**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

1[brielt.bella@student.ukdc.ac.id](mailto:brielt.bella@student.ukdc.ac.id)

*Corresponding Author*: 2ryan@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 algoritma SGD klasifikasi terhadap objek yang menggunakan jenis data berita. algoritma tersebut sangatlah cocok digunakan bagi peneliti, karena algoritma itu untuk melakukan penggolongan berita sangatlah akurat.

Beberapa peneliti telah melakukan riset atau penelitian sebelumnya diantaranya peneliti menggunakan algoritma klasifikasi, yaitu algoritma *Support Vector Machine* yang digunakan untuk Sentimen *Hate Speech* Pada Portal Berita Online dengan nilai akurasi sebesar 53,88%[3], kemudian penelitian selanjutnya menggunakan algoritma *K-Nearest Neighbor* dalam klasifikasi teks berita dengan nilai akurasi sebesar 87%[4]. Terdapat penelitian serupa tentang mengetahui berita hoax dengan menggunakan algoritma *Naïve Bayes* dengan nilai akurasi sebesar 91,82%[5], kemudian penelitian mengenai pengkategorian berita menggunakan algoritma *Random Forest Multi-Label*[6]. Penelitian selanjutnya menggunakan *deep learning Multi-layer Perceptron* dan *Term Frequency-Invers Document Frequency* dalam sentimen komentar terhadap calon presiden 2019 dengan nilai akurasi sebesar 88%[7], terdapat penelitian mengenai pendeteksi spam SMS menggunakan algoritma *Gradient Descent Classifier* dengan nilai akurasi sebesar 97%[8]. Dari beberapa penelitian terdahulu dapat disimpulkan bahwasanya algoritma *Gradient Descent Classifier* klasifikasi sangatlah baik digunakan terhadap analisa teks judul berita dengan 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 menentukan jenis berita secara otomatis. Dengan hasil model klasifikasi penentuan jenis berita yang dilakukan oleh sistem secara otomatis, hal ini dapat membantu kinerja jurnalis Liputan 6 dalam memproduksi 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 klasifikasi SGD dalam mengklasifikasi jenis berita yang dilakukan secara otomatis. Dengan menggunakan beberapa tahapan yaitu : (1) Collecting Data; (2) Pengolahan data; (3) Feature Extraction;(4) Data Splitting Training dan Testing; (5) Penerapan Algoritma SGD; (6) Hasil Analisa;

Collecting Data

Text Preprocessing

Data Splitting Training dan Testing

Feature Extraction

Hasil Analisa

Penerapan Algoritma SGD

**Gambar 1.** Alur Penelitian

**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 analytics*. Pada data ini Terdapat 800 data yang diperoleh dari *google analytics* 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 pengunjung. Pada tahap ini data yang sudah didapat, data tersebut dilakukan Text Preprocessing. Data yang telah dilakukan pengelolaan terdapat data judul berita dan jenis berita. Sehingga data tersebut dapat dilakukan penyisihan pada pengunjung. Teknik yang digunakan dalam penelitian ini yaitu (1) Case Folding; (2) Tokenizing; (3) Stemming;

Case Folding

Tokenizing

Stemming

**Gambar 2**. alur Text Preprocessing

**2.3 *Feature extraction***

*Feature extraction* merupakan mengidentifikasi satu cara untuk mengenali suatu objek dengan melihat ciri-ciri khusus yang dimiliki objek tersebut. Tujuan dari *feature extraction* ini yaitu melakukan perhitungan terhadap jumlah frekuensi kata yang muncul dari setiap *record* data.  Teknik yang digunakan adalah TF-IDF. TF-IDF merupakan metode yang melakukan untuk menentukan perhitungan nilai dan memberikan bobot pada term didalam dokumen. 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 tolok ukur TFIDF.

**Gambar 3.** Alur Perhitungan TF-IDF

Menghitung nilai TF

Menghitung nilai IDF

Menghitung tolak ukur TFIDF

**2.4 Data Splitting Training dan Testing**

Data Splitting biasanya digunakan data yang terdiri menjadi 2 bagian untuk membentuk sebuah subhimpunan data[9]. Data yang harus dilakukan pertama yaitu mengevaluasi uji data dan lainnya dengan menggunakan model. Metode ini menggunakan data yang akan menunjang penelitian ini yaitu (1) Training; (2) Testing;

**2.5 Penerapan Algoritma *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 jika terdapat kesalahan dalam memprediksi akan diperbaiki hingga literasi berhenti dengan aturan gradien [10]. Berikut Formula SGD.

Keterangan Formula :

(1)

Keterangan Formula :

= bobot

= learning rate

= gradient descent

**3. HASIL DAN PEMBAHASAN**

**3.1 Collecting Data**

Penelitian ini melakukan *collecting data* dengan data *realtime*. Data tersebut digunakan untuk pengambilan data yaitu *Google Analytics*. Data yang diambil berupa file dengan format CSV, didalamnya terdapat data judul berita, jenis berita, dan jumlah pengunjung *website*. Jumlah keselurahan data sebesar 800, berikut peneliti mengambil sampel dengan data 5 teratas.

