**Penerapan Algoritma** **Stochastic Gradient Descent untuk Klasifikasi Kategori Berita Berdasarkan Judul (Studi Kasus: Liputan6)**

*Brielt Bella Gracitwo1; Ryan Putranda Kristianto2*

1,2FakultasTeknik, Ilmu Informatika, Universitas Katolik Darma Cendika, Surabaya, Indonesia

brielt.bella@student.ukdc.ac.id

***ABSTRACT***

*Liputan6 Is a news site that has been registered and a media place for the Indonesian Press Council. Liputan6 generates news articles every hour and the data in real-time. The need to determine the type of news automatically to speed up the work of journalists in making news articles. The wclassification algorithm is used to determine the type of news automatically. The results of this study research to obtain accuracy obtained after knowing the type of news automatically in journalism, the accuracy value obtained by researchers is 65%.*

***Keywords:*** *Liputan6, News, Classification, Stochastic Gradient Descent*

***ABSTRAK***

Liputan6 Merupakan tempat berita yang sudah terdaftar dan tempat media dewan pers indonesia. Setiap jam Liputan6 menghasilkan artikel berita dan data tersebut bersifat *real time.* Perlunya penentuan jenis berita secara otomatis untuk mempercepat pekerjaan jurnalis dalam pembuatan artikel berita. Digunakan algoritmaklasifikasi *Stochastic Gradient Descent*dalam menentukan jenis berita secara otomatis. Hasil dari penelitian ini adalah penelitian mendapatkan akurasi yang didapat setelah mengetahui jenis berita secara otomatis pada jurnalistik, nilai akurasi yang didapat oleh peneliti sebesar 65%

***Kata kunci:*** *Liputan6, Berita, Klasifikasi, Stochastic Gradient Descent*

**1. PENDAHULUAN**

Berita merupakan sebuah informasi atas kejadian terkini tertentu, sehingga masyarakat bisa mengetahui akan peristiwa tersebut. Pada era sekarang berita sangat mudah didapatkan karena adanya teknologi internet. Terdapat survei yang dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) terdapat 210.026.769 dari 272.682.600 penduduk Indonesia pengguna internet [1]. Dari hasil survei tersebut bahwasanya hamper semua penduduk Indonesia cenderung membaca berita melalui daring. Banyak sekali portal berita daring yang ada pada saat ini salah satunya Liputan6.  Liputan6 merupakan perusahan berita yang bergerak pada portal berita dan bekerjasama dengan stasiun televisi Surya Citra Televisi (SCTV) dalam menyiarkan berita[2]. Namun, terkadang  ada beberapa portal berita yang hanya mencantumkan beritanya saja tanpa memperhatikan suatu hal penting bagi pengguna yaitu penggolongan jenis berita dan berita tersebut menggunakan data *real time,* sehingga hal tersebut membuat pengguna cenderung bingung untuk memilih jenis berita yang ingin  dibaca karena terlihat acak.  Oleh karena itu penelitian ini bertujuan untuk mengetahui judul yang tercantum pada portal berita tersebut termasuk ke dalam jenis berita apa dengan cara melakukan penggolongan dan penerapannya menggunakan klasifikasi pada algoritma *Stochastic Gradient Descent*(SGD).

Algoritma SGD Merupakan salah satu algoritma pembelajaran mesin yang bersifat non parametrik pemodelan berbasis machine learning yang digunakan untuk mengetahui klasifikasi terkait penggolongan berita. pada metode SGD klasifikasi terhadap objek yang menggunakan jenis data berita. Metode tersebut sangatlah cocok digunakan bagi peneliti, karena metode itu untuk melakukan penggolongan berita sangatlah akurat.

