**Pemodelan Topik pada Berita Kompas Menggunakan Latent Dirichlet Allocation (Studi Kasus Bola Kategori Internasional)Shinta Zahira Hayathun Nufus1, Suci Faaza Naafia2, Rukmiati3,**

**Program Studi Ilmu Komputer**

**Universitas Muhammadiyah Bima**

**Abstrak**

Penelitian ini bertujuan untuk mengidentifikasi topik-topik utama yang sering muncul dalam berita bola internasional di media daring Kompas.com. Sebanyak 28 berita dikumpulkan secara manual dari kategori Internasional dan diproses menggunakan tahapan preprocessing teks seperti pembersihan data, normalisasi, tokenisasi, dan stemming. Selanjutnya, dilakukan ekstraksi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF) untuk merepresentasikan kata-kata penting dalam bentuk numerik. Proses pemodelan topik dilakukan menggunakan algoritma Latent Dirichlet Allocation (LDA) yang dievaluasi berdasarkan nilai coherence score dan pendekatan penilaian manusia. Hasil evaluasi menunjukkan bahwa model dengan jumlah topik 7 menghasilkan nilai coherence tertinggi sebesar 0.513965. Setiap topik kemudian diinterpretasikan berdasarkan kumpulan kata dengan probabilitas tertinggi untuk memahami isu-isu utama yang dibahas, seperti performa tim nasional, pertandingan internasional, peran federasi sepak bola, dan dinamika klub. Visualisasi topik dilakukan menggunakan pyLDAvis untuk memperjelas distribusi dan hubungan antar topik. Hasil penelitian menunjukkan bahwa metode LDA efektif dalam mengelompokkan topik berita bola secara semantik dan dapat digunakan sebagai alat bantu untuk menganalisis kecenderungan isi pemberitaan olahraga internasional.

Kata Kunci : Latent Dirichlet Allocation, TF-IDF, pemodelan topik, coherence score, pyLDAvis.

**Abstract**

This study aims to identify the main topics frequently discussed in international football news on the online media platform Kompas.com. A total of 28 news articles were collected through web scraping from the International Football category and processed using standard text preprocessing techniques, including data cleaning, normalization, tokenization, and stemming. Text features were extracted using the Term Frequency–Inverse Document Frequency (TF-IDF) method to numerically represent important terms. Topic modeling was carried out using the Latent Dirichlet Allocation (LDA) algorithm, which was evaluated based on the coherence score and human judgment for interpretability. The evaluation results indicate that the model with 7 topics achieved the highest coherence score of 0.513965. Each topic was interpreted by examining the top-ranked words with the highest probabilities to understand the dominant issues, such as national team performance, international matches, roles of football federations, and club and coach dynamics. Topic visualization was performed using pyLDAvis to enhance the clarity of topic distribution and inter-topic relationships. The findings demonstrate that LDA is an effective method for semantically clustering sports news topics and can be used as a tool to analyze content trends in international sports journalism.

Keywords : Latent Dirichlet Allocation, TF-IDF, topic modeling, coherence score, pyLDAvis.

1. **PENDAHULUAN**

Sepakbola merupakan salah satu cabang olahraga yang paling banyak digemari oleh sebagian besar manusia yang ada di bumi ini. Sepakbola digemari oleh semua lapisan masyarakat baik dari tingkat daerah, nasional, dan internasional, dari usia anak-anak, dewasa hingga orang tua, mereka senang memainkan sendiri atau sebagai penonton[1]. Seiring perkembangan zaman maka sepakbola juga mengalami perubahan terutama terlihat sekali pada peraturan pertandingan, perlengkapan lapangan, kelengkapan pemain, perwasitan, dan organisasi sepakbola. Kesemuanya itu bertujuan agar sepakbola lebih bisa dinikmati, digemari dan menjadi suatu suguhan[2].

Berita merupakan suatu bentuk dari penyajian suatu informasi yang dipublikasikan kepada khalayak ramai[3]. Pemberitaan dari suatu berita itu sendiri sangat dinanti oleh khalayak yang membutuhkan informasi terlebih lagi apabila berita tersebut berita yang baru, terkini, atau hangat (up to date). Pemberitaan itu sendiri adalah laporan lengkap ataupun interpretative (telah disajikan sebagai mana dianggap penting oleh redaksi pemberitaan) ataupun berupa pemberitaan penyelidikan (investigative reporting) yang merupakan pengkajian fakta-fakta lengkap dengan latar belakang, trend/kecenderungan, yang mungkin terjadi di masa yang akan datang[4]. Komunikasi massa memberi sarana untuk mengambil keputusan dan membentuk opini yang bisa digunakan untuk lebih memahami diri mereka sendiri[5].

