PENERAPAN *NEURAL NETWORK BACKPROGPAGATION* UNTUK KLASIFIKASI ARTIKEL *CLICKBAIT*

Rakhmad Maulidi[1], Muhammad Fahmi Ayilillahi[2], Laila Isyiriyah[3], Jozua F. Palandi[4]
Program Studi Teknik Informatika, STIKI Malang[1,2,3,4]
maulidi@stiki.ac.id

Abstrak. Clickbait sering digunakan untuk meningkatkan trafik pengunjung website, namun disalahgunakan oleh pengelola website dengan mengabaikan kepuasan pembaca berita dengan cara menampilkan judul yang hiperbola atau isi kontennya tidak sesuai dengan yang tertera pada judul berita. Pengklasifikasian artikel *clickbait* berbahasa indonesia menggunakan metode neural network dengan menggunakan fitur ekstraksi TF-IDF, Top word, tanda baca. Data training yang digunakan sejumlah 800 judul. Pengujian tingkat akurasi model menggunakan 185 judul dari dari uji, kemudian dilanjutkan dengan pengujuan menggunakan 20 judul data baru menghasilkan tingkat akurasi mencapai 85%.

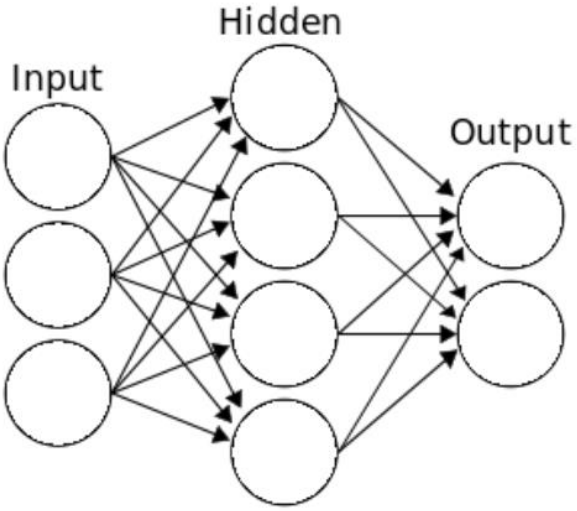
Kata Kunci: *Neural Network, Backpropagation , Klasifikasi dokumen, Clickbait.*

PENDAHULUAN

Clickbait merupakan sebuah bentuk konten web pada judul yang ditulis dengan formula dan tata bahasa yang bertujuan untuk memancing pembaca untuk mengklik tautan (Palau & Sampio, 2016). Clickbait digunakan meningkatkan trafik pengunjung website sehingga menjadi skema baru di dunia jurnalistik daring yang terus berkembang(Pogue, 2014). Clickbait kini digunakan secara berlebihan oleh pengelola website dengan cara memberi judul artikel yang tidak sesuai isi dari artikel (Patel, 2014). Semakin maraknya penyalahgunaan clickbait pada artikel berita online menyebabkan banyak pengguna atau pembaca berita online merasa tertipu karena konten artikel berita tersebut tidak sebagus dengan apa yang dicantumkan pada judul, bahkan banyak artikel yang konten beritanya tidak memuat apa yang dicantumkan pada judul berita. Hal ini dikarenakan persaingan dunia internet dalam mencari penghasilan semakin ketat, sehingga banyak penyedia berita online tidak mengutamakan kepuasan pembaca dan cenderung dinilai negatif oleh pembaca (Scacco & Muddiman, 2016)

Klasifikasi artikel clickbait berdasarkan judulnya bermanfaat bagi pembaca dan media online dengan mencari pola dari judul dari artikel. Salah satu cara yang digunakan adalah mendeteksi frekuensi kata yang sering muncul pada sebuah artikel, seperti yang dilakukan oleh (Wijaya & Santosa, 2016)) dengan menggunakan metode Naïve Bayes. Sayangnya pola judul artikel *clickbait* berkembang secara dinamis, media online menggunakan teknik linguistic untuk membuat judul artikel misalnya dengan permainan kata(misalnya menggunakan kata-kata yang tidak baku atau kata-kata sensasional), morphosyntax(misalnya menggunakan struktur kalimat yang sederhana), atau judul artikel yang "*eye catching*" (Palau & Sampio, 2016).