Beberapa peneliti telah melakukan riset atau penelitian sebelumnya diantarnya peneliti menggunakan metode klasifikasi. Seperti yang dilakukan oleh peneliti (Kurniawan et al., 2017) dengan judul “Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes”, Dengan hasil penelitian, peneliti mendapatkan hasil nilai precision 0.962961, recall 0.789164[3]. Peneliti selanjutnya dilakukan oleh peneliti (Ulfah and Anam 2020) dengan judul “Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM)”[4], peneliti mendapatkan hasil penelitan dengan nilai akurasi sebesar 53,88% berdasarkan 105 komentar. Penelitan selanjutanya dilakukan oleh peneliti (Komalasari, 2018) dengan judul “Klasifikasi Berita Menggunakan Metode K-Nearest Neighbor ”[5]. Peneliti memperoleh hasil dari penelitian dengan nilai akurasi 87%. Peneliti dapat menyimpulkan bawasanya algoritma klasifikasi KNN sangatlah baik digunakan, tipe data yang digunakan dalam penelitian ini data bersifat *real-time*.

Berdasarkan penelitian ini, peneliti melakukan eksperimen dengan menentukan jenis berita. Pada penelitian ini peneliti menggunakan data *real-time*, dengan menerapkan algoritma klasifikasi *Stochastic Gradient Descent* yaitu untuk menetukan jenis berita secara otomatis. Dengan hasil model klasifikasi penentuan jenis berita yang dilakukan oleh sistem secara otomatis, hal iwni dapat membantu kinerja jurnalis Liputan6 dalam meproduksi artikel berita dengan cepat dan tepat tanpa menentukan jenis berita secara manual.

**2. METODE/PERANCANGAN PENELITIAN**

Metode Penelitian merupakan cara yang dilakukan oleh peneliti untuk mencapai tujuan tertentu, dengan cara memecahkan masalah serta menemukan penyelesaian dengan menggunakan prosedur atau peneliti terdahulu. Berdasarkan pemaparan diatas, maka peneliti melakukan perhitungan dengan menggunakan kalasifikasi SGD dalam mengklasifikasi jenis berita yang dilakukan secara otomatis. Dengan menggunakan beberapa metode yaitu : (1) Collecting Data; (2) Pengolahan data; (3) Feature Extraction;(4) Data Splitting Training dan Testing; (5) Hasil Analisa*;*

**Tabel 1.** Alur Penelitian

Collecting Data

Text Preprocessing

Data Splitting Training dan Testing

Feature Extraction

Hasil Analisa

**2.1 Collecting Data**

Collecting Data merupakan awal dari penelitian yang data tersebut. Data ini digunakan untuk mengklasifi jenis berita. Pada penelitian ini menggunakan *software system* informasi yaitu *google anyltic*. Pada data ini Terdapat 800 data yang diperoleh dari *google anyltic* berupa file CSV.

**2.2 Text Preprocessing**

Setelah melakukan pengambilan data, tahapan berikutnya adalah melakukan Text Preprocessing tersebut terdapat judul berita, jenis berita, dan jumlah penggunjung. Pada tahap ini data yang sudah didapat, data tersebut dilakukan Text Preprocessing. Data yang setelah dilakukan penglolaan terdapat data judul berita dan jenis berita. Sehingga data tersebut dapat dilakukan penyisian pada pengunjung.

**2.3 *TF-IDF***

*Terms Frequency-Inverse Document* Frequency atau disingkat TF-IDF merupakan metode yang melakukan untuk menentukan perhitungan nilai dan memberikan bobot pada term didalam dokumen[6] . Metode ini menggabungkan 2 perhitungan yaitu : (1)TF ini akan melakukan perhitungan; (2)IDF melakukan setiap dokumen. Proses Algoritma TF-IDF yang digunakan pada penelitian ini yaitu : (1)Menghitung Nilai TF; (2) Menghitung Nilai IDF; (3) Menghitung tolak ukur TF-IDF.

Tabel 1 : Alur Perhitungan TF-IDF

Menghitung nilai TF

Menghitung nilai IDF

Menghitung tolak ukur TF-IDF

berikut perhitungan rumus TF-IDF.persamaan yang digunakan untuk menghitung nilai Term frequency pada rumus 1 sebagai berikut.

Keterangan Rumus :

TF(d,t) = F(d,t) (1)

F(d,t) = Kemunculan kata t di dalam dokumen d

Setelah dilakukan perhitungan TF, lakukan perhitungan IDF pada rumus 2 sebagai berikut.

