# DETEKSI CLICKBAIT DENGAN *SENTENCE SCORING BASED ON FREQUENCY* DI DETIK.COM

Budi Wijaya Rauf, Suwanto Raharjo, Heri Sismoro

*Program Studi Magister Teknik Informatika, Universitas Amikom Yogyakarta*
Jl. Ring Road Utara, Sleman, Yogyakarta
Budi0029@students.amikom.ac.id, wa2n@akprind.ac.id, herisismoro@amikom.ac.id

***Abstract*** - The clickbait phenomenon has become one of the powerful ways to increase the number of readers for a website. With the increasing number of visitors to the site, the higher the income on the website. However, this clickbait technique is like a double-edged knife. Many people who do not like this technique because of incompatibility of the title and content of the article being read. This study aims to detect clickbait articles on the Indonesian news site detik.com by using python and sentence scoring based on frequency algorithms to find a match between the title and content of the article. Appropriate titles and contents will be given a value of 1 (one) and those that are not appropriate will be given a value of 0 (zero), the results of the test system are compared with existing datasets and produce an accuracy rate of 75%, 57% precision and 80% recall.
***Keywords*** - Clickbait Detection, Sentence Scoring Based on Frequency, Python.

***Abstrak –*** Fenomena *clickbait* sudah menjadi salah satu cara ampuh untuk meningkatkan jumlah pembaca untuk sebuah website. Dengan meningkatnya jumlah pengunjung pada situs maka semakin tinggi pula pendapatan pada website tersebut. Namun, Teknik *clickbait* ini seperti pisau bermata dua. Banyak masyarakat yang tidak suka dengan Teknik ini dikarenakan ketidaksesuaian judul dan isi artikel yang dibaca. Penelitian ini bertujuan untuk mendeteksi artikel clickbait pada situs berita Indonesia detik.com dengan menggunakan python dan algoritma sentence scoring based on frequency guna mencari kecocokan antara judul dan isi dari artikel tersebut. Judul dan isi yang sesuai akan diberikan nilai 1 (satu) dan yang tidak sesuai akan diberikan nilai 0 (nol), hasil uji coba system tersebut dibandingkan dengan dataset yang telah ada dan menghasilkan tingkat akurasi sebesar 75%, presisi 57% dan recall 80%.
***Kata Kunci*** - Deteksi Clickbait, Sentence Scoring Based on Frequency, Python.

## I. PENDAHULUAN

Dengan semakin pesat perkembangan teknologi,maka semakin gampang pula orang-orang dapat mengakses informasi. Jika dahulu informasi seperti berita terkini, kecelakaan, terorisme, politik dan kejadian tertentu hanya dapat diakses di media cetak seperti koran, kini informasi tersebut dapat dengan mudah diakses di website berita atau di sosial media.

Kemudahan akses informasi ini membuat para penulis berita berbondong-bondong menyebar artikel milik mereka ke sosial media dan untuk menambah jumlah pembaca biasanya meraka menggunakan judul yang bernada provokatif untuk membuat pembacanya menjadi penasaran, strategi tersebut biasa dinamakan *clickbait.* Pada umumnya artikel dengan *clickbait* ini fokus pada subjek seperti selebriti, rumor, akun fiktif daripada isu-isu yang yang serius dan akademik. (Pegnante, 2016)

Fenomena *clickbait* ini terjadi karena penulis/editor berupaya agar menambah pembaca dengan membuat judul yang menarik sehingga membuat orang yang melihatnya tidak mampu untuk melewatkan judul tersebut. Hal ini terjadi karena adanya gap informasi (*information gap*) antara yang diketahui oleh pembaca dan yang ingin diketahui pembaca. Loewenstein (1994) mengemukakan teori *Information Gap* yang pada dasarnya melihat celah antara "apa hal yang kita ketahui dan yang ingin kita ketahui", kesenjangan itu yang menimbulkan konsekuensi emosional. Dengan kesenjangan itu pun membuat individu merasakan kehilangan sesuatu. Akhirnya individu tersebut termotivasi untuk mencari tahu informasi yang ia rasa kehilangan untuk menghilangkan rasa kehilangannya tersebut[1].