Topik modeling merupakan salah satu teknik unsupervised machine learning untuk menemukan tema tersembunyi dari kumpulan dokumen teks besar guna mengelompokkan tema-tema tersebut menjadi satu topik[6]. Pertama kali diperkenalkan oleh David Blei, Latent Dirichlet Allocation (LDA) merupakan salah satu pemodelan topik yang paling populer, LDA digunakan untuk menampilkan sebuah topik menggunakan probabilitas dari setiap kata sehingga dapat membantu menggambarkan dokumen-dokumen menjadi lebih terstruktur.

Penelitian ini melakukan pengambilan data pada berita online khususnya berkaitan dengan bola kategori internasional. Evaluasi penelitian ini menggunakan nilai coherence dan pendekatan human untuk menganalisis jumlah topik terbaik. Keterbaruan dari penelitian ini, terletak pada hasil dari pemodelan topik yang akan divisualisasikan secara interaktif menggunakan pyLDAvis untuk memberikan kemudahan mengetahui distribusi kata dari setiap topik. Tujuan dari penelitian ini adalah untuk menganalisis topik yang paling sering dibahas dalam media berita online mengenai bola. Penelitian ini memberikan manfaat untuk menggali wawasan lebih mendalam terkait bola khususnya Internasional. Hasil dari penelitian ini dapat dijadikan sebagai sumber informasi bagi perusahaan untuk meningkatkan kualitasnya di masa mendatang. Disamping itu, masyarakat juga mendapatkan informasi yang relevan terkait bola tersebut.

1. **METODOLOGI PENELITIAN**

Tahapan penelitian kami Tahapan pertama yaitu pengumpulan data mengenai berita bola pada media berita online yaitu Kompas. Kemudian dilanjutkan dengan tahapan data preprocessing untuk membersihkan kata-kata yang tidak diperlukan. Tahap selanjutnya melakukan pemodelan topik menggunakan Latent Dirichlet Allocation (LDA). Tahap terakhir pembuatan visualisasi dari hasil pemodelan topik menggunakan pyLDAvis.

1. Pengumpulan Data

Penelitian ini menggunakan data dari media berita online. Portal berita yang digunakan sebagai sumber pada penelitian ini dipilih berdasarkan reputasi dan kredibilitas media dalam menyampaikan informasi. Media berita online yang digunakan sebagai sumber data adalah kompas.com. Penelitian ini secara khusus mengambil berita dari kategori Bola Internasional untuk membatasi cakupan pencarian dan fokus pada topik tertentu. Pengumpulan data dilakukan melalui proses web scraping menggunakan bahasa pemrograman Python, dengan menargetkan judul, isi, kategori, dan tautan berita dari halaman-halaman yang relevan di situs kompas.com. Data hasil scraping kemudian disimpan dalam file berformat .xlsx sebagai basis data untuk tahap preprocessing dan analisis lebih lanjut.Preprocessing Data

Dalam penelitian ini, penelitian ini melakukan preprocessing data untuk memahami dan

mengidentifikasi masalah terkait data, sehingga data dapat lebih diandalkan untuk kemudian digunakan. Pada tahapan ini penelitian ini memanfaatkan bahasa pemrograman menggunakan software berbasis Python.

Adapun tahapan preprocessing data diantaranya yaitu:

1. Cleaning Data

Membersihkan teks dari karakter tak penting (tanda baca, angka, simbol, html, dll).

1. Tokenisasi

Memecah teks menjadi satuan kecil (token), biasanya per kata.

1. Stemming

Digunakan untuk mengubah kata menjadi kata dasar yang memiliki makna.

