IJSA (2025), Vol. 9, No. 1, 86-99
DOI: https://doi.org/10.29244/ijsa.v9i1p86-99

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Digital Newsworthiness Scores Model Using a Combination of Unsupervised and Supervised Learning Approaches*

Pemodelan Skor Kelayakan Berita Digital dengan Pendekatan Kombinasi *Unsupervised* dan *Supervised Learning*

Reza Felix Citra[1‡], Aji Hamim Wigena[1], and Bagus Sartono[1]

[1] Study Program on Statistics and Data Science, IPB University, Indonesia
‡corresponding author: rezafelix@gmail.com

Abstract

The rapid evolution of digital technology has transformed the media landscape, making news more accessible while also introducing challenges related to content quality and accuracy. The rise of misinformation and fake news has diminished public trust in traditional media. The research objective of this study is to develop a systematic method for evaluating the quality and potential impact of news articles before publication. By adapting credit risk scoring principles, a model was used to predict the suitability of news content based on factors such as title length, number of images, news category, and publication timing. A variable target was firstly formed using three clustering methods: K-Means, K-Modes, and K-Medoids. The results indicated that K-Means outperformed the other methods, leading us to use its outcomes for determining publication suitability. Subsequently, stepwise logistic regression was applied to implement the credit risk scoring approach, allowing for variable selection and assessment of importance. Ultimately, ten variables were identified to generate a newsworthiness score, with minimum and maximum scores of 997 and 1407, respectively. The average scores for articles deemed publishable and not publishable were 1137 and 1110. A cutoff score of 1123 was established based on these averages, categorizing 6708 articles (57.9%) as suitable for publication. These findings aim to assist media organizations in refining their content curation processes, thereby enhancing the overall quality of news consumption.

Keywords: accuracy, credit risk scoring, k-means, logistic regression, silhouette.

* Received: Dec 2024; Reviewed: Dec 2024; Published: Jun 2025

1. Pendahuluan

Perkembangan pesat teknologi digital telah merevolusi industri media, mengubah cara mengonsumsi berita. Berita digital kini menjadi sumber informasi utama yang mudah diakses oleh masyarakat. Namun, di balik kemudahan akses ini, muncul tantangan baru terkait kualitas, kedalaman informasi, akurasi, serta interaktifitas dari elemen-elemen lain yang mendukung berita. Proliferasi berita palsu, bias, dan hoaks juga telah mengikis kepercayaan publik terhadap media massa (Muqsith, 2021).

Dalam konteks ini, pengembangan sebuah model penilaian dapat membantu redaksi media dalam mengidentifikasi konten berita digital yang berkualitas dan relevan. Sayangnya, sampai saat ini belum ada model penilaian untuk konten berita yang dapat membantu penerbit menjaga kualitas terbitannya. Model penilaian yang ada saat ini adalah penilaian resiko kredit (*Credit Risk Scoring*) yang biasa digunakan dalam industri keuangan atau ekonomi dalam menilai kemungkinan seseorang mendapatkan kredit. Dengan mengadaptasi konsep *credit risk scoring* ini diharapkan akan diperoleh sebuah pendekatan inovatif untuk menilai kelayakan sebuah berita sebelum dipublikasikan.

Penggunaan sistem skor pada penilaian resiko kredit baru secara resmi digunakan pada tahun 1989 (Lauer, 2017). Saat itu Equifax (sebuah biro kredit skala besar di Amerika) dan FICO (sebuah perusahaan teknologi) berhasil menciptakan standar penilaian kredit setelah bekerja sama kurang lebih 20 tahun. *Credit risk scoring* telah terbukti efektif dalam memprediksi risiko kredit nasabah (Abdou & Pointon, 2011). Sejak itu penggunaan dan pengembangan skor kredit ini menjadi lebih masif dan luas.

Sayangnya, penggunaan sistem skor di luar bidang keuangan dan perbankan belum banyak terjadi. Berdasarkan penelitian Onay & Öztürk (2018) terhadap sekitar 258 literatur tentang skor kredit hanya ada 4 penelitian yang mencoba menerapkan metode ini dengan hal lain, yaitu: tentang hubungan penyakit kardiovaskular dan skor kredit (Israel et al., 2014), hubungan antara rumah hemat energi dan skor kredit (Sanderford et al., 2014), pengembangan kerangka kepatuhan untuk menilai risiko pencucian uang dan pendanaan terorisme (Sathye & Islam, 2011), dan pemodelan pengaturan kontainer dan pengacakan operasi (Gharehgozli & Zaerpour, 2018).

Begitu juga dalam hal penilaian kelayakan berita. Sampai saat ini, belum ada yang pernah mencoba menerapkan sistem skor ini. Dalam penelitian ini, kasus penilaian berita akan dianalogikan seperti penilaian calon nasabah pada kasus di bidang keuangan dan perbankan. Dalam hal ini, kegagalan mendeteksi berita yang kurang baik dapat menyebabkan pelanggan berita kecewa, yang jika terjadi terus menerus dapat menyebabkan pelanggan berhenti berlangganan.

Penerapan sistem skor di bidang jurnalistik diperlukan data dari setiap berita yang dihasilkan dan klasifikasi dari setiap berita tersebut. Tidak seperti dalam kasus kredit, klasifikasi penentuan berita yang layak terbit belum mempunyai standar. Selama ini penentuan berita dilakukan oleh seorang editor, sehingga masih bersifat subyektif karena tergantung dari tipe dan karakter editor. Editor yang berbeda kemungkinan besar mempunyai penilaian yang berbeda pada suatu berita yang sama.

Untuk menghindari efek subyektifitas, penentuan tersebut akan dilakukan dengan menggunakan salah satu metode dalam algoritma pembelajaran tak terawasi (*unsupervised learning*), yaitu metode pengelompokkan (*clustering*). Metode ini akan menghasilkan sebuah peubah target dengan 2 kategori, yaitu kategori berita yang

layak terbit dan yang belum. Data yang akan digunakan untuk pembentukan peubah target ini adalah jumlah total berapa kali berita diakses (*pageviews*), total waktu sebuah berita diakses (*Total Engaged Time*), dan rata-rata lama akses per berita per pembaca (*Average Engage page*).