Disisi lain, masyarakat Indonesia sangat menyukai berita dengan nada provokatif atau berita hangat yang sebenarnya tidak terlalu penting. Hal ini dibuktikan dengan jumlah klik berita dan *view* dari yang bertambah pesat ketika menggunakan judul *clickbait*. Hal inilah yang dimanfaatkan oleh para penulis berita di Indonesia, sehingga membuat beberapa berita yang tersebar di media sosial memiliki berita dengan judul yang bombastis.

Penelitian *clickbait* ini sudah banyak dilakukan sebelumnya, namun untuk di Indonesia belum banyak penelitian yang fokus terkait hal ini. Penelitian yang dilakukan oleh Yayat D. Hidayat (2019) yang membahas tentang penggunaan judul *clickbait* pada artikel yang tersebar di media online. Penelitian ini menggunakan metode penelitian kualitatif dengan objek penelitian, judul-judul di media online Indonesia yang dipilih berdasarkan kriteria-kriteria tertentu yang sesuai dengan masalah penelitian[2].

Jurnal dari M Rizky Kertanegara (2018) membahas penggunaan judul clickbait pada situs berita Dream.co.id untuk menarik pembaca dengan tujuan mendapatkan Key Performance Indicator (KPI) atau indikator kinerja kunci yang tinggi yang nantinya berguna untuk kepentingan beriklan. Berdasarkan hasil temuan, sebagian besar artikel yang paling banyak dilihat menggunakan clickbait headline namun tetap memenuhi standar Kode Etik Jurnalistik dari Dewan Pers[3].

Penelitian yang dilakukan oleh Godham Eko Saputra (2019) yang yang membahas komik strip dan fenomena *clickbait* yang ada di sosial media, Penelitian ini menggunakan metode kualitatif dalam meneliti fenomena *clickbait* melalui proses studi literatur, observasi serta mencari pengalaman langsung dengan mengakses artikel dengan judul *clickbait*. Untuk analisa permasalahan menggunakan framing dengan memaparkan beberapa realita terpilih untuk menemukan statement sebagai dasar perancangan karya. Metode Design Thinking dipilih untuk mengembangkan komik strip sebagai media untuk memberikan informasi seputar fenomena clickbait[4].

Nathan Hurst (2016) melakukan penelitian ini bertujuan untuk mengetahui apakah berita dengan artikel *clickbait* dapat menunjukan kredibilitas dengan membandingkan 2 artikel, 2 sumber berita dan 2 tingkat kredibilitas. Hasilnya adalah berita dengan *clickbait* menghasilkan hasil yang negatif dengan tingkat kreidibilitas rendah[5].

Penelitian dari Iva Belestin (2017) meneliti terhadap judul artikel yang bersifat *clickbait*, penelitian ini bersifat kuantitatif dengan jumlah sampel partisipan berjumlah 406 orang. Analisa juga dilakukan dengan meneliti 1619 artikel dari portal berita paling populer yang ada di Serbia. Hasil penelitian ini menunjukkan bahwa ada sebanyak 33,11% judul yang bersifat *clickbait* yang ada di situs berita terkenal seperti Blic, Novosti dan Kurir. Judul clickbait untuk majalah tabloid sebanyak 29,4%. Untuk genre berita paling populer untuk judul *clickbait* adalah *entertainment* 49,77% dan *lifestyle* 45,97%. Genre judul *clickbait* terendah ada pada seksi politik yaitu hanya 0,97%. Dari penelitian ini menunjukkan bahwa portal berita dengan jumlah *clickbait* terbanyak adalah Kurir (50,56%)[6].

Penelitian yang dilakukan oleh Abhijnan Chakraborty(2016) bertujuan untuk mendeteksi secara otomatis artikel *clickbait* dengan merancang ekstensi browser yang akan memperingatkan pengguna ketika membuka situs-situs di internet akan adanya kemungkinan *clickbait* pada judul artikel. Ekstensi ini juga memiliki fitur untuk memblok artikel yang bersifat *clickbait* untuk para pengguna, Ekstensi ini dapat dijalankan online maupun offline serta telah dicoba ke beberapa portal berita, hasilnya ekstensi ini berjalan baik dan mendeksi serta memblok artikel dengan tingkat akurasi deteksi 93% dan akurasi pemblokiran 89%[7].