Normalisasi

1. Mengubah kata ke bentuk baku (misal: “tdk” → “tidak”).
2. Fitur Ekstraksi Menggunakan TF IDF

Sebelum dilakukan pemodelan topik, penelitian ini terlebih dahulu menerapkan proses ekstraksi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF berfungsi untuk mengukur seberapa penting suatu kata dalam sebuah dokumen relatif terhadap keseluruhan dokumen, dengan memberikan bobot tinggi pada kata-kata yang spesifik dan jarang muncul di dokumen lain. Proses ini dilakukan terhadap teks yang telah melalui preprocessing, sehingga hanya kata-kata bersih dan bermakna yang digunakan sebagai fitur. Hasil ekstraksi menghasilkan representasi numerik dalam bentuk matriks, di mana setiap baris merepresentasikan dokumen dan setiap kolom merupakan kata unik. Matriks TF-IDF ini kemudian digunakan sebagai input dalam proses pemodelan topik menggunakan LDA, sehingga algoritma dapat mengenali pola kemunculan kata dan membentuk kelompok topik secara lebih akurat. Dengan demikian, TF-IDF berperan penting dalam menyaring dan merepresentasikan informasi inti dari kumpulan berita untuk keperluan analisis topik.

1. Topic Modeling Menggunakan Latent Dirichlet Allocation

Topic modeling merupakan salah satu teknik unsupervised machine learning untuk menemukan tema tersembunyi dari kumpulan dokumen teks besar guna mengelompokkan tema-tema tersebut menjadi satu Pemodelan topik menganggap bahwa setiap dokumen merupakan kombinasi dari sekumpulan topik dan kata. Latent Dirichlet Allocation (LDA) merupakan salah satu metode dari pemodelan topik yang banyak digunakan. Latent mengacu pada segala sesuatu yang tersembunyi dalam data, Dirichlet merujuk pada distribusi probabilitas yang menggambarkan distribusi topik dalam dokumen dan kata-kata dalam topik, dan Allocation berarti mengalokasikan topik atau kata-kata menjadi lebih spesifik ke dalam dokumen. LDA digunakan untuk menampilkan sebuah topik menggunakan probabilitas dari setiap kata sehingga dapat membantu menggambarkan dokumen-dokumen menjadi lebih terstruktur. Latent Dirichlet Allocation dianggap lebih unggul dalam menghasilkan topik bermakna logis, kemampuan interpretasi, dan performansi prediktif.

Pada tahapan ini, penelitian ini menggunakan dua parameter acuan yaitu number of topics dan words of topic, untuk menghasilkan model dengan jumlah topik terbaik. Penelitian ini terlebih dahulu menentukan klaster words of topic atau jumlah kata per topik, yaitu berjumlah 10 topik, memiliki tujuan agar model yang dibangun menghasilkan model yang jelas dan tidak bersinggungan satu sama lain. Kemudian, penelitian ini menganalisis number of topics atau jumlah topik dengan melakukan evaluasi menggunakan nilai coherence, dimana kumpulan kata yang dihasilkan oleh topic modeling dinilai berdasarkan tingkat kemudahan pemahaman manusia.

Nilai coherence dapat mengukur hubungan antar kata dalam pemodelan topik, evaluasi topik terbaik dinilai berdasarkan nilai coherence tertinggi yang diperoleh. Di samping itu, penelitian ini juga melakukan evaluasi menggunakan pendekatan human judgement untuk menginterpretasikan distribusi kata dari hasil pemodelan topik. Dalam konteks penilaian manusia, penelitian ini menggunakan pendekatan intrusi kata untuk mengevaluasi topik yang dinilai memiliki makna sama dan saling berkaitan satu sama lain sehingga dapat diidentifikasi oleh manusia. Pendekatan ini merujuk pada sejauh mana topik yang dihasilkan oleh model dapat direpresentasikan dengan logis atau alami menurut pandangan manusia. Dalam penelitian ini, penelitian ini memanfaatkan library gensim menggunakan software berbasis Python.