Neural Network adalah sebuah metode yang menyerupai jaringan saraf pada manusia, yang memiliki kemampuan untuk belajar dari beberapa contoh karena memiliki karakteristik dapat belajar dari data - data sebelumnya dan mengenal pola data yang selalu berubah. Selain itu, metode ini tidak terprogram sehingga semua keluaran (output) dari metode ini ditarik berdasarkan pengalaman pada proses pelatihan/pembelajaran. Neural Network sendiri memiliki berbagai metode untuk pengklasifikasian, salah satu metode yang umum digunakan adalah propagasi balik (*backpropagation*).



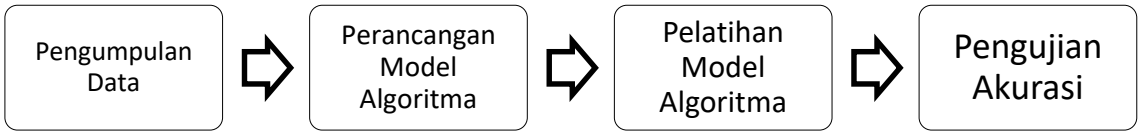
Gambar 1 Model Neural Network

Backpropagation atau propagasi balik merupakan suatu teknik pelatihan supervised learning yang paling banyak digunakan. Pada jaringan backpropagation, setiap unit yang berada pada input layer terhubung dengan setiap unit pada hidden layer, lalu setiap unit pada hidden layer terhubung dengan setiap unit pada output layer, seperti yang telah tertera pada Gambar Keunggulan dari metode ini adalah mampu mengenali dan mengekstraksi pola – pola yang berjumlah besar dan kompleks (Pradipa & Rahmawan, 2013)

Pada penelitian ini bertujuan untuk mengklasifikasikan artikel yakni berita berbahasa Indonesia dengan menggunakan Neural Network backpropagation dengan cara mendeteksi judul artikel untuk dikelompok kedalam artikel clickbait atau bukan dengan menggunakan ekstrasi fitur TF-IDF, tingkat kemunculan kata terbaik dan tanda baca.

METODE PENELITIAN

Metodelogi penelitian untuk klasifikasi artikel clickbait dengan cara memasukkan data kedalam rancangan model Neural Network untuk diproses sehingga akan menghasilkan keluaran berupa klasifikasi artikel. Gambar 1 merupakan Ringkasan proses penelitian dimana masing-masng sub proses akan dijelaskan dalam beberapa subbab berikut



Gambar 2 Tahapan Penelitian

A. Pengumpulan Data

Pengumpulan data dilakukan dengan mengumpulkan 1000 artikel berita, yang nanti 80% artikel akan menjadi data latih, dan 20% artikel akan menjadi data uji. Artikel berita pada penelitian ini dikumpulkan dari 13 belas situs berita online berbahasa Indonesia berikut:

1. http://www.detik.com/
2. http://www.republik.in/
3. http://www.tribunnews.com
4. http://www.republika.co.id/
5. https://www.brilio.net/
6. http://www.otomania.com/

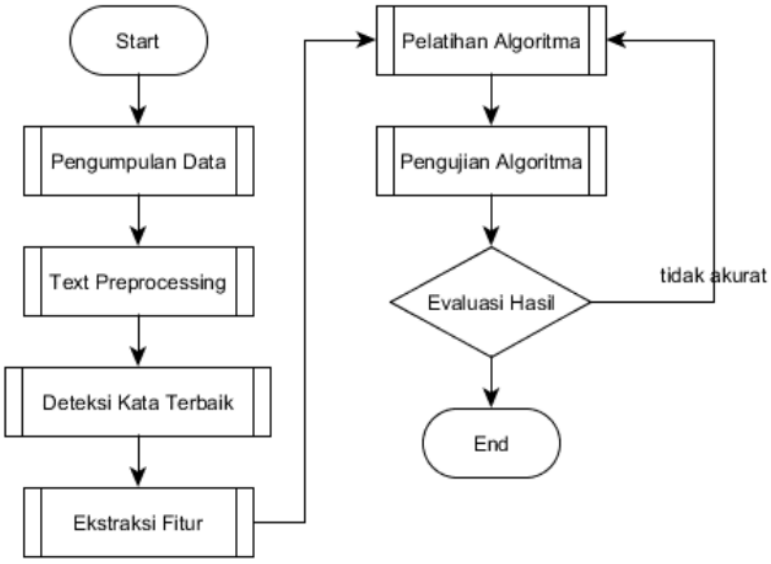
7. http://www.kompas.com/
8. http://www.bintang.com/
9. http://www.okezone.com/
10. https://www.kapanlagi.com/
11. https://www.kapanlagi.com/
12. http://www.jawapos.com/
13. https://www.tempo.co/