A. Sentence Scoring Based on Frequency

*1. Preprosessing Data*

Pada fase pra-proses data, kalimat yang diambil dari dokumen akan dibagi menjadi per kata, berikut merupakan tahapan-tahapan yang ada pada pra proses data[8]:

a. Segmentasi Kalimat
Pada tahap ini, artikel yang diambil akan dibagi menjadi beberapa segmentasi kalimat. Kalimat-kalimat tersebut akan dibagi berdasarkan tanda baca yang ada contohnya seperti titik (.), koma (,), tanda seru(!) dan lainnya.

b. Tokenisasi
Setelah diproses menjadi beberapa segmentasi, kalimat tersebut lalu akan ditokenisasi atau dibagi lagi menjadi kata per kata. Tokenisasi ini dilakukan dengan mengidentifikasi kata-kata dari kalimat tersebut dengan spasi ( ), koma(,) maupun titik(.). Pada tahap ini juga kalimat yang telah ditokenisasi akan disimpan ke dalam database dan akan digunakan untuk menghitung frekuensi kata.

c. Penghapusan Stop Word
Stop word atau kata yang sifatnya membantu sebuah kalimat yang ada pada dokumen. Penghapusan stop word ini berguna untuk menghilangkan kata-kata yang dianggap tidak penting oleh system seperti kata yang, ada, pada, atau, dan, untuk dan sebagainya.

d. Stemming
Stemming merupakan proses dimana kata yang telah ditokenisasi akan dikembalikan ke

bentuk dasarnya. Contoh kata tersebut adalah, menghidupkan menjadi hidup, berjuang menjadi juang dan kata-kata yang ditambahkan imbuhan lainnya.

*2.* Sentence Scoring

Setelah fase pra-proses data telah dilakukan, maka proses selanjutnya ialah memberikan nilai terhadap kata yang berdasarkan 5 poin berikut ini[9]:

a. Frekuensi
Frekuensi ialah jumlah kata yang muncul dalam sebuah artikel. Jika sebuah kata selalu ada pada artikel maka dapat dikatakan bahwa kata tersebut memiliki efek besar dalam konten di artikel tersebut. Semakin tinggi frekuensi kata tersebut, maka semakin tinggi pula nilai kata itu. Perhitungan yang selalu digunakan untuk mengkakulasikan frekuensi sebuah kata di artikel adalah TF (Term Frequency) dan IDF (Inverse Document Frequency). Total frekuensi yang telah dihitung akan berdampak besar untuk nilai terhadap kata tersebut.

b. Nilai posisi kalimat
Posisi kalimat juga memberikan nilai lebih dalam sebuah artikel. Seperti kalimat awal dari paragraph dan kalimat akhir dari paragraph. Kalimat dengan nilai yang tinggi berada pada kalimat pertama dan terakhir dari sebuah paragraph, nilai kedua tertinggi untuk kalimat kedua dan kedua terakhir dari sebuah paragraf serta nilai nol (0) untuk kalimat selanjutnya.

c. Kesamaan dengan judul
Pada proses inilah focus pada penelitian ini. Dengan membandingkan kesamaan judul dan konten maka akan terlihat apakah artikel tersebut merupakan *clickbait* atau bukan. Selain itu, ketika kata di judul dan di kalimat sama maka akan diberikan nilai satu (1) terhadap kalimat tersebut, jika tidak maka nilainya nol (0) .