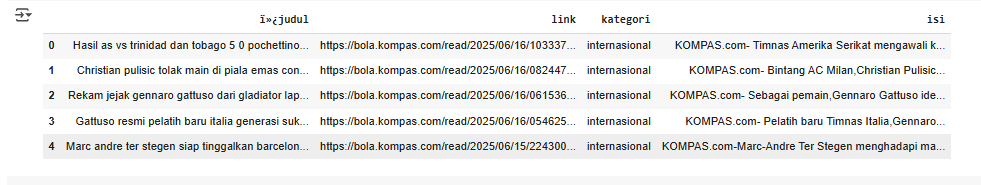
1. Visualisasi Topic Modeling

Hasil analisis topic modeling kemudian divisualisasikan menggunakan sistem visualisasi interaktif berbasis web yaitu LDAvis untuk membantu memahami makna setiap topik dan hubungan antara topik yang satu dengan topik lainnya. Penelitian ini menggunakan library pyLDAvis pada software berbasis Python. Dengan menggunakan PyLDAvis, distribusi kata dan relevansinya dengan masing-masing topik

dapat divisualisasikan dengan jelas. Kemudian, hasil visualisasi LDAvis disajikan dalam format .html yang dapat memberikan kemudahan untuk mengeksplorasi hubungan antara topik dan kata-kata yang terbentuk dari topik yang ada.

1. **HASIL DAN PEMBAHASAN**
2. Hasil Pengumpulan Data

Penelitian ini mengumpulkan data diwebsite Kompas.com dengan kategori Internasional untuk menampilkan berita terkait Bola. Setelah itu, pengambilan data dilakukan secara manual, data yang penelitian ini ambil adalah isi berita secara menyeluruh., dengan mengecualikan gambar yang terdapat dalam isi berita. Data set disimpan pada file dalam format .xlsx. Total data yang diperoleh dari portal berita adalah 28 data berita mentah.

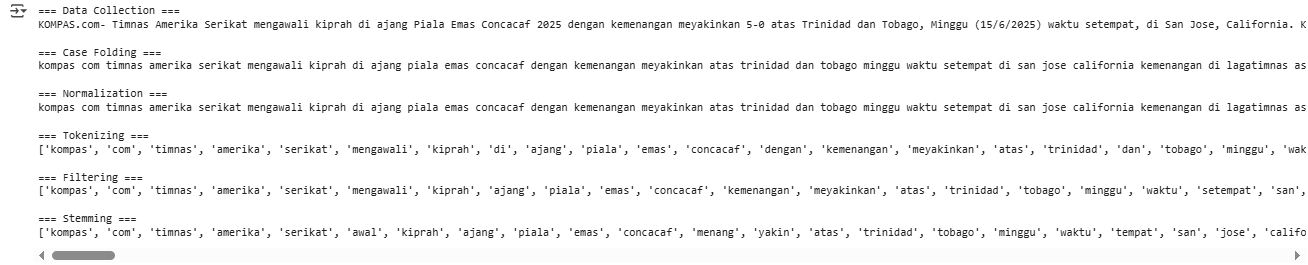


Gambar 1

1. Hasil Preprocessing Data

Pada tahapan ini, penelitian ini melakukan preprocessing data terhadap 28 data berita mentah. Untuk melakukan tahap preprocessing penelitian ini menggunakan bahasa pemrograman pada software berbasis Python. Tahapan pertama adalah case folding, digunakan untuk mengkonversi huruf kapital menjadi huruf kecil. Setelah itu, dilakukan cleaning data untuk menghapus karakter seperti tanda baca, angka, garis miring, data duplikat dan tidak konsisten. Tahapan kedua yaitu normalization, digunakan untuk mengubah kata yang tidak normal menjadi normal. Tahap selanjutnya adalah tokenizing, digunakan untuk memisahkan kata pada teks. Tahap keempat yaitu Filtering, merupakan tahap eliminasi kata hasil token untuk mewakili dokumen

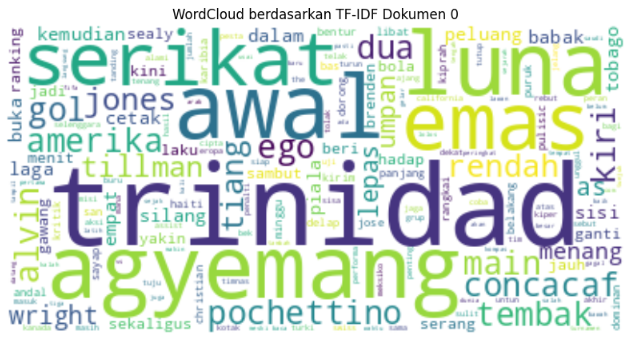
sesuai dengan stopword removal. Tahap terakhir yaitu stemming digunakan untuk mengubah kata menjadi makna dasar. Setelah data bersih, data kemudian disimpan sebagai ekstensi csv. Adapun contoh dari hasil preprocessing data berita dapat dilihat pada gambar 2.