Setelah pembentukan peubah target, baru penerapan metode *credit risk scoring* dapat dilakukan. Metode *credit risk scoring* termasuk dalam algoritma pembelajaran terawasi (*supervised learning*), sehingga diperlukan sebuah peubah target sebagai penentu keputusan. Dalam hal penilaian resiko kredit, peubah target adalah perilaku nasabah dalam membayar kredit, yaitu lancar atau tidak lancar. Pada penelitian ini, peubah target adalah berita yang layak diterbitkan dan yang belum layak.

Prinsip dasar metode ini adalah memberikan skor numerik kepada setiap nasabah berdasarkan sejumlah faktor yang relevan. Demikian pula, dengan penerapan prinsip yang sama dapat dikembangkan sebuah model yang mampu memberikan skor kelayakan kepada setiap konten berita. Skor ini akan didasarkan pada berbagai faktor yang melekat pada sebuah berita, seperti panjang judul (jumlah kata pada judul), kategori berita (rubrik), penulis (divisi), jumlah gambar/foto/infografis, jumlah video/grafik interaktif, dan sebagainya.

Model penilaian ini akan semakin relevan untuk perusahaan yang memproduksi berita digital berbayar. Model ini dapat membantu redaksi dalam menentukan konten mana yang layak untuk dijadikan konten premium dan dapat menarik minat pelanggan berbayar. Dengan demikian, model ini tidak hanya dapat meningkatkan kualitas konten secara keseluruhan, tetapi juga dapat menjadi sumber pendapatan baru bagi media. Adanya model penilaian ini diharapkan redaksi media dapat meningkatkan kualitas konten, yaitu dengan menyajikan berita yang memiliki nilai berita tinggi dan relevan dengan minat pembaca. Standar kualitas yang terjaga akan meningkatkan kepercayaan publik dan menjaga reputasi penerbit sebagai media yang kredibel dan terpercaya. Efek turunannya diharapkan dapat juga mendorong pertumbuhan bisnis, yaitu dengan mengoptimalkan pendapatan dari konten berbayar.

Tujuan dari penelitian ini adalah menerapkan *metode credit risk scoring* untuk membantu pengelola media digital dalam menentukan kelayakan berita yang akan diterbitkan. Penerapan metode ini diharapkan dapat membantu editor menentukan berita mana yang sudah dan belum memenuhi kriteria sebelum dipublikasikan. Tujuan penelitian dapat dicapai dengan memenuhi tiga tujuan praktis berikut. Pertama, akan dibentuk sebuah peubah baru sebagai peubah target, yang akan digunakan sebagai penentu kelayakan berita. Kedua, menentukan peubah yang akan menjadi penyumbang nilai dalam skor kelayakan. Terakhir adalah menentukan batas nilai (skor) berita yang akan menjadi pembatas antara berita yang sudah layak terbit dan belum.

Perhitungan skor akan dibuat otomatis ke dalam sistem, sehingga penulis dapat mengetahui skor tulisannya sesaat setelah mengunggah tulisan. Begitu juga editor dapat dengan cepat melihat skor kelayakan berita dari setiap tulisan yang masuk. Namun, perlu dipahami bahwa model penilaian ini tidak bermaksud untuk menggantikan peran editor dalam mengevaluasi berita. Model ini hanya sebagai alat bantu untuk meningkatkan efisiensi dan akurasi dalam proses pengambilan keputusan. Selain itu hasil penelitian diharapkan dapat memberikan kontribusi yang signifikan bagi pengembangan jurnalisme digital yang lebih berkualitas.

2. Metodologi

2.1 Data

Data yang digunakan adalah data berita dan data performa. Data berita adalah daftar semua berita digital yang diterbitkan dari tanggal 1 Januari 2022 hingga 31 Maret 2022 di kanal kompas.id. Selama periode tersebut total 11.584 berita yang diterbitkan.

Sedangkan data kedua adalah data performa dari setiap berita diukur menggunakan aplikasi Chartbeat. Chartbeat adalah sebuah panel kontrol (*dashboard*) yang berisi performa dari setiap halaman (berita) dari sebuah website. Dari aplikasi Chartbeat akan ditarik tiga buah data untuk setiap berita, yaitu jumlah total berapa kali berita diakses (*pageviews*), total waktu sebuah artikel pernah diakses (*Total Engaged Time*), dan rata-rata lama akses per artikel per pembaca (*Average time on page*).

2.2 Metode Penelitian

Proses pembentukan skor kelayakan berita melalui beberapa tahapan sebagai berikut:

1. Eksplorasi data performa. Eksplorasi data ditujukan untuk membandingkan sebaran dan variasi dari setiap peubah. Dari hasil di tahap ini akan ditentukan perlu tidaknya penanganan data awal, misalnya ketika ditemukan adanya keanehan ataupun ketidaksesuaian data dengan kebutuhan analisa nanti. Beberapa alat statistik yang bisa digunakan, antara lain: frekuensi, tabulasi silang (*crosstabulations*), grafik pencar (*scatter plot*), diagram kotak (*box plot*). Penanganan awal dilakukan bila ditemukan anomali atau perbedaan yang sangat besar antar peubah.