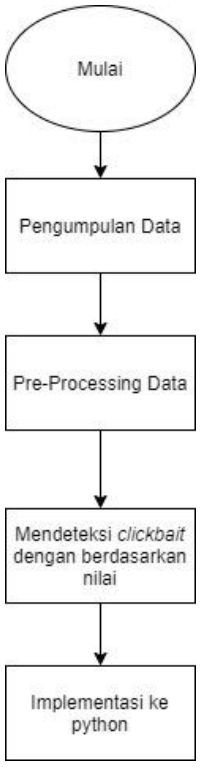
d. Panjang Kalimat
Panjang tidaknya sebuah kalimat juga mempengaruhi nilai dari kalimat tersebut. Pada umumnya kalimat yang sangat panjang maupun kalimat yang sangat pendek tidak memiliki nilai. Karena kalimat yang panjang akan memiliki informasi yang kurang informatif dan yang pendek tidak memiliki informasi yang cukup.

e. Pengurangan Kalimat
Tahap terakhir dari fase sentence *scoring* adalah menghilangkan frasa dan klausa yang kurang relevant dengan tema artikelnya. Pengambilan keputusan dalam mengurangi kalimat merupakan hasil dari *syntactic knowledge*, konteks dan corpus analisis.

*3.* Peringkat Kalimat
Pada fase ini kalimat yang telah diberikan nilai dari fase sebelumnya akan diurutkan berdasarkan nilai tertinggi[10].

Berdasarkan penelitian diatas menunjukkan bahwa artikel bersifat *clickbait* membawa dampak negatif terhadap pembaca. Untuk mengatasi hal tersebut penulis ingin membuat "Deteksi Clickbait dengan *sentence scoring based on frequency* di Detik.com".

## II. METODE PENELITIAN




Gambar 1. Metode Penelitian

Penelitian ini dirancang dengan menggunakan pendekatan kualitatif yang mana data uji dan data training diambil dari situs berita detik.com. Hal ini bertujuan untuk mengetahui seberapa banyak artikel yang mengandung sifat *clickbait* dibandingkan dengan yang tidak. Data yang telah diambil nanti akan diuji menggunakan metode Sentence Scoring Based on

Frequency. Berikut merupakan metode penelitiannya seperti pada gambar 1:

1. Mengumpulkan Data

   Penelitian ini dilakukan dengan mengumpulkan data yang ada pada penelitian perpustakaan (*library research)* serta penelitian di lapangan (*field research*).

2. Pre-Processing Data

   Data yang telah masuk akan di pra-proses terlebih dahulu. Pra-proses ini merupakan tahap awal agar data dapat digunakan. Langkah-langkahnya mencakupi :

   a. Segmentasi

   b. Tokenisasi

   c. Penghapusan Stop word

3. Mendeteksi *clickbait* dengan mencocokan judul berdasarkan hasil nilai

   Setelah dilakukan pra-proses data, maka selanjutnya akan diberikan nilai dengan menggunakan metode *sentence scoring based on frequency* dengan berdasarkan kesesuaian judul dan konten artikel.

4. Implementasi ke Python

   Pada tahap ini akan dilakukan implementasi secara langsung ke dalam bentuk pemrograman menggunakan Bahasa python. Dengan memasukkan data yang telah ada seperti artikel yang berasal dari detik.com dan juga algoritma sentence scoring based on frequency.

## III. HASIL DAN PEMBAHASAN

Dari dataset yang sudah diambil dari Sekolah Menengah Atas Kolese De Britto Yogyakarta meliputi 30 data yang diambil secara random dengan atribut yang digunakan adalah Pendapatang Orangtua, Tanggungan Keluarga, Pendidikan Orangtua, Umur Orang tua dapat dilihat pada Tabel.1 berikut :

Tabel 2. Dataset dari detik.com

| No | Judul | Clickbait |
|---|---|---|
| 1 | Bos tiktok sebut facebook tukang tiru | Tidak |
| 2 | Catatan fenomena Embun es di Dieng, berapa kali terjadi | Tidak |
| 3 | Heboh predator kain jarik ini 5 pakaian yang kerap jadi objek fetish | Clickbait |
| 4 | Kevin Systrom sempat takut Zuckerberg hancurkan Instagram | Clickbait |
| 5 | Kocak foto ekspektasi vs realita beli barang online di took | Tidak |
| 6 | Pegawai lab di wuhan dites corona massal ini hasilnya | Clickbait |
| 7 | Penjualan iphone meroket 255 di China tapi… | Clickbait |
| 8 | Saat orang berharta ribuan triliun dicecar kongres AS | Tidak |
| 9 | Lapor polisi, ahok merasa akun @itokurnia sudah keterlaluan | Tidak |
| 10 | Sudah pakai apd kok masih bisa terinfeksi corona? Ini kata dokter | Clickbait |

Dataset yang telah dikumpulkan lalu akan diexport ke dalam excel yang nantinya akan diubah ke bentuk CSV. File CSV tersebut lalu akan digunakan sebagai file data di pemrograman python dengan *framework* Spyder(anaconda3) untuk nantinya akan dieksekusi untuk mendapatkan nilai hasil pencocokan judul dan isi artikel.