Gambar 2

1. Hasil dari Fitur Ekstraksi TF IDF

Hasil dari proses ekstraksi fitur menggunakan TF-IDF dalam penelitian ini berupa matriks nilai bobot dari setiap kata yang muncul dalam masing-masing dokumen berita. Setiap baris pada matriks mewakili satu dokumen (berita), sedangkan setiap kolom menunjukkan kata-kata unik yang dianggap penting berdasarkan perhitungan bobot TF-IDF. Bobot yang tinggi menunjukkan bahwa kata tersebut memiliki frekuensi kemunculan yang signifikan dalam dokumen tertentu, namun tidak sering muncul di dokumen lain, sehingga dianggap sebagai kata yang mewakili topik khusus dari dokumen tersebut. Berdasarkan hasil tersebut, diperoleh kata-kata dengan bobot tertinggi yang secara kontekstual berkaitan dengan tema berita bola internasional, seperti “trinidad”, “awal”, “luna”, “agyemang”, ”emas” Kata-kata ini menunjukkan kecenderungan topik utama dalam korpus berita yang dianalisis. Dengan demikian, hasil fitur ekstraksi TF-IDF tidak hanya berfungsi sebagai input numerik untuk algoritma LDA, tetapi juga secara eksploratif memberikan gambaran awal mengenai kata-kata dominan yang membentuk struktur isi dari berita yang dikaji.

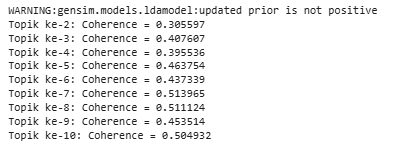


Gambar 3

Pada Tabel 3 di atas, topik yang paling banyak muncul mengenai berita bola internasional diantaranya meliputi trinidad, awal, luna, agyemang dan emas. Temuan tersebut sejalan dengan tujuan utama dari penelitian ini yaitu untuk mengetahui topik yang paling sering dibahas mengenai berita bola kategori internasional

1. Hasil Topic Modeling Menggunakan Latent Dirichlet Allocation (LDA)

Pada tahapan topic modeling menggunakan Latent Dirichlet Allocation (LDA), penelitian ini memanfaatkan library Gensim yang dirancang untuk mengekstraksi topik-topik semantik secara otomatis dari kumpulan dokumen teks. Implementasi dilakukan menggunakan bahasa pemrograman Python. Sebelum menjalankan model LDA, penelitian ini menetapkan jumlah kata per topik (words per topic) sebanyak 10 kata. Hal ini bertujuan untuk memperoleh karakteristik topik yang jelas, tidak tumpang tindih, serta dapat diinterpretasikan oleh manusia. Selanjutnya, dilakukan evaluasi model LDA berdasarkan nilai coherence, yaitu ukuran statistik yang mencerminkan keterkaitan semantik antar kata dalam satu topik. Penelitian ini mengevaluasi LDA dengan jumlah topik mulai dari 2 hingga 10, dan untuk meningkatkan stabilitas hasil, setiap model dijalankan sebanyak 50 passes. Semakin tinggi nilai coherence, maka semakin baik kualitas topik yang terbentuk dan semakin mudah pula untuk diinterpretasikan secara semantik oleh manusia. Nilai coherence dari masing-masing jumlah topik ditunjukkan pada Gambar 4.



Gambar 4

Berdasarkan informasi yang terdapat pada Tabel 3, terlihat bahwa setiap jumlah topik menghasilkan nilai coherence yang berbeda-beda. Penelitian ini mengevaluasi model LDA dengan variasi jumlah topik dari 2 hingga 10, menggunakan metode Coherence Score dengan 50 passes. Hasil evaluasi menunjukkan bahwa model dengan jumlah topik = 7 menghasilkan nilai coherence tertinggi sebesar 0.514, sehingga dipilih sebagai model terbaik. Model LDA dengan 7 topik tersebut selanjutnya dianalisis berdasarkan sekumpulan kata yang memiliki probabilitas kemunculan tertinggi dalam tiap topik. Setiap topik diberi label numerik (Topik 0 s.d. Topik 6) dan diinterpretasikan secara manual untuk mengetahui tema atau isu utama yang diwakili oleh topik tersebut. Proses interpretasi dilakukan dengan pendekatan penilaian manusia terhadap daftar kata teratas yang dihasilkan oleh model. Penelitian ini melakukan evaluasi dengan fokus pada nilai coherence tertinggi yaitu jumlah topik 7 dengan nilai coherence sebesar 0.513965. Hasil pemodelan topik dari jumlah topik 10, dapat dilihat pada gambar 4.