Artikel yang telah dikumpulkan akan dikelompokkan, dan dilabeli secara manual sesuai dengan kategorinya masing-masing yakni clickbait dan bukan clickbait. Pelabelan dilakukan 5 orang responden yang akan memberikan poin pada setiap artikel yang telah dikumpulkan yakni bukan clickbait (0), ragu – ragu (1), dan clickbait (2), kemudian dilanjutkan pemberian skor per data, pemberian skor diambil dari nilai rata-rata responden per data, sehingga apabila rata – rata yang didapatkan adalah lebih dari atau sama dengan 1 maka hasil yang didapatkan adalah clickbait, sebaliknya apabila kurang dari satu maka hasilnya bukan clickbait.

B. Perancangan Model Algoritma

Cara kerja dari rancangan model algoritma neural network untuk klasifikasi artikel yang dikembangkan tergambarkan pada gambar 3. Tahap Text preprocessing merupakan tahap awal dalam mengolah data input sebelum memasuki tahap utama. Text Preprocessing dilakukan untuk tujuan penyeragaman dan kemudahan untuk diolah pada tahap utama yaitu dengan menghilangkan noise, memperjelas fitur data, mengkonversi data asli agar diperoleh data yang sesuai dengan kebutuhan (Milatina, Syukur, & Supriyanto, 2012).

Text preprocessing pada penelitian ini, meliputi casefolding, tokenizing, dan stemming. Tahap Pendeteksian kata terbaik digunakan untuk mengidentifikasi kata apa saja yang sering muncul pada artikel clickbait dan tidak sering muncul pada bukan clickbait maupun kata yang sering muncul pada artikel bukan clickbait dan tidak sering muncul pada artikel clickbait. Tahap Ekstraksi fitur digunakan untuk mengambil ciri unik yang terdapat pada data masukan, Penelitian ini menggunakan beberapa fitur yang digunakan yakni TF-IDF yang TF-IDF, tingkat kemunculan kata terbaik dan tanda baca. Kemudian fitur tingkat kemunculan kata terbaik yang digunakan untuk mendeteksi jumlah kata terbaik pada sebuah judul dan fitur terakhir jumlah tanda baca yaitu pendeteksian jumlah tanda baca pada judul dimungkinkan berpengaruh pada hasil klasifikasi.