Versi python yang digunakan oleh penulis adalah python 3.7 dan *import library* yang digunakan dalam penelitian ini ialah pandas dan nltk. Prosesnya dapat dilihat pada gambar 2.


Gambar 2. *Import library*

Setelah melakukan *import library*, maka selanjutnya akan dilakukan proses tokenisasi atau pemecahan kata perkata dari kalimat utuh. Berikut merupakan proses tokenisasi pada gambar 3.


Gambar 3. Tokenisasi dan Stopword

Tokenisasi berguna untuk mempermudah system untuk mengenali kata-kata dan memberikan nilai terhadap kata tersebut. Pada gambar 3 juga terdapat proses penghapusan stop word berbahasa Indonesia, penghapusan ini harus dilakukan agar kata-kata yang kurang penting dihilangkan dalam system sehingga

dapat mempermudah dalam proses pemberian nilai dan pencocokan judul dan isi.

Setelah dilakukan proses tokenisasi dan penghapusan stop word maka selanjutnya adalah mencari hasil nilai dengan mencari kecocokan judul dan nilai. Berikut adalah hasil nilainya dapat dilihat pada tabel 2.

Tabel 2. Hasil nilai kecocokan judul dan konten

| No | Judul | Nilai |
|---|---|---|
| 1 | Bos tiktok sebut facebook tukang tiru | 1 |
| 2 | Catatan fenomena Embun es di Dieng, berapa kali terjadi | 0 |
| 3 | Heboh predator kain jarik ini 5 pakaian yang kerap jadi objek fetish | 1 |
| 4 | Kevin Systrom sempat takut Zuckerberg hancurkan Instagram | 0 |
| 5 | Kocak foto ekspektasi vs realita beli barang online di took | 0 |
| 6 | Pegawai lab di wuhan dites corona massal ini hasilnya | 0 |
| 7 | Penjualan iphone meroket 255 di China tapi… | 0 |
| 8 | Saat orang berharta ribuan triliun dicecar kongres AS | 0 |
| 9 | Lapor polisi, ahok merasa akun @itokurnia sudah keterlaluan | 1 |
| 10 | Sudah pakai apd kok masih bisa terinfeksi corona? Ini kata dokter | 0 |

Keterangan :
Nilai 0 : Clickbait
Nilai 1 : Tidak Clickbait
Pada tabel 2 dapat dillihat hasil dari sistem yang dijalankan menggunakan program bahasa python. Setelah hasilnya keluar maka akan dibandingkan dengan dataset yang telah ada sebelumnya seperti pada tabel 3.

Tabel 3. Membandingkan dataset dengan data hasil uji coba

| No | Judul | Dataset | Uji coba sistem |
|---|---|---|---|
| 1 | Bos tiktok sebut facebook tukang tiru | Tidak Clickbait | Tidak Clickbait |
| 2 | Catatan fenomena Embun es di Dieng, berapa kali terjadi | Tidak Clickbait | Clickbait |
| 3 | Heboh predator kain jarik ini 5 pakaian yang kerap jadi objek fetish | Clickbait | Tidak Clickbait |
| 4 | Kevin Systrom sempat takut Zuckerberg hancurkan Instagram | Clickbait | Clickbait |
| 5 | Kocak foto ekspektasi vs realita beli barang online di took | Tidak Clickbait | Clickbait |
| 6 | Pegawai lab di wuhan dites corona massal ini hasilnya | Clickbait | Clickbait |
| 7 | Penjualan iphone meroket 255 di China tapi… | Clickbait | Clickbait |
| 8 | Saat orang berharta ribuan triliun dicecar kongres AS | Tidak Clickbait | Clickbait |
| 9 | Lapor polisi, ahok merasa akun @itokurnia sudah keterlaluan | Tidak Clickbait | Tidak Clickbait |
| 10 | Sudah pakai apd kok masih bisa terinfeksi corona? Ini kata dokter | Clickbait | Clickbait |

Setelah mendapatkan hasil uji coba maka tingkat akurasinya akan dihitung dengan confusion matrix seperti pada tabel 4.