Keterangan Rumus :

IDF(t) = log(N/df(t)) (2)

n = total dokumen latih

df(t) = dokumen frecuency dari kata t

*Terms Frequency-Inverse Document* Frequency menjadi tolak ukur untuk melakukan perhitungan statistic dalam proses analisia dokumen hal tersebut pada rumus 3 sebagai berikut.

Keterangan Rumus :

wij = tfij x idij (3)

wij = bobot term terhadap dokumen

tfij = jumlah kemunculan kata (*term*) dalam setiap dokumen

idij = jumlah kemunculan kata (*term*) dalam sebuah dokumen

**2.4 *Stochastic Gradient Descent***

Algoritma *Stochastic Gradient Descent* (SGD) merupakan algoritma yang digunakan untuk menemukan nilai minimum lokal dari sebuah fungsi pembelajaran. SGD ini dapat mengetahui kesalahan estimasi yang dikoreksi dengan iterasi estimasi menggunakan aturan gradien dari fungsi yang akan diminalkan. Proses Algoritma *Stochastic Gradient Descent* yaitu melakukan prediksi untuk tahap pertama, kemudian jikalau ada kesalah dalam memprediksi akan di perbaiki hingga literasi berhenti dengan aturan gradien [7]. Berikut rumus SGD.

Keterangan Rumus :

(4)

= bobot

= learning rate

= gradient sescent

**2.5 Hasil Analisa**

hasil Analisa yang digunakan pada penelitian ini untuk menentukan jenis berita di liputan 6 dengan membuat pemodelan klasifikasi jenis berita yang menggunakan metode *Stochastic Gradient Descent* , sehingga metode yang digunakan peneliti tersebut dapat membantu jurnalis liputan 6 dalam memproduksi artikel jenis berita dengan cepat dan tidak menentukan jenis berita secara manual.

**3. HASIL DAN PEMBAHASAN**

**3.1 Collecting Data**

Penelitan ini mengambil data realtime dengan menggunakan pihak ketiga yaitu *Google Analytics*. Data yang diambil berupa file dengan format CSV, didalamnya terdapat data judul berita, jenis berita, dan jumlah pengunjung *website*. Jumlah keselurahn data sebesar 2287, berikut peneliti mengambil sampel dengan data 5 teratas.

Tabel 2 : Output read data csv

|  |  |  |
| --- | --- | --- |
| **Judul** | **Jenis Berita** | **User Aktif** |
| Wajah Istri Sultan Kelantan yang Pernah Bersengketa dengan Miss Moscow Terlihat untuk Pertama Kali | *Lifestyle* | 7 |
| Nonton Video Porno Bareng Pasangan Beri Dampak Positif atau Justru Negatif? | *Health* | 7 |
| Cuek dengan Isu Perselingkuhan Behati Prinsloo Tertangkap Kamera Temani Adam Levine Manggung | *Health* | 5 |
| *Skrining Hormon Tiroid* pada Bayi Baru Lahir Bisa Cegah Gangguan Tumbuh Kembang Anak | *Health* | 5 |
| Wajah Istri Sultan Kelantan yang Pernah Bersengketa dengan Miss Moscow Terlihat untuk Pertama Kali | *Lifestyle* | 5 |

**3.2 Text Preprocessing**

1. *Data Prosesing*

Setelah dilakukan Collecting Data, data akan dilakukan uji coba untuk pemilihan. Data yang dibuthkan dalam kalasifikasi jenis berita menggunakan TF-IDF. Tahapan pertama pengelolahan data yaitu melakukan pembersihan data yang duplikasi berdasarkan judul berita, sehingga didapatakan data yang unik. Hal yang penting dalam penelitian ini untuk klasifikasi jenis berita menggunkan 2 kolom yaitu Judul Berita dan Jenis Berita. Berikut sampel 5 data teratas yang digunakan dalam klasifikasi jenis berita

Tabel 3 : output pemilihan data dan menghapus

|  |  |
| --- | --- |
| **Jenis** | **Jenis Berita** |
| Wajah Istri Sultan Kelantan yang Pernah Bersengketa dengan Miss Moscow Terlihat untuk Pertama Kali | *Lifestyle* |
| Nonton Video Porno Bareng Pasangan Beri Dampak Positif atau Justru Negatif? | *Health* |
| Cuek dengan Isu Perselingkuhan Behati Prinsloo Tertangkap Kamera Temani Adam Levine Manggung | *Health* |
| Skrining Hormon Tiroid pada Bayi Baru Lahir Bisa Cegah Gangguan Tumbuh Kembang Anak | *Health* |
| Wajah Istri Sultan Kelantan yang Pernah Bersengketa dengan Miss Moscow Terlihat untuk Pertama Kali | *Lifestyle* |

Setelah dilakukan pemilihan data, data tersebut dikelola untuk menghilangkan judul yang sama.