Gambar 3 Tahapan Klasifikasi NN Backpropagation

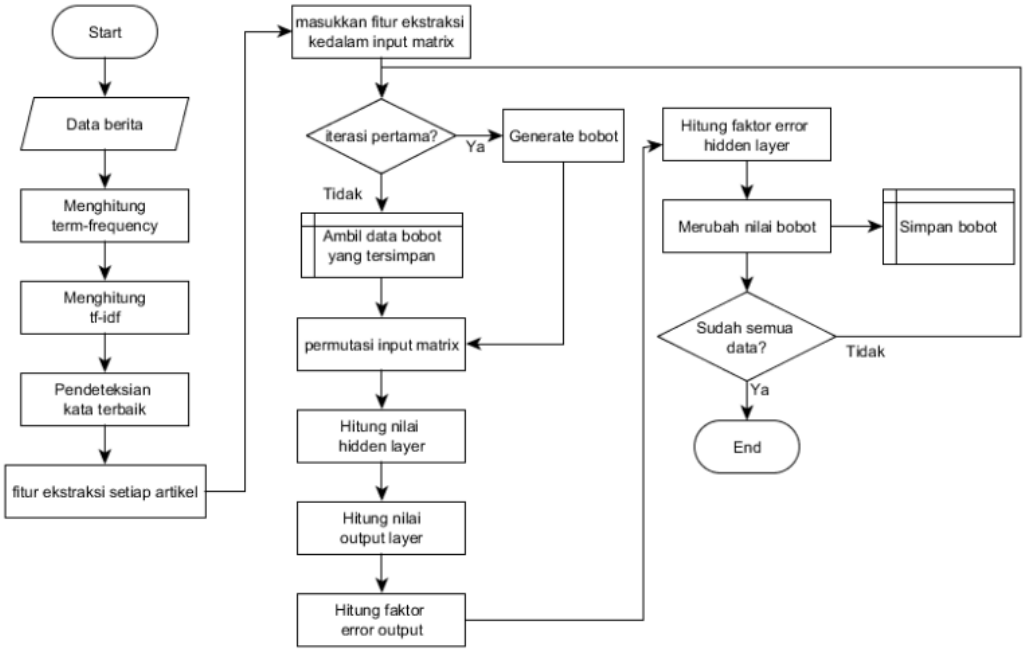
C. Pelatihan Model Algoritma

Tahapan pelatihan model algoritma menggunakan data training sejumlah 800 judul artikel yang dengan komposisi 400 artikel clickbait dan 400 artikel bukan clickbait. Proses pengambilan data traning ke dalam model dengan cara acak. Pelatihan dilakukan dengan menggunakan formula yakni kombinasi antara jumlah data dan configurasi model seperti pada table 1.

Table 1 Formula Pelatihan Algoritma

No	Jumlah Data	Epoch	Laju	Kata Terbaik	Hidden Neuron
1	800	500	0.5	120	60
2	800	700	0.45	120	60
3	800	700	0.45	120	120
4	800	700	0.45	120	180
5	800	700	0.45	80	40
6	800	1500	0.45	120	60
7	800	1000	0.45	50	25

Hasil fitur ekstraksi akan dijadikan sebagai nilai masukan setiap artikel pada tahap pelatihan algoritma. Tahapan pelatihan algoritma secara umum dapat dilihat pada Gambar 4.



Gambar 4 Proses detail pelatihan dalam 1 epoch

D. Pengujian Akurasi

Untuk dapat menganalisa akurasi dari sebuah klasifikasi dapat digunakan confusion matrix yaitu sebuah matrik dari prediksi yang akan dibandingkan dengan kelas yang asli dari data inputan dengan cara menentukan nilai True Negative (TN), True Positive (TP), False Negative (FN), False Positive (FP), False Positive (FP). Matrik ini ini dikembangkan oleh (Powers, 2011) yang sudah umum digunakan dalam penelitian machine learning. Pengujuin dilakukukan dengan menggunakan sejumlah data uji seperti yang terdapat pada table 2. Pengujuin akurasi menggunakan sejumlah data uji yang telah dikumpulkan, kemudian dilakukan pengujian lanjutan dengan menggunakan data baru yakni judul berita yang diambil secara langsung dari website.

HASIL DAN PEMBAHASAN

Table 2 Hasil pengujian terhadap formula pelatihan

Formula	Jumlah Data Uji (RANDOM)	TP	FN	FP	TN	Total CB	Total Bukan CB	Conf. Matrix	Recall	Precession
1	185	78	23	25	59	101	84	74,0%	77,2%	56,9%
2	185	79	22	30	54	101	84	71,9%	78,2%	59,3%
3	185	83	18	21	63	101	84	78,9%	82,2%	56,8%
4	185	77	24	12	72	101	84	80,5%	76,2%	51,7%
5	185	68	33	11	73	101	84	76,2%	67,3%	48,2%
6	185	82	19	24	60	101	84	76,8%	81,2%	57,7%
7	185	76	25	22	62	101	84	74,6%	75,2%	55,0%

Hasil pengujian dengan menggunakan data uji menunjukkan formula nomer 3 menghasilkan nilai recall yang terbaik yakni 82.2% dan sedangkan formula nomer 2 menghasilkan nilai precession yang terbaik yakni 59.3%. Sedangkan formula nomer 4 menghasilkan nilai confusion matrix yang paling tinggi yakni 80.5%.

