujaran Kebencian sehingga menjadi dataset pada system Proses pengolahan kata oleh bagian klasifikasi dengan Convolutional Neural Network dengan dilakukan beberapa layer Neural Network melalui beberapa filter. Proses perbandingan nilai hasil dari dropout layer yang nanti akan di class kan berdasarkan kategori : ujaran kebencian atau bukan ujaran kebencian.

3.3 Perancangan

Pada sistem ini, data yang didapatkan dari Twitter diproses terlebih dahulu sebelum disimpan didalam *dataset txt*. Setelah itu, sistem akan melakukan *convolution* pada setiap layer untuk dijadikan pemetaan yang jelas untuk mendapatkan klasifikasi ujaran kebencian atau tidak. Sistem juga dapat menghitung seberapa baik klasifikasi ini melalui perhitungan *accuracy, precesion,* dan *recall.*

Gambar 3.3 Diagram Konteks sistem

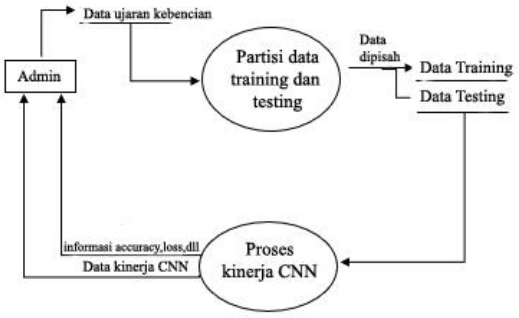
a) Proses Pengambil Keputusan

Setelah *dataset tweet* ujaran kebencian tersimpan kedalam *dataset txt*, maka akan di proses ketika input gambar yang mengandung teks. Setelah itu, algoritma CNN akan melakukan proses pengklasifikasian dan menampilkan keluaran berupa gambar tersebut dikategorikan sebagai ujaran kebencian atau bukan.

Gambar 3.4 Data Flow Diagram Level 1 (1)

b) Proses Perhitungan Kinerja

Proses ini digunakan untuk mengukur tingkat keakuratan sistem ini. Partisi data *Training* dan *testing* digunakan untuk membagi jumlah data yang menjadi acuan perhitungan keakuratan. Setelah data berhasil di partisi, sistem akan melakukan klasifikasi pada setiap data *testing* lalu membandingkan hasilnya dengan kondisi yang sebenarnya sehingga perbandingan tersebut dapat menghasilkan nilai *accuracy, precision,* dan *recall*



Gambar 3.5 Data Flow Diagram Level 1 (2)

4. Implementasi dan Pengujian Sistem

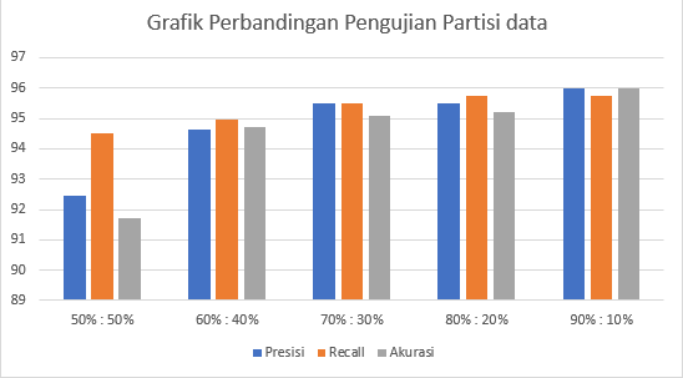
4.1 Pengujian Kinerja Sistem

Pengujian ini berfungsi untuk mengetahui kerja algoritma CNN dalam mengklasifikasikan data ke dalam kelas yang telah ditentukan sebelumnya. Pada percobaan pengujian ini diberikan data testing untuk menguji tabel keputusan yang sudah terbentuk. Untuk kerjanya diperoleh dengan memberikan nilai *confusion matrix* dengan menghitung nilai *precision*, *recall*, dan accuracy dari hasil pengujian

Tabel 4.1 Rangkuman Hasil Pengujian

Pengujian ke-	Data Partision (%)		Tabel Confusion Matrix		**Kelas Prediksi**	
	Training Data	Testing Data			UK	BUK
1	50	50	**Kelas Asli**	UK	320	33
				BUK	18	298
2	60	40		UK	251	14
				BUK	12	244
3	70	30		UK	194	9
				BUK	9	176
4	80	20		UK	139	5
				BUK	9	111
5	90	10		UK	74	3
				BUK	2	49

Dari hasil partisi data tersebut didapatkan nilai partisi data dengan tingkat akurasi tertinggi yaitu pada partisi data 90% : 10% dengan nilai 96%



Gambar 4.1 Grafik pengujian kinerja sistem

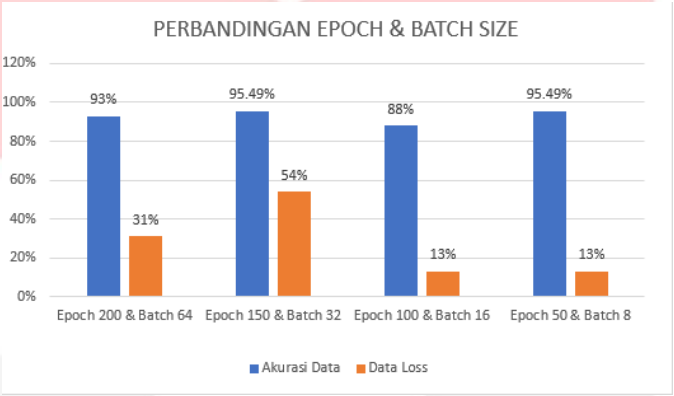
Didapatkan bahwa pengujian kelima dengan data partisi Training Data 90% dan Datatest 10% mendapatkan nilai Presisi, Recall dan Akurasi yang tinggi di antara yang lain yaitu sebesar 96%. Maka tahap selanjutnya adalah melakukan pengujian dengan parameter *epoch* dan *batch size* yang sesuai untuk mendapatkan nilai akurasi yang lebih tinggi.