   Pengolahan pendahuluan dilakukan untuk mendapatkan gambaran dari hasil klasifikasi data perlu dilakukan untuk mendeteksi kemungkinan terjadinya pengelompokan data yang tidak seimbang (*imbalanced data*). Jika hal ini terjadi perlu dilakukan penyeimbangan data dengan menggunakan teknik *undersampling* dan/atau *oversampling*. Melihat kelebihan dan kekurangan dari masing-masing teknik, dalam penelitian ini akan menggunakan kombinasi dari keduanya. Pada kelompok minoritas akan dilakukan oversampling sedangkan pada kelompok mayoritas akan dilakukan undersampling. Kombinasi ini diharapkan dapat mengurangi efek bias dari masing-masing teknik.
2. Pengelompokkan data performa. Setelah data siap, langkah selanjutnya adalah melakukan pengelompokan data menjadi sebuah peubah yang nanti akan digunakan sebagai peubah target pada proses selanjutnya. Tujuan melakukan pengelompokan data adalah membagi semua berita menjadi 2 kelompok, yaitu berita yang sudah layak terbit dan yang belum. Dalam penelitian ini akan digunakan 3 algoritma pengelompokan data, yaitu *K-Means, K-Medoids*, dan *K-Modes.*

   Pemilihan algoritma ini didasari pada kenyataan bahwa penerapan model *credit risk scoring* dalam bidang jurnalistik ini merupakan yang pertama kali dilakukan (N. & Boitan, 2009). Atas dasar tersebut digunakan algoritma standar ketika awal pengembangan model ini yang sekaligus merupakan metode statistika standar yang paling umum digunakan.

Analisis pengelompokan digunakan untuk mengidentifikasi pola dan struktur dalam data yang hasilnya akan dipergunakan sebagai peubah target pada perhitungan skoring di langkah selanjutnya. Metode yang dipergunakan di sini telah banyak dibahas dalam literatur, di mana *K-Means* dikenal sebagai salah satu algoritma pengelompokan paling populer dan efektif dalam mengelompokkan data berdimensi tinggi (Hartigan & Wong, 1979; Huang, 1997; Seitshiro & Govender, 2024).

3. Evaluasi model pengelompokkan. Hasil pengelompokan terbaik dari algoritma *K-Means, K-Medoids,* dan *K-Modes* akan ditentukan berdasarkan ukuran *Silhouette*, WSS (*Within-Cluster Sum of Squares*) dan *gap statistics* (Sari et al., 2019). Dari ketiga ukuran ini akan ditentukan hasil pengelompokan mana yang lebih baik dibandingkan yang lain. Dalam hal tidak ada pengelompokan yang unggul mutlak, maka hasil akhir akan ditentukan berdasarkan hasil klasifikasi terbanyak dari ke-3 ukuran tersebut (*Majority Vote*).
   Hasil akhir dari pengelompokan adalah sebuah peubah target dikotomi sebagai representatif dari berita yang layak terbit dan yang belum. Dengan target ini, metode c*redit risk scoring* dapat diterapkan untuk menghasilkan skor pada setiap berita.
4. Pengolahan data karakteristik berita. Sebelum menentukan skor berita, proses ekplorasi data dilakukan pada data kedua, yaitu data karakteristik berita. Sama seperti data performa, penangan data perlu dilakukan bila ditemukan anomali pada data. Transformasi dan ekstraksi data dari peubah hasil seleksi juga perlu dilakukan, terutama untuk peubah yang berisi teks, seperti judul, nama penulis, keyword, isi tulisan, dan sebagainya.
5. Transformasi peubah. Transformasi dilakukan agar informasi dari setiap peubah tersebut bisa dimanfaatkan dan bisa digunakan dalam algoritma pembelajaran mesin. Konversi teks menjadi panjang teks dan jumlah kata dilakukan untuk mengkuantifikasi peubah-peubah ini menjadi bentuk numerik. Hal lain yang juga dilakukan adalah dengan pengkategorian ulang, seperti mengubah tanggal publikasi menjadi kategori hari dalam seminggu, dan mengkoding ulang nama penulis menjadi divisi tempat penulis bekerja.
   Dari proses ini selain mendeteksi dan pemberbaiki kesalahan dalam data, juga melihat konsistensi dan keterkaitan antar peubah. Proses ini sangat penting dilakukan terutama dalam data besar, sebab semakin besar data semakin besar juga kemungkinan ada data yang tidak sesuai, atau perlu disesuaikan dulu sebelum diolah lebih lanjut. Peubah yang mempunyai pencilan juga perlu dilakukan penyesuaian dan transformasi merupakan supaya data lebih standar dan seragam.
6. Pengkategorian peubah. Metode skor resiko kredit memerlukan peubah dalam bentuk kategori. Untuk itu semua peubah (numerik) dikonversi menjadi kategori menggunakan metode prinsip panjang deskripsi minimum (*Minimum Description Length Principle*/MDLP). Algoritma ini akan meminimalkan jumlah kelas untuk tiap peubah, sehingga model akan menjadi lebih sederhana dan mengurangi risiko *overfitting* (Kamimura et al., 2023).
   Beberapa peubah kategorik, seperti Kategori Rubrik dilakukan pengkodingan ulang, agar jumlah kategorinya sesedikit mungkin. Untuk peubah skala nomimal

seperti Tanggal dan Bulan dibiarkan apa adanya, karena tidak mungkin dilakukan pengkategorian ulang.