Tabel 4 : Output data untuk menghilangkan judul yang sama

|  |  |  |
| --- | --- | --- |
|  | Judul | Jenis |
| Count | 800 | 800 |
| Unique | 800 | 4 |
| Top | Wajah Istri Sultan Kelantan yang Pernah Bersen | Lifestyle |
| Freq | 1 | 200 |

1. Case Folding

Selanjutnya mengubah huruf kapital menjadi huruf kecil dan menghilangkan tanda baca agar dapat diproses lebih lanjut.

Tabel 5 : Output mengubah judul dan jenis menjadi huruf kecil

|  |  |  |
| --- | --- | --- |
|  | **Judul** | **Jenis** |
| 0 | wajah istri sultan kelantan yang pernah bersen... | *lifestyle* |
| 1 | nonton video porno bareng pasangan beri dampak... | *health* |
| 2 | cuek dengan isu perselingkuhan behati prinsloo... | *health* |
| 3 | skrining hormon tiroid pada bayi baru lahir bi... | *health* |
| 4 | dibayangi resesi harga emas bakal melambung di... | *bisnis* |

1. Tokenizing dan Stemming

Proses selanjutnya melakukan *tokenizing* dengan memisahakan kata satu dengan yang lain. Kata yang sudah di pisah akan dilakukan proses *stemming* agar mendapatkan kata baku sebagai parameter untuk melakukan TF-IDF.

Tabel 6 : Output tokenizing dan stemming

|  |  |  |
| --- | --- | --- |
|  | **Judul** | **Jenis** |
| 0 | wajah istri sultan lantan pernah sengketa miss.. | *lifestyle* |
| 1 | nonton video porno bareng pasang beri dampak p... | *health* |
| 2 | cuek isu selingkuh behati prinsloo tangkap kam.. | *health* |
| 3 | skrining hormon tiroid bayi baru lahir cegah g... | *health* |
| 4 | bayang resesi harga emas bakal lambung | *bisnis* |

**3.3 *TF-IDF***

Setelah melakuan pengelolahan data peneliti selanjutnya melakukan *splitting* data untuk mengacak data dan membagi kedalam 2 segment. 2 segment tersebut untuk menentukan data *training* dan *test*. Langkah berikutnya peneliti memberikan nilai TF-IDF pada semua judul berita agar dapat meningkatakan nilai akurasi dari SGD.

**3.4 *Stochastic Gradient Descent***

Tahap berikutnya setelah melakukan TF-IDF peneliti yaitu imlementasi SGD klasifikasi. Adapun parameter yang dibutuhkan yaitu data tfidf\_train dan pada splitting menggunakan y\_train. Kemudian dimasukan kedalam model *stochastic gradient descent* dengan cara seperti tabel 13 berikut.

Kemudian dilakukan prediksi dengan data latih tf-idf\_test kedalam model SGD klasifikasi. Dengan cara seperti tabel 14 berikut.

Tabel 7 : Output prediksi data test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0** | 0.64 | 0.80 | 0.71 | 20 |
| **1** | 0.62 | 0.75 | 0.68 | 20 |
| **2** | 0.59 | 0.50 | 0.54 | 20 |
| **3** | 0.79 | 0.55 | 0.65 | 20 |
| ***Accuracy*** | - | - | 0.65 | 80 |
| ***Macro Avg*** | 0.66 | 0.65 | 0.65 | 80 |
| ***Weighted Avg*** | 0.66 | 0.65 | 0.65 | 80 |

**3.5 Hasil Analisa**

Berdasarkan dari hasil prediksi SGD pada klasifikasi jenis berita liputan 6 mendapatkan hasil dengan nilai akurasi 65%. Nilai tersebut yang didapatkan dari hasil klasifikasi kurang memenuhi dalam metode SGD. Hasil yang memenuhi dalam metode SGD yaitu memiliki nilai akurasi diatas 80%. Hal tersebut dalam metode ini tidak dapat memproduksi artikel jenis berita dengan cepat. dan tidak menentukan jenis berita secara manual.