4.2 Pengujian *Kinerja Sistem Epoch & Batch Size*

Tabel 4.2 Pengujian Kinerja Sistem

Pengujian ke-	Precision	Recall	Accuracy
50% : 50%	92,45%	94.45%	91.17%
60% : 40%	94.65%	94.95%	94.7%
70% : 30%	95.5%	95.5%	95.1%
80% : 20%	95.5%	95.75%	95.2%
90% : 10%	96%	95.75%	96%

Dari tabel pengujian kinerja sistem diatas dapat dilihat grafik perbandingannya dibawah ini :



Gambar 1 Hasil Tuning PID

Setelah gambar yang mengandung tweet diklasifikasi oleh CNN maka tahap selanjutnya akan dilakukan proses pengujian hasil klasifikasi. Pengujian akurasi dari hasil klasifikasi dilakukan dengan teknik confusion matrix dengan cara pembagian partisi data. Hasil pengujian akurasi didapatkan dengan sistem bahwa data partisi 90% : 10 % mempunyai tingkat akurasi paling tinggi dibanding yang lain, hal ini dikarenakan semakin banyak jumlah data Training maka sistem lebih mengenali lebih banyak kosa kata yang dipelajari sehingga sistem lebih cerdas mengenali klasifikasinya.

Sesudah melakukan pengujian sistem dengan cara partisi data didapatkan tadi bahwa partisi 90% : 20% memiliki nilai akurasi yang tinggi yaitu bernilai 96% maka pengujian kedua adalah mengubah parameter *epoch* dan *batch size.* Setelah dilakukan pengujian selama empat tahap didapatkan bahwa pada tahap keempat *epoch 50, batch size 8* dan tahap kedua *epoch 100, batch size 16* nilai akurasi mencapai paling tinggi

Setelah proses perhitungan CNN dilakukan, maka hasil klasifikasi akan ditampilkan dengan melihat dari hasil label apakah data bernilai (1) berarti ujaran kebencian (0) bukan ujaran kebencian. Berikut adalah tampilan dari hasil klasifikasi yang sudah dilakukan oleh CNN.

Gambar 4.2 Visualisasi hasil *tweet* dari CNN

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan dari hasil Tugas Akhir ini, dapat ditarik beberapa kesimpulan yaitu :

1. Proses klasifikasi data ujaran kebencian hanya diberikan 2 jenis kelas (ujaran kebencian atau bukan ujaran kebencian) menggunakan algoritma Convolutional Neural Network.
2. Semakin banyak data Train maka semakin bagus hasil Accuracy yang dihasilkan
3. Jumlah Filter dan Ukuran dimensi mempengaruhi proses training dan hasil training
4. Ukuran Epoch dan Batch size mempengaruhi kinerja sistem dan hasil accuracy
5. Pada proses pengujian kinerja sistem didapatkan rata-rata precision sebesar 99.46%, recall sebesar 97.99%, dan Accuracy sebesar 99.8%.
6. Berhasil memvisualisasikan hasil output ujaran kebencian dan bukan

5.1 Saran

1. Berdasarkan dari Penelitian Tugas akhir ini maka, penulis menyarankan untuk penelitian selanjutnya adalah sebagai berikut :
2. Penelitian lebih lanjut dapat ditambah jumlah dataset ujaran kebencian yang sudah di verifikasi oleh Dinas Balai Bahasa setempat
3. Penelitian lebih lanjut dapat menambah pengenalan ujaran kebencian melalui sebuah visual gambar secara langsung tanpa perantara teks.
4. Penelitian lebih lanjut dapat menambahkan integrasi aplikasi mobile sehingga pengguna tidak hanya tertuju pada web melainkan juga pada perangkat smartphone nya masing – masing sehingga mudah untuk melakukan pengambilan data melalui foto maupun screenshot gambar.

5.

Daftar Isi

[1] Komisi Nasional Hak Asasai Manusia Republik Indonesia. 2015. Buku Saktu Penanganan Ujaran Kebencian. Jakarta

[2] KEPALA KEPOLISIAN NEGARA REPUBLIK INDONESIA. 2015. Surat Edaran Nomor SE/6/X/2015. Jakarta.

[3] Shamanth Kumar, Fred Morstatter, Huan Liu Springer. Twitter Data Analytics, Science & Business Media, 11 Nov 2013. America

[4] Herbert F, Schantz. The History of OCR: Optical Character Recognition. 1982. Recognition Technologies Users Association. America.

[5] Abubakr H. Ombabi, Onsa Lazzez, Wael Ouarda, Adel M. Alimi. Deep learning framework based on Word2Vec and CNN for users interests classification. Computer Science and Information Technology (SCCSIT), 2017. Sudan.

[6] Denny Britz. Understanding Convolutional Neural Networks for NLP. Wildml. 2015.

[7] Stanford university. "Feature extraction using convolution". 2013. America..

[8] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1746–1751.

[9] Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification,

[10] Santos, C. N. dos, & Gatti, M. (2014). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In COLING-2014 (pp. 69–78)..