7. Eksplorasi hubungan antara data performa dan data karakteristik berita. Setelah proses persiapan peubah selesai, tahapan selanjutnya dilakukan penggabungan dari kedua tabel yang akan digunakan. Hasil tabel penggabungan inilah yang akan digunakan dalam model. Sebelum proses pemodelan, kembali dilakukan pengolahan pendahuluan untuk mendapatkan gambaran awal dari data gabungan, seperti sebaran, pencilan, hubungan antar peubah, dan visualisasinya.
8. Metode skor resiko kredit. Perhitungan skor resiko kredit menggunakan dua metrik, yaitu *Weight of Evidence* (WOE) dan *Information Value* (IV). Kedua metrik ini digunakan untuk mengevaluasi kekuatan setiap peubah dalam memprediksi layak atau belum sebuah berita untuk diterbitkan. Walaupun berkaitan, keduanya mempunyai fungsi yang berbeda. WOE mengukur seberapa kuat setiap kategori dalam setiap peubah dalam mendeteksi kelayakan berita. Sedangkan IV digunakan untuk mengukur kekuatan prediksi dari setiap peubah. IV dihitung dengan menjumlahkan hasil kali selisih antara proporsi layak dan tidak dalam setiap kategori dengan WOE untuk kategori tersebut.
9. Uji Multikolinearitas. Sebelum masuk ke analisis regresi, dilakukan uji pendahuluan terhadap data. Uji pendahuluan dilakukan dengan menghitung skor VIF (*Variance Inflation Factor*) untuk melihat apakah ada multikolinearitas.
   Perhitungan tingkat kepentingan dari setiap peubah. Proses seleksi peubah akan dilakukan bersamaan dengan penerapan regresi logistik. Dalam penelitian ini seleksi peubah akan diserahkan pada regresi logistik sebab berdasarkan temuan Seitshiro & Govender (2024), model regresi logistik tanpa transformasi WOE lebih dominan daripada model regresi logistik dengan transformasi WOE.
   Ada beberapa metode regresi logistik yang dapat melakukan seleksi peubah, salah satunya adalah regresi logistik *stepwise*. Menurut Harrell (2015), metode *stepwise* sangat baik digunakan untuk data besar dengan peubah banyak, karena kemampuannya dalam mereduksi jumlah peubah. Cara kerjanya yang menggabungkan metode *forward* dan *backward*, membuatnya lebih *superior*, walaupun masih tetap ada beberapa permasalahan. Masalah seperti *overfitting*, dan bias masih tetap dimungkinkan terjadi, sehingga dasar teori dan penyesuaian dengan permasalah tetap diperlukan.
10. Evaluasi Model Regresi Logistik. Untuk mengevaluasi hasil regresi logistik *stepwise*, digunakan dua ukuran, yaitu AIC (*Akaike Information Criterion*) dan BIC (*Bayesian Information Criterion*). AIC dan BIC adalah dua kriteria informasi yang sering digunakan dalam pemilihan model statistik. Keduanya digunakan untuk membandingkan berbagai model statistik dan memilih model yang paling sesuai dengan data.
11. Penentuan titik batas (*cut off*). Tahapan selanjutnya adalah menentukan titik batas yang akan menjadi nilai pembatas antara skor nilai berita yang sudah layak terbit dan yang belum. Menurut (Zhang, 2018) penentuan titik batas ini sangat penting karena dapat mempengaruhi keputusan akhir. Tidak ada aturan baku untuk menentukan titik batas yang optimal. Pilihan titik batas sangat bergantung pada konteks permasalahan dan tujuan penelitian.
12. Evaluasi model. Dalam proses penentuan model regresi terbaik, akan dilakukan

juga evaluasi dan optimisasi model untuk mendapatkan cut off optimal (Karmakar, 2023). Data gabungan yang sudah berbentuk kelas dari proses MDLP akan dibagi menjadi 80% data training dan 20% data testing. Pemilihan Regresi Logistik *stepwise* (Atodaria & Pentar, 2024) dapat melakukan seleksi peubah, yang memastikan hanya peubah yang mempunyai tingkat kepentingan tinggi terhadap target yang akan digunakan dalam pembentukan skor kelayakan berita.
Model Regresi Logistik terbaik akan ditentukan berdasarkan nilai sensitifitas (*sensitivity/Recall*), spesifisitas (*specificity*), akurasi (*accuracy*), dan AUC (*Area Under the ROC Curve*).

13. Penentuan batas nilai skor kelayakan berita. Hasil dari Regresi logistik ini akan mengidentifikasi peubah mana saja yang berpengaruh, dan setelah itu dilakukan penjumlahan nilai skor sebagai pembatas antara berita yang sudah layak terbit dan yang belum. Batas skor inilah yang nanti akan menjadi penilai obyektif untuk setiap berita yang akan dipublikasikan.

3. Hasil dan Pembahasan

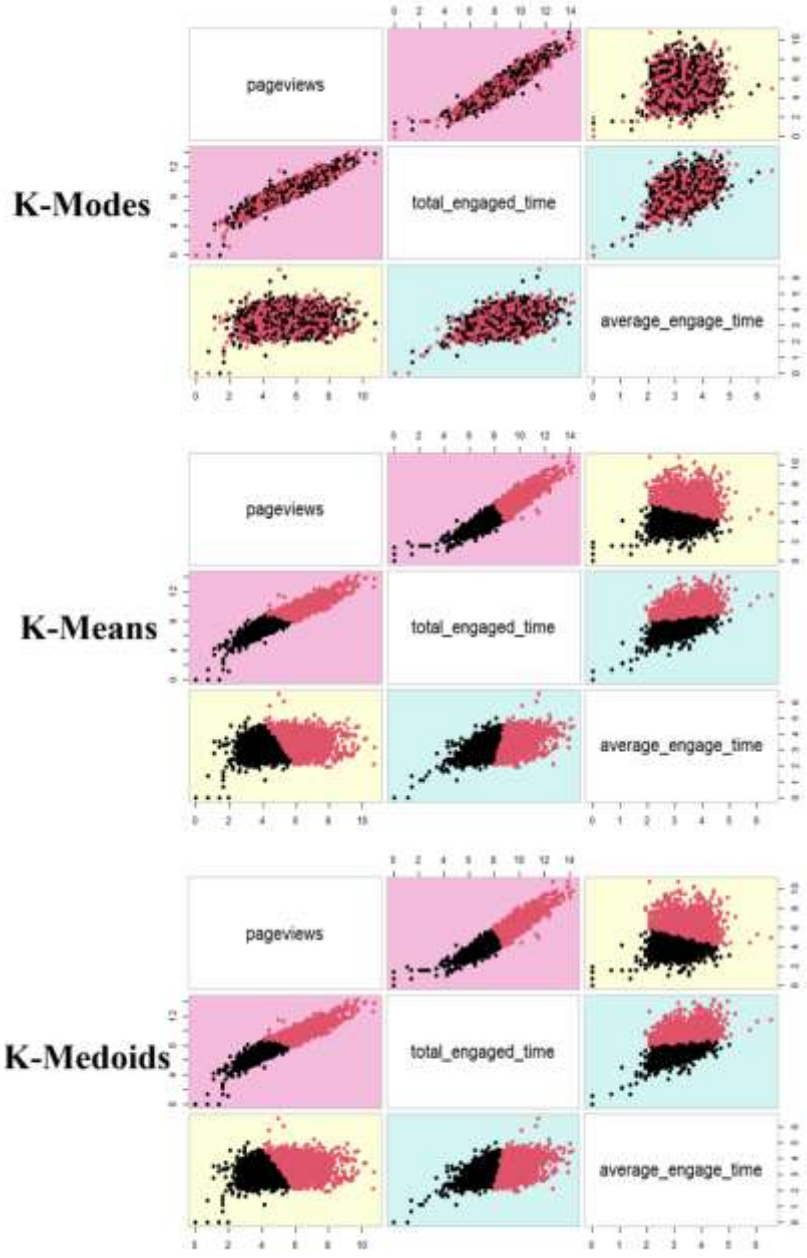
Pengolahan data performa diawali dengan eksplorasi data. Dari proses ini diketahui bahwa data memiliki banyak pencilan, sehingga dilakukan beberapa teknik penanganan data. Teknik penanganan data yang digunakan adalah Standarisasi (Standz), Winsorisasi (Winsor), Winsorisasi yang distandarisari (Standz Winsor), dan transformasi logaritma (Log) (Kwak & Kim, 2017).