**Tabel 1**. 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**

Setelah dilakukan Collecting Data, data akan dilakukan uji coba untuk pemilihan. Data yang dibutuhkan dalam klasifikasi jenis berita. Tahapan pertama *Text Preprocessing* yaitu melakukan pembersihan data yang duplikasi berdasarkan judul berita, sehingga didapatkan data yang unik. Hal yang penting dalam penelitian ini untuk klasifikasi jenis berita menggunakan 2 kolom yaitu Judul Berita dan Jenis Berita. Berikut sampel 5 data teratas yang setelah dikelola dan digunakan dalam klasifikasi jenis berita.

**Tabel 2.** 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 3.** 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 |

Selanjutnya mengubah huruf kapital menjadi huruf kecil dan menghilangkan tanda baca agar dapat diproses lebih lanjut, teknik ini dinamakan Case Folding.

**Tabel 4.** 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* |

Proses selanjutnya melakukan *tokenizing* dengan memisahakan kata satu dengan yang lain. Kata yang sudah dipisah akan dilakukan proses *stemming* agar mendapatkan kata baku sebagai parameter untuk melakukan pemilahan judul berita.

**Tabel 5.** 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 *Feature extraction***

Setelah melakuan *Text Processing* data peneliti selanjutnya melakukan *Feature Extraction* untuk mengubah judul menjadi huruf kecil atau *lower case*, dan menghilangkan tanda baca pada judul berita.

**3.4 Data *Splitting Training* dan *Testing***

Penelitian ini menggunakan dataset Liputan 6, penulis membagi dataset tersebut kedalam 2 bagian yaitu dataset training dan testing, dimana dataset training sebesar 90% dan dataset testing sebesar 10%.

**3.5**  ***Penerapan Algoritma* *Stochastic Gradient Descent***

Tahap berikutnya data tersebut menggunakan algoritma *Stochastic Gradient Descent* untuk implementasi klasifikasi pada jenis berita. data tersebut dikelola dan mendapatkan hasil pada tabel 6.

**Tabel 6.** 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 |

**3.6 Hasil Analisa**

Berdasarkan dari hasil prediksi SGD pada klasifikasi jenis berita liputan 6 mendapatkan hasil 65%.hasil tersebut memiliki nilai akurasi yang baik dan nilai tersebut ada pada tabel 6. Berdasarkan pengujian diatas maka algoritma SGD dapat memprediksi jenis kategori artikel berita.

**4. KESIMPULAN DAN SARAN**

Berdasarkan hasil uji coba klasifikasi jenis berita pada liputan 6 menggunakan algoritma SGD dengan mendapatkan nilai akurasi 65%. Nilai akurasi tersebut dapat dikatakan baik. Hasil ini digunakan untuk klasifikasi jenis berita liputan 6 secara otomatis. Pemetaan pada klasifikasi ini dapat diimplementasikan dengan memproduksi jenis berita dengan cepat dan menentukan jenis berita secara otomatis. Penulis memiliki saran untuk penulis selanjutnya untuk melakukan optimasi dengan menggunakan algoritma yang berbeda dalam melakukan klasifikasi jenis berita.

**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 Putranda Kristianto terimakasih sudah menjadi dosen pendamping dan membantu pelaksanaan penelitian ini.

**DAFTAR PUSTAKA**

[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] 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.

[4] B. P. Komalasari, “Klasifikasi Berita Menggunakan Metode KNearest Neighbor (K-NN),” vol. 5, no. 2, pp. 259–268, 2020, [Online]. Available: https://perpustakaan.ft.unram.ac.id/index.php?p=show\_detail&id=7862

[5] H. Muhabatin *et al.*, “Classification of Hoax News Using Naïve Bayes Algorithm Based on PSO,” *Informatics Educ. Prof.*, vol. 5, no. 2, pp. 156–165, 2021.

[6] B. H. Mahendra, Adiwijaya, and U. N. Wisesty, “Kategorisasi Berita Multi-Label Berbahasa Indonesia Menggunakan Algoritma Random Forest,” *e-Proceeding Eng.*, vol. 6, no. 2, pp. 9030–9041, 2019.

[7] F. Syadid, “Analisis Sentimen Komentar Netizen Terhadap Calon Presiden Indonesia 2019 Dari Twitter Menggunakan Algoritma Term Frequency-Invers Document Frequency (Tf- Idf) Dan Metode Multi Layer Perceptron (Mlp) Neural Network,” *Skripsi Univ. Islam Negeri Syarif Hidayatullah Jakarta*, p. 72, 2019.

[8] R. Dwiyansaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta, “Deteksi Sms Spam Berbahasa Indonesia Menggunakan Tf-Idf Dan Stochastic Gradient Descent Classifier ( Indonesian Sms Spam Detection Using Tf-Idf And Stochastic Gradient Descent,” *J. Teknol. Informasi, Komput. dan Apl.*, vol. 3, no. 2, pp. 200–207, 2021, [Online]. Available: https://jtika.if.unram.ac.id/index.php/JTIKA/article/view/145

[9] D. Z. Abidin, S. Nurmaini, R. Firsandava Malik, Erwin, E. Rasywir, and Y. Pratama, “RSSI Data Preparation for Machine Learning,” *Proc. - 2nd Int. Conf. Informatics, Multimedia, Cyber, Inf. Syst. ICIMCIS 2020*, pp. 284–289, 2020, doi: 10.1109/ICIMCIS51567.2020.9354273.

[10] 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.