Tabel 4. Confusion Matrix

| Correct Classification | Classification | |
|---|---|---|
| | Positif | Negatif |
| Positif | 4 | 1 |
| Negatif | 3 | 2 |

Keterangan :

1. Klasifikasi positif dengan positif : 4 karena dataset positif dan hasil uji sistem positif
2. Klasifikasi negatif dengan positif : 3 karena dataset negatif dan hasil uji sistem positif
3. Klasifikasi positif dengan negatif : 1 karena dataset positif dan hasil uji sistem negative
4. Klasifikasi negatif dengan negatif : 2 karena dataset negatif dan hasil uji sistem negative
5. Hasil pengujian adalah sebagai berikut :
   Akurasi = 4+2/(4+3+1+2)*100% = 75%
   Presisi = 4/(4+3)*100% = 57%
   Recall = 4/(4+1)*100% = 80%

## IV. KESIMPULAN

Berdasarkan penelitian, implementasi dan pengujian, maka dapat diambil kesimpulan sebagai berikut :

1. Penerapan algoritma sentence scoring based on frequency dalam mendeteksi clickbait dapat dilakukan.
2. Tingkat akurasi yang diperoleh dari membandingkan dataset dan data uji coba dari sistem adalah 75% untuk akurasi, 57% untuk presisi dan 80% untuk recall.

Saran yang dapat diberikan untuk penelitian ini ialah :

1. Dapat dikombinasikan dengan metode lain untuk meningkatkan akurasi seperti metode naïve bayes atau k-means.
2. Mengambil sampel dataset lebih dari satu website berita.
3. Mencoba melakukan penelitian pada implementasi di python guna mengoptimalkan hasil dari penelitian.

## DAFTAR PUSTAKA


[1] R., Golman, G., and Loeweinstein, "An Information-Gap Theory of Feelings about Uncertainty" *Department of Social and Decision Sciences, Carnegie Mellon University*, PA 15213.

[2] Y., D., Hidayat, "Clickbait di media online Indonesia" *Jurnal Pekommas*, vol. 4, no.1, pp 1-10, 2019.

[3] M., R., Kartanegara, "Penggunaan clickbait headline pada situs berita dan gaya hidup muslim dream.co.id" *Mediator: Jurnal Komunikasi*, vol.11, no.1, pp 31-43, 2018.

[4] G., E., Saputro and T., Haryadi, "Komik Strip dan Fenomena Clickbait" *Jurnal Titik Imaji, vol. 2, no.1, 2019*.

[5] N. Hurst, "To clickbait or not to clickbait? an examination of clickbait headline effects on source credibility" *University of Missouri-Columbia*, 2016.

[6] I., Belestin, B., R., Njegovan and M., S., Vukadinovic, "Clickbait titles: Risky formula for attracting readers and advertisers"*XVII International Scientific Conference on Industrial Systems*, 2017.

[7] A., Chakraborty, B., Paranjape, S., Kakarla and N., Ganguly, "Stop Clickbait : Detecting and preventing clickbaits in online news media" *Department of computer science and engineering*, 2016.

[8] T., S., R., Raju and B., Allarpu, "Text Summarization using sentence scoring method" *IRJET*, vol.4, no.4, 2017.

[9] R., Ferreira and L., D., S., Cabral, "Assessing sentence scoring techniques for extractive text summarization" *Science Direct*, 2013.

[10] T., S., Bell, and R., H., Wilson, "Sentence recognition materials based on frequency of word use and lexical Confusability"*J am acad audiol*, Vol.12, pp 514-522, 2001.