Dari setiap jenis data kemudian digunakan tiga metode pengelompokan untuk menggabungkan tiga peubah dalam data performa menjadi satu peubah yang akan digunakan sebagai peubah target pada proses selanjutnya. Hasilnya dapat dilihat pada Tabel 1.

Berdasarkan hasil perbandingan ketiga metode pengelompokan terhadap lima jenis data dapat disimpulkan bahwa metode *K-Means* menunjukkan performa yang lebih baik dibandingkan dengan K-Mode dan *K-Medoids*. Hal ini terlihat dari nilai *Silhouette* yang lebih tinggi, yang menunjukkan bahwa objek dalam kluster *K-Means* memiliki kedekatan yang lebih baik satu sama lain dan pemisahan yang lebih jelas dari kluster lainnya. Selain itu, WSS untuk *K-Means* juga menunjukkan nilai terendah, yang menandakan variabilitas dalam pengelompokan yang lebih kecil. Untuk *Gap Statistic*, sebenarnya nilai yang dihasilkan ketiga metode hampir sama, dengan sedikit keunggulan pada K-Mode.

Tabel 1: Perbandingan Kualitas Pengelompokan Terhadap Beberapa Jenis Preprocessing Data

Teknik	Metode	Data Awal	Standz	Winsor	Standz Winsor	Log
Silhouette	K-Mode	-0,0014	0,0004	-0,0001	-0,0001	0,0003
Silhouette	K-Means	**0,9811**	**0,5537**	0,7413	**0,5617**	**0,4732**
Silhouette	K-Medoids	0,7248	0,4735	**0,7442**	0,5556	0,4697
WSS	K-Mode	1,69E+13	34737,17	3,14E+11	7.818,87	36.085,08
WSS	K-Means	**6,27E+12**	**27894,64**	**6,01E+10**	**2.626,96**	**17.160,55**
WSS	K-Medoids	1,47E+13	28323,01	6,16E+10	2.643,06	17.191,67
gap_stat	K-Mode	**3,6153**	3,1852	-0,0117	**0,2912**	**1,1991**
gap_stat	K-Means	3,0946	3,3150	**0,1061**	0,2539	1,0896
gap_stat	K-Medoids	3,2863	**3,3487**	0,0934	0,2560	1,0902



Keterangan:
***merah**: korelasi kuat,*
***biru**: korelasi sedang,*
***kuning**: korelasi lemah/ tidak ada korelasi*

Gambar 1: Perbandingan Hasil K-Mode, *K-Means*, dan *K-Medoids* secara visual

Untuk penentuan hasil akhir, digunakan pendekatan pilihan terbanyak (*Majority Voting/Concensus*). Menurut Vega-Pons & Ruiz-Shulcloper (2011), dengan pendekatan pilihan terbanyak dari beberapa metode dapat meningkatkan kekekaran (*robustness*) hasil, mengurangi dampak optimum lokal dan menghasilkan klaster yang lebih stabil. Dari sini diperoleh bahwa kluster yang terbentuk dari metode *K-Means* lebih optimal dibandingkan dengan metode lainnya.

Hasil ini diperkuat lagi dengan perbandingan hasil visualisasi pengelompokan yang dihasilkan dari ketiga metode, yang dapat dilihat pada Gambar 1. Hasil visualisasi pengelompokan pada *K-Mode* tidak terlihat batas yang jelas antar kelompok yang dihasilkan. Berbeda dengan hasil *K-Means* dan *K-Medoids* yang secara visual sudah berhasil membagi data menjadi dua kelompok.

Setelah penentuan peubah target, tahapan selanjutnya adalah pengolahan data karakteristik berita. Pengolahan data berita diawali dengan penggabungan dengan data performa dan peubah target. Proses pertama pada data gabungan adalah eksplorasi data. Jenis data sangat bervariasi sehingga perlu dilakukan standarisasi dan transformasi data. Transformasi data menggunakan algoritma MDLP (*Minimum Description Length Principle*) yang akan mengubah data menjadi kategori, dan sekaligus meminimalkan jumlah kelas untuk tiap peubah. Hasil dari algoritma MDLP dapat dilihat pada Tabel 2.