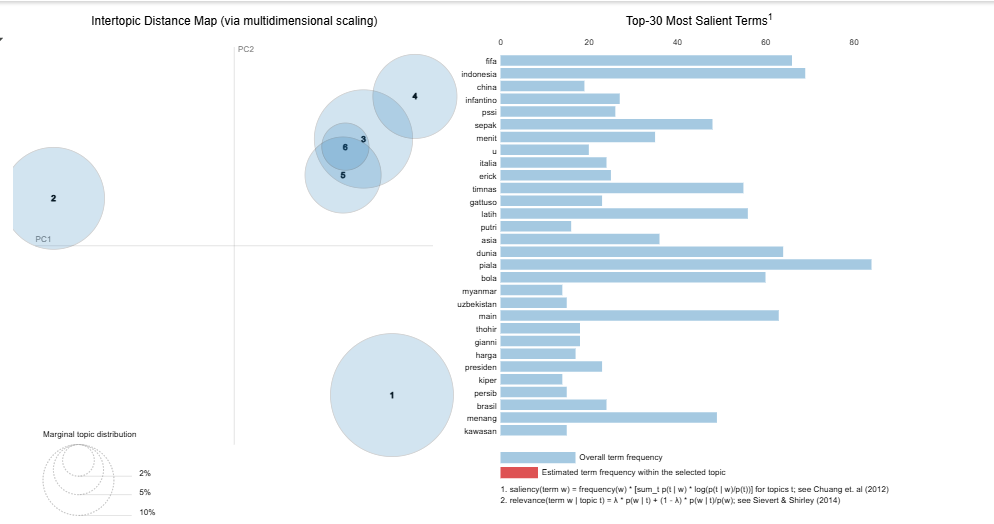
|  |  |  |
| --- | --- | --- |
| Topik | Probabilitas \* Kata | Analisis Topik (Interpretasi Manual) |
| 0 | 0.001\*"bola" + 0.001\*"fifa" + 0.001\*"sepak" + 0.001\*"indonesia" + 0.001\*"main" + 0.001\*"asia" + 0.001\*"sebut" + ... | Distribusi kata umum, tidak spesifik – kemungkinan *residual noise* |
| 1 | 0.019\*"timnas" + 0.018\*"piala" + 0.017\*"indonesia" + 0.011\*"latih" + 0.011\*"china" + 0.010\*"main" + ... | Performa Timnas Indonesia di Piala dan laga internasional |
| 2 | 0.018\*"main" + 0.011\*"jadi" + 0.010\*"persib" + 0.009\*"stegen" + 0.008\*"hasil" + 0.008\*"argentina" + ... | Hasil pertandingan sepak bola dan klub (Persib, Argentina) |
| 3 | 0.011\*"dunia" + 0.010\*"gelar" + 0.009\*"piala" + 0.009\*"amerika" + 0.008\*"serikat" + 0.007\*"liga" + ... | Turnamen Dunia dan Liga di Amerika Serikat (MLS, Copa America) |
| 4 | 0.044\*"fifa" + 0.023\*"sepak" + 0.023\*"bola" + 0.023\*"indonesia" + 0.018\*"infantino" + 0.017\*"erick" + ... | Hubungan FIFA, PSSI, dan kepemimpinan (Erick Thohir – Infantino) |
| 5 | 0.032\*"italia" + 0.031\*"gattuso" + 0.024\*"latih" + 0.014\*"timnas" + 0.012\*"baru" + 0.010\*"gennaro" + ... | Pelatih dan perubahan timnas Italia (Gattuso, Gennaro) |
| 6 | 0.018\*"menang" + 0.018\*"piala" + 0.017\*"dunia" + 0.014\*"brasil" + 0.010\*"uzbekistan" + 0.009\*"hasil" + ... | Hasil Piala Dunia U-17: kemenangan Brasil dan tim lainnya |