**4. KESIMPULAN DAN SARAN** [11 pts/Bold]

Berdasarkan hasil uji coba klasifikasi jenis berita pada liputan 6 menggunakan metode SGD dengan mendapatkan nilai akurasi 65%. Hal ini dengan menggunakan metode SGD mendapatkan nilai akurasi yang kurang akurat sehingga untuk menentukan jenis berita secara otomatis tidak dapat diimplementasikan. Hal ini peneliti memiliki sara untuk peneliti selanjutnya untuk melakukan optimasi dengan menggunakan metode yang berbeda dalam melakukan klasifikasi jenis berita pada liputan 6.

**UCAPAN TERIMAKASIH**

Terima kasih disampaikan kepada Liputan 6 atas telah mengizinkan untuk dilakukan penelitian ini, serta terimakasih juga kepada Bapak Simon Handianto Alim yang telah membantu pengambilan data liputan 6, serta kepada Ryan Putrandra Kristanto terimakasih sudah menjadi dosen pendamping dan mmebantu pelaksanaan penelitian ini.

**DAFTAR PUSTAKA**

Kutipan berturut-turut dalam tanda kurung [1]. Kalimat tanda baca berikut braket [2]. Merujuk hanya untuk nomor referensi, seperti pada [3] -Jangan menggunakan "Ref. [3] "atau" referensi [3]. Minimal daftar pustaka sebanyak 15 Kutipan.

0[1] MUHAMMAD ARIF, “Profil Internet Indonesia 2022,” *AJII*, 2022. https://apjii.or.id/content/read/39/559/Laporan-Survei-Profil-Internet-Indonesia-2022 (accessed Nov. 01, 2022).

[2] “Tentang Kami,” *Liputan6*, 2022. https://www.liputan6.com/info/tentang-kami (accessed Nov. 01, 2022).

[3] B. Kurniawan, M. A. Fauzi, and A. W. Widodo, “Klasifikasi Berita Twitter Menggunakan Metode Improved Naïve Bayes,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 10, pp. 1193–1200, 2017.

[4] A. N. Ulfah and M. K. Anam, “Analisis Sentimen Hate Speech Pada Portal Berita Online Menggunakan Support Vector Machine (SVM),” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 7, no. 1, pp. 1–10, 2020, doi: 10.35957/jatisi.v7i1.196.

[5] B. P. Komalasari, “Klasifikasi Berita Menggunakan Metode KNearest Neighbor (K-NN),” vol. 5, no. 2, pp. 259–268, 2018, [Online]. Available: https://perpustakaan.ft.unram.ac.id/index.php?p=show\_detail&id=7862

[6] A. B. Prasetijo *et al.*, “Hoax Detection System with SVM and SGD-2017.pdf,” pp. 45–49, 2017.

[7] S. M. Tambunan, Y. Nataliani, and E. S. Lestari, “Perbandingan Klasifikasi dengan Pendekatan Pembelajaran Mesin untuk Mengidentifikasi Tweet Hoaks di Media Sosial Twitter,” *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 2, p. 112, 2021, doi: 10.26418/jp.v7i2.47232.

**Referensi harus ditulis dalam gaya IEEE**.. Penyusunan rujukan dalam daftar pustaka berurut berdasarkan urutan sitasi yang digunakan (sekuensial) dan diberi nomor angka arab dalam kurung siku. Penulisan unsur-unsur keterangan pustaka mengikuti kaidah dengan urutan: (1) nama pengarang ditulis dengan urutan nama akhir, nama awal dan nama tengah, tanpa gelar akademik. (2) tahun penerbitan. (3) Judul. (4) tempat penerbitan. (5) nama penerbit.

**LAMPIRAN**

Jika diperlukan, tulisan dapat dilengkapi dengan lampiran.