Tabel 2: Perbandingan Jumlah Kategori Sebelum dan Sesudah Modifikasi

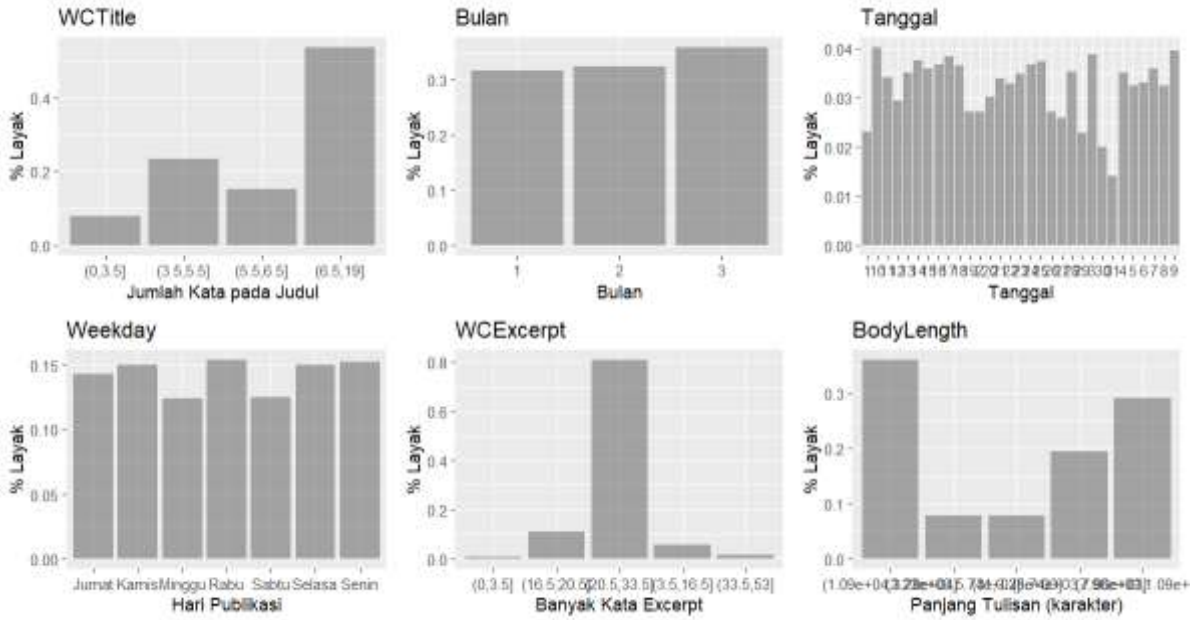
Jenis Data	Nama Peubah Awal	Kategori Awal	Nama Peubah Baru	Kategori Hasil
Numerik	TitleLength	107	WOE_TitleLength	4
Numerik	WCTitle	18	WOE_WCTitle	4
Numerik	**Bulan**[a]	3	**WOE_Bulan**	3
Numerik	**Tanggal**[a]	31	**WOE_Tanggal**	31
Numerik	**Weekday**[a]	7	**WOE_Weekday**	7
Numerik	WCKeyword	18	WOE_WCKeyword	2
Numerik	ExcerptLength	213	WOE_ExcerptLength	5
Numerik	WCExcerpt	47	WOE_WCExcerpt	5
Numerik	BodyLength	7287	WOE_BodyLength	5
Numerik	Image	19	WOE_Image	4
Numerik	Video	4	WOE_Video	2
Numerik	Time (hours)	9708	WOE_TimeHours	8
Numerik	DayDiff	3171	WOE_DayDiff	2
Karakter	Kategori	68	WOE_Kategori2	23
Karakter	DayName[b]	7	-	-
Karakter	Authors[c]	1290	-	-
Karakter	Divisi	7	WOE_Divisi	7
Karakter	TimeKategory	7	WOE_TimeKat	7

Keterangan:
[a] *Peubah skala nominal tidak dilakukan modifikasi*
[b] *Peubah DayName sudah diwakili peubah Weekday*
[c] *Peubah Authors sudah diwakili peubah Divisi*

Beberapa peubah kategori yang tidak dapat mengikuti ketentuan ini, diupayakan mereduksi jumlahnya sampai ke angka yang memungkinkan. Contoh pada jumlah

rubrik dilakukan pengkategorian ulang secara manual, dan berhasil mengurangi jumlah rubrik dari 68 menjadi 23 rubrik utama.

Proses berikutnya adalah visualisasi semua peubah untuk melihat efek hasil pengkategorian pada setiap peubah terhadap peubah target. Berdasarkan perbandingan visual (Gambar 2) dari tingkat kelayakan berita terhadap kategori peubah, dapat dilihat bahwa setiap kategori pada peubah mempunyai tingkat kelayakan yang berbeda. Ada yang mempunyai pola tertentu, tapi ada juga yang tidak. Hal ini mengindikasikan tidak semua peubah cukup baik disertakan dalam model. Untuk itu pada penghitungan model regresi logistik digunakan metode *stepwise* agar peubah yang kurang penting akan tereleminasi secara langsung.



Gambar 2: Perbandingan Tingkat Kelayakan pada Beberapa Kategori Peubah

Regresi logistik ini merupakan metode awal yang digunakan dalam pengembangan skor resiko kredit (Altman, 1968) dan merupakan metode yang paling umum digunakan dibandingkan metode lainnya (Thomas et al., 2002). Trinh (2024), membandingkan sembilan metode klasifikasi dalam pembentukan skor resiko kredit juga menyatakan bahwa regresi logistik masih menjadi salah satu metode terbaik dan lebih stabil dibandingkan metode lainnya.

Hasil akhir model regresi logistik *stepwise* berhasil mengurangi jumlah peubah dari 18 menjadi 10 buah, dengan taraf signifikansi 0,01. Dari 10 peubah yang terpilih, kategori rubrik (Kategori2) menjadi peubah yang paling penting, lalu panjang isi tulisan (BodyLength) dan waktu publikasi (TimeHour). Peubah lainnya mempunyai tingkat kepentingan yang lebih rendah. Untuk peubah lain di luar kesepuluh peubah ini, bukan berarti tidak penting. Kemungkinan peubah lain tersebut juga mempunyai tingkat kepentingan terhadap target, namun mungkin sudah terwakili oleh peubah yang lainnya.

Tahapan selanjutnya adalah menentukan titik batas (*cut off*). Untuk itu dipergunakan nilai sensitifitas (*sensitivity/Recall*), spesifisitas (*specificity*), akurasi (*accuracy*), dan AUC (*Area Under the ROC Curve*) yang dihitung dari matriks

klasifikasi.