Pengujian selanjutnya pada model untuk diuji untuk masing-masing formula dengan menggunakan data baru yang diambil dari website detik.com sejumlah 20 judul seperti pada table 3 berikut ini.

Table 3 Data Pengujian manual 20 data detik.com

No	Judul	Target	Terdeteksi
1	Habis Gerhana Parsial, Terbitlah Gerhana Matahari Total	Bukan Clickbait	Bukan clickbait
2	Berapa Jumlah Penyandang Disabilitas Nganggur? Ini Kata Menaker	Clickbait	Clickbait
3	Alasan Kesehatan, Keponakan Novanto Absen Panggilan KPK	Bukan Clickbait	Clickbait
4	Bikin Mupeng, Resort Ini Punya Bioskop Sampai Perosotan di Pulau	Clickbait	Clickbait
5	Terciduk! Asyik Joget Bareng Pedangdut, Pria Ini Kepergok Istri	Clickbait	Clickbait
6	Kamu Nggak Akan Sangka Kostum 'Game of Thrones' Ternyata Dibeli di Sini	Clickbait	Clickbait
7	Cerita Patung Bunda Maria di Balik Air Terjun Sanggau	Bukan Clickbait	Clickbait
8	Dear Noel, Liam Gallagher Masih Ingin Bersama Oasis	Bukan Clickbait	Bukan Clickbait
9	Madrid vs MU: Bersiap Menjadi Tim Terbaik di Eropa	Bukan Clickbait	Bukan Clickbait
10	Heboh meme the power of emak emak simak penjelasan psikolog	Clickbait	Clickbait
11	Begal Pasutri Dago Juga Menjambret Mahasiswi di Bandung	Bukan Clickbait	Bukan Clickbait
12.	Bos First Travel Ditangkap, Polisi Buru Tersangka Lain	Bukan Clickbait	Bukan Clickbait
13	Akhir Kisah Petualangan 'Raja Jambret' Sadis di Bandung	Bukan Clickbait	Bukan Clickbait
14.	SItuasi Panas AS-Korut: Apakah Kita Perlu Khawatir?	Clickbait	Clickbait
15	Begini Kondisi Slamet Jemaah yang Stroke dan Dikabarkan Telantar	Clickbait	Clickbait
16	Giliran Rumah Kabid Bina Marga Kota Malang Digeledah KPK	Bukan Clickbait	Bukan Clickbait
17	Kira-kira Orang Amerika Tahu Letak Korea	Clickbait	Clickbait

	Utara Nggak Ya?		
18	Soffi Bertemu Hantu di Palaguna	Bukan Clickbait	Bukan Clickbait
19.	Kecewa, Akankah Habiburokhman Coba Simpang Semanggi Lagi?	Clickbait	Clickbait
20	Uang di Rekening Tinggal Rp 1,3 Juta, ke Mana Dana First Travel?	Clickbait	Bukan Clickbait

Table 4 Hasil pengujian menggunakan 20 data detik.com

Formula	TP	FN	FP	TN	Total CB	Total Bukan CB	Conf. Matrix	Recall	Precession
1	7	3	3	7	10	10	70%	70%	50%
2	6	4	3	7	10	10	65%	60%	46,2%
3	7	3	1	9	10	10	80%	70%	43,8%
4	6	4	1	9	10	10	75%	60%	40%
5	6	4	2	8	10	10	70%	60%	42,9%
6	8	2	3	7	10	10	75%	80%	53,3%
7	9	1	2	8	10	10	85%	90%	52,9%

Hasil pengujian yang dilakukan dengan menggunakan data baru menunjukkan bahwa formula nomer 7 menghasilkan tingkat akurasi yang terbaik, yakni dengan nilai confusion matrix tertinggi yakni 85% dimana nilai recallnya sejumlah 90% dan precession sejumlah 52.9%. Formula terbaik tersebut menggunakan 1000 epoch, 0.45 laju dan 50 kata terbaik dan 25 hidden neoron. Nilai recall terbaik didapatkan oleh formula nomer 7, sedangkan nilai precession terbaik didapatkan oleh formula nomer 6 menggunakan 1500 epoch, 0.45 laju dan 120 kata terbaik dan 60 hidden neuron.