Tabel 1

Berdasarkan hasil evaluasi nilai coherence yang dilakukan terhadap variasi jumlah topik dari 2 hingga 10, diperoleh bahwa jumlah topik 7 menghasilkan nilai coherence tertinggi, yaitu sebesar 0.513965. Oleh karena itu, model LDA dengan 7 topik dipilih sebagai model terbaik untuk penelitian ini. Setiap topik dianalisis lebih lanjut melalui pendekatan penilaian manusia untuk mengetahui konsistensi semantik kata-kata dalam masing-masing topik. Proses interpretasi dilakukan dengan mengamati 10 kata dengan probabilitas tertinggi yang mewakili karakteristik topik. Dari hasil interpretasi tersebut, topik 1–6 menggambarkan isu-isu yang cukup jelas seperti "performa tim nasional Indonesia,", "hasil pertandingan sepak bola dunia,", "kompetisi internasional seperti Piala Dunia,", "dinamika klub dan pelatih internasional (Gattuso, Brasil),", "serta keterlibatan federasi sepak bola (FIFA, PSSI).". Namun, topik 0 menunjukkan distribusi kata-kata yang sangat umum, seperti “bola”, “fifa”, “sepak”, “main”, dan “jadi”, dengan bobot sangat rendah. Topik ini sulit diinterpretasikan karena kata-kata tersebut muncul secara merata dalam seluruh dokumen dan tidak menunjukkan hubungan semantik yang kuat. Dalam konteks penilaian manusia, topik seperti ini dianggap residual atau noise, sehingga tidak mewakili tema tertentu secara spesifik. Hasil ini menunjukkan bahwa meskipun sebuah model memiliki nilai coherence yang tinggi, kualitas topik juga sangat ditentukan oleh hubungan semantik antar kata. Oleh karena itu, evaluasi model LDA tidak hanya bergantung pada skor coherence secara kuantitatif, tetapi juga harus mempertimbangkan kemampuan interpretasi manusia terhadap topik yang dihasilkan.

1. Hasil Visualisasi Pemodelan Topik

Setelah melakukan pemodelan topik menggunakan Latent Dirichlet Allocation (LDA), selanjutnya penelitian ini melakukan visualisasi jumlah topik 7 dengan menggunakan pyLDAvis untuk mengetahui distribusi kata dari setiap topik yang dihasilkan. Gambar 7 di bawah ini menampilkan visualisasi dari pemodelan topik.



Gambar 5

Berdasarkan Gambar 5 di atas, menunjukan bahwa hasil visualisasi jumlah topik 7 terbagi menjadi dua bagian. Pada bagian kiri terdapat bentuk lingkaran yang digunakan untuk menggambarkan topik dengan menampilkan hubungan antara topik-topik tersebut. Besarnya lingkaran menunjukkan bahwa topik yang muncul secara berurutan semakin berpengaruh terhadap dokumen. Pada bagian kanan terdapat diagram batang horizontal. Diagram ini digunakan untuk menggambarkan kata-kata yang dominan dibahas dalam topik. Panjang batang mewakili frekuensi topik yang muncul dari satu kata. Bagian kiri dan kanan pada visualisasi tersebut saling terhubung satu sama lain, sederhananya bagian kiri untuk memilih sebuah topik, selanjutnya akan ditunjukkan kata-kata yang paling sering muncul oleh bagian kanan untuk menginterpretasikan topik yang dipilih sebelumnya.

Pada bagian kanan visualisasi yaitu bar chart, terdapat 30 kata (terms) penting yang menampilkan persentase kemunculan kata untuk masing-masing topik. Kata-kata yang ditampilkan diantaranya yaitu ter, stegen, Galatasaray, Barcelona, musim, sane, akan, Bayern, klub, main, kiper, Bundesliga, jerman, depan, sebut, transfer, posisi, utama, tahun, leroy, gaji, turkiye, panas, mungkin, jadi, tambah, tiga, besar, baca dan

satu. Dapat disimpulkan bahwa berdasarkan visualisasi menggunakan pyLDAvis, jumlah topik 7 menghasilkan distribusi kata yang relevan terkait berita Bola Internasional.

1. **KESIMPULAN**

Penelitian ini bertujuan untuk mengetahui topik-topik yang paling sering dibahas dalam berita