Berdasarkan pengelompokan kelas peubah target sebelumnya, diperoleh hasil bahwa jumlah kelas negatif dan positif cukup berimbang. Dalam kasus seperti ini, penentuan titik batas dapat didasari pada nilai akurasi saja. Sedangkan nilai AUC dipergunakan untuk melihat kemampuan model secara keseluruhan dalam membedakan antara dua kelas (layak atau belum).

Tabel 3: Perbandingan Nilai Kebaikan Model Berdasarkan Nilai *Cut off*

Titik Batas	Akurasi	Sensitifitas	Spesifisitas	AUC
0,1	0,64604	**0,98949**	0,18145	0,58547
0,2	0,66590	**0,98549**	0,23358	0,60953
0,3	0,69525	**0,96697**	0,32769	0,64733
0,4	0,72230	**0,90741**	0,47190	0,68966
0,5	0,72144	**0,81481**	0,59513	0,70497
0,6	**0,72547**	**0,72222**	**0,72986**	**0,72604**
0,7	0,67856	0,56306	**0,83480**	0,69893
0,8	0,58245	0,32883	**0,92552**	0,62718
0,9	0,47799	0,09760	**0,99255**	0,54508

Tabel 3 menunjukkan bahwa nilai akurasi tertinggi adalah 0,72547, dengan AUC terbesar adalah 0,72604. Kedua nilai ini terjadi pada titik batas yang sama, yaitu 0,6, sehingga nilai titik batas inilah yang akan digunakan dalam penentuan skor berita.

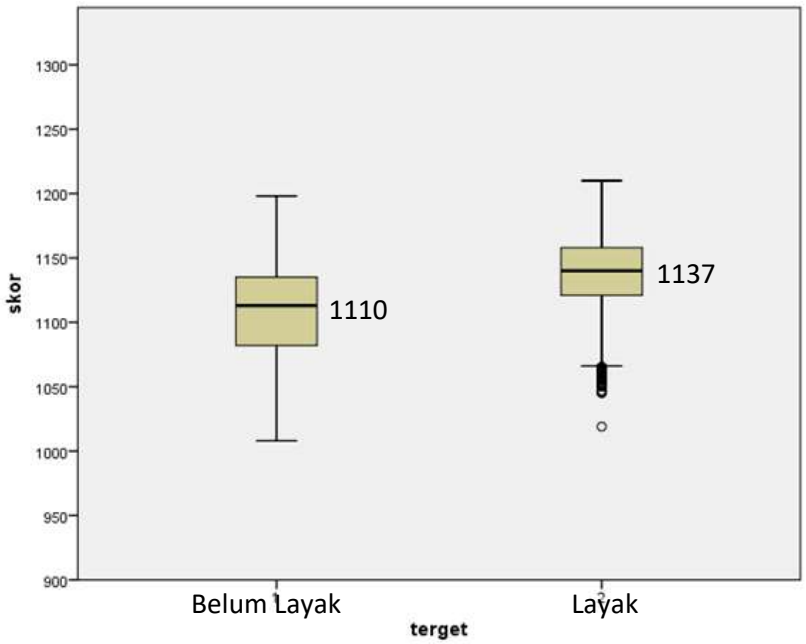
Perhitungan skor setiap kategori peubah khusus pada peubah yang lolos seleksi pada analisis regresi logistik *stepwise*. Skor dihitung berdasarkan karakteristik berita yang telah dikelompokkan sebelumnya. Berdasarkan skor dapat diketahui kategori mana yang akan menghasilkan nilai yang lebih baik dibandingkan dengan kategori yang lain. Daftar skor inilah yang akan menjadi panduan penulisan berita bagi penulis. Penulis akan mengumpulkan skor pada setiap kategori peubah yang sudah dipenuhi. Total skor yang dikumpulkan menjadi skor kelayakan berita yang dihasilkan penulis. Total skor ini akan dibandingkan dengan batas skor yang ditetapkan editor atau penerbit sebagai seleksi awal berita yang akan diterbitkan.

Dalam penentuan batas skor kelayakan berita, pertama dilakukan perhitungan skor semua berita berdasarkan jumlah total skor pada tiap elemen berita. Hasil sebaran skor secara visual dapat dilihat pada Gambar 3. Sebaran skor antara yang layak dan belum layak saling beririsan. Hal ini menandakan total skor yang sama memungkinkan masuk kelompok yang layak ataupun sebaliknya. Pada kasus seperti ini, penentuan skor sebagai batas kelayakan berita untuk terbit menjadi agak sulit. Rata-rata skor dari kelompok layak sekitar 1137 poin, hanya sedikit lebih tinggi dibandingkan kelompok sebaliknya yang rata-ratanya 1110. Untuk kasus ini diperlukan penentuan nilai rata-rata antara kedua kelompok (Thomas et al., 2002) sebagai batas minimal skor kelayakan berita, yaitu: 1123.

Dengan skor batas ini, penulis dan editor berita mendapatkan acuan awal yang obyektif bagi tulisan yang akan diterbitkan. Tulisan yang sudah mendapatkan skor lebih dari 1123 dapat dianggap sudah memenuhi syarat minimal untuk diterbitkan, sehingga editor bisa fokus dalam mengoreksi elemen-elemen kualitatif, seperti sudut pandang, logika pemikiran, akurasi dan pemilihan kata, baik di dalam isi berita, judul maupun kutipan.

Total skor ini juga dapat dimanfaatkan sebagai penentu berita layak “premium” atau “non-premium”, yaitu dengan penetapan skor yang lebih tinggi. Berita yang sudah mencapai batas skor “premium” memiliki kualitas yang tinggi sesuai kebutuhan konsumen dan akan mempunyai performa yang optimal ketika nanti dipublikasikan.