Dari hasil pengujian model dengan menggunakan data uji dan data baru menunjukan perbedaan urutan hasil akurasi confusion matrix. Pada pengujuian menggunakan data uji, formula nomer 4 menghasilkan nilai confusion matrix tertinggi, sedangkan pengujian menggunakan data baru menghasilnya formula nomer 7 sebagai peraih nilai confusion matrix tertinggi. Hal ini disebabkan pengambilan data dalam waktu yang berbeda antara data uji dan data baru yang berselisih 4 bulan, hal ini menunjukkan bahwa pola judul artikel clickbait berubah(dinamis), tidak statis, hal ini sejalan dengan penelitian yang dilakukan oleh (Palau & Sampio, 2016).

PENUTUP

Kesimpulan

1. Nilai confusion matrix tertinggi yakni 85% dimana nilai recallnya sejumlah 90% dan precession sejumlah 52.9%.
2. Faktor yang mempengaruhi keakurasian algoritma sangat bervariasi diantaranya jumlah epoch, jumlah hidden neuron, data yang dilatih fitur yang digunakan, dan laju pelatihan.
3. formula dengan akurasi paling optimum yaitu jumlah epoch 1000, data yang dilatih 800, laju pelatihan 0.45, jumlah kata terbaik 50, jumlah hidden neuron 25.
4. Pola judul article clickbait terbukti tidak statis sejalan dengan penelitian yang dilakukan oleh (Palau & Sampio, 2016)

Saran

1. Rata – rata akurasi pada penelitian ini adalah berkisar 75% sehingga perlu dikaji ulang pemilihan fitur, dan formula yang bagus untuk mendeteksi artikel clickbait.
2. Penelitian selanjutnya terkait dengan clickbbait dapat mendeteksi dan mengklasifikasi apakah judul artikel sesuai dengan konten pada artikel tersebut.

DAFTAR RUJUKAN

Palau, D., & Sampio. (2016). Reference press metamorphosis in the digital context: clickbait and tabloid strategies in elpais. com. *Communication & Society, 29*(2).

Milatina, Syukur, A., & Supriyanto, C. (2012). Pengaruh Text Preprocessing Pada Clustering Dokumen Teks Berbahasa Indonesia. *Jurnal Teknologi Informasi, 8*.

Patel, N. (2014, Juli 16). *The real problem with clickbait*. Dipetik Mei 10, 2018, dari Poynter A global leader in journalism: http://www.poynter.org/2014/the-realproblem-with-clickbait/258985/

Pogue, D. (2014, Mei 22). *You Have No*. Dipetik Januari 1, 2018, dari Yahoo!: https://www.yahoo.com/tech/you-haveno-idea-whats-behind-these-clickbait-

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning*, 37-63.

Pradipa, E., & Rahmawan, E. (2013). *Klasifikasi Pola Konten E-mail Menggunakan Jaringan Syaraf Tiruan Metode Backpropagation Untuk Pengecekan Spam E-mail dengan Data Acuan DMC 2003.* Semarang: Universitas Dian Nuswantoro.

Scacco, J., & Muddiman, A. (2016). *Investigating the influence of Investigating the influence of.* Diambil kembali dari The Center for Media Engagement - Moody College of Communication at the University of Texas at Austin: https://engagingnewsproject.org/wpcontent/uploads/2016/08/ENP-Investigating-the-Influence-of-ClickbaitNews-Headlines.pdf

Wijaya, A., & Santosa, H. (2016). Naive Bayes Classification pada Klasifikasi Dokumen Untuk. *Journal of Applied Intelligent System*, 48-55.