Gambar 3: Perbandingan Sebaran Skor antara Tulisan yang Layak atau Tidak Layak



4. Simpulan dan Saran

Simpulan

Penerapan c*redit risk scoring* sebagai Skor kelayakan berita memungkinkan untuk dilakukan. Penentuan skor batas kelayakan berita memungkinkan editor dengan cepat menyeleksi berita yang akan dapat segera diterbitkan. Simpulan ini berdasarkan hasil pada tiga tahapan penerapan. Tahapan pertama adalah, pembentukan peubah baru dilakukan melalui transformasi logaritma untuk mengurangi perbedaan variasi antar peubah dan mengatasi pencilan dalam data. Metode pengelompokan terbaik adalah algoritma *K-Means* yang menggabungkan data performa menjadi sebuah peubah baru sebagai target dalam penentu kelayakan berita. Tahapan Kedua, tingkat kepentingan dan kesederhanaan model diukur dengan model regresi logistik *stepwise*. Model ini menghasilkan 10 peubah. Tahapan Ketiga adalah batas skor minimal berita untuk diterbitkan adalah 1123.

Saran

Pengembangan metode ini masih sangat terbuka, misalnya dengan penggunaan algoritma pembelajaran mesin lain yang lebih kompleks, seperti *Random Forest, Support Vector Machines* (SVM), dan *Neural Networks*. Pemilihan algoritma perlu disesuaikan dengan data yang akan digunakan.

Pengembangan lain adalah memasukkan aspek selain kuantitatif dari berita dalam model. Aspek kualitatif mencakup antara lain aktualitas, obyektifitas, sudut pandang, nada berita (tone positif/negatif), hingga kelengkapan berita yang menjawab informasi

tentang 5W+1H (*Who, What, Where, When, Why, + How*), akan meningkatkan sisi kualitas dalam pembentukan standar kelayakan berita.

Daftar Pustaka

Abdou, H. A., & Pointon, J. (2011). Credit Scoring, Statistical Techniques and Evaluation Criteria: a Review of The Literature: Credit Scoring, Techniques & Evaluation Criteria: A Literature Review. *Intelligent Systems in Accounting, Finance and Management*, *18*(2–3): 59–88. https://doi.org/10.1002/isaf.325

Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction Of Corporate Bankruptcy. *The Journal of Finance*, *23*(4): 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Atodaria, Z., & Pentar, S. (2024). *Credit Risk Analysis Using Logistic Regression Modeling.* 57.

Gharehgozli, A. H., & Zaerpour, N. (2018). Stacking outbound barge containers in an automated deep-sea terminal. *Eur. J. Oper. Res.*, *267*: 977–995.

Harrell , F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. https://doi.org/10.1007/978-3-319-19425-7

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1): 100. https://doi.org/10.2307/2346830

Huang, Z. (1997). *Clustering Large Data Sets with Mixed Numeric and Categorical Values*. Retrieved from https://api.semanticscholar.org/CorpusID:3007488

Israel, S., Caspi, A., Belsky, D., Harrington, H., Hogan, S., Houts, R., … Moffitt, T. (2014). Credit scores, cardiovascular disease risk, and human capital. *Proceedings of the National Academy of Sciences of the United States of America*, *111*. https://doi.org/10.1073/pnas.1409794111

Kamimura, E. S., Pinto, A. R. F., & Nagano, M. S. (2023). A recent review on optimisation methods applied to credit scoring models. *Journal of Economics, Finance and Administrative Science*, *28*(56): 352–371. https://doi.org/10.1108/JEFAS-09-2021-0193

Karmakar, A. (2023). *Machine Learning Approach to Credit Risk Prediction: A Comparative Study Using Decision Tree, Random Forest, Support Vector Machine and Logistic Regression*. https://doi.org/10.13140/RG.2.2.31652.14725

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4): 407. https://doi.org/10.4097/kjae.2017.70.4.407

Lauer, J. (2017). *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. https://doi.org/10.7312/laue16808

Muqsith, M. A. (2021). Teknologi Media Baru: Perubahan Analog Menuju Digital. *ADALAH*, *5*(2): 33–40. https://doi.org/10.15408/adalah.v5i2.17932

N., D. & Boitan. (2009). A Cluster Analysis Approach for Bank's Risk Profile: The Romanian Evidence. *European Research Studies Journal*, *XII*(Issue 1): 109–118. https://doi.org/10.35808/ersj/213

Onay, C., & Öztürk, E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance*, *26*(3): 382–405. https://doi.org/10.1108/JFRC-06-2017-0054

Sanderford, A. R., McCoy, A. P., Keefe, M. J., & Zhao, D. (2014). *Adoption Patterns of Energy Efficient Housing Technologies 2000-2010: Builders as Innovators?*

Sari, P. D., Aidi, M. N., & Sartono, B. (2019). Credit Scoring Analysis using LASSO Logistic Regression and Support Vector Machine (SVM). *International Journal of Engineering and Management Research*, *7*(4): 393–397.

Sathye, M., & Islam, J. (2011). Adopting a risk-based approach to AMLCTF compliance: the Australian case. *Journal of Financial Crime*, *18*(2): 169–182. https://doi.org/10.1108/13590791111127741

Seitshiro, M. B., & Govender, S. (2024). Credit risk prediction with and without weights of evidence using quantitative learning models. *Cogent Economics & Finance*, *12*(1): 2338971. https://doi.org/10.1080/23322039.2024.2338971

Thomas, L. C., Edelman, D., & Crook, J. N. (2002). *Credit scoring and its applications*. https://doi.org/10.1137/1.9780898718317

Trinh, L. T. (2024). A comparative analysis of consumer credit risk models in Peer-to-Peer Lending. *Journal of Economics, Finance and Administrative Science*, *29*(58): 346–365. https://doi.org/10.1108/JEFAS-04-2021-0026

Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, *25*(03): 337–372. https://doi.org/10.1142/S0218001411008683

Zhang, Z. (2018). *Estimating The Optimal Cutoff Point For Logistic Regression*. Retrieved from https://digitalcommons.utep.edu/open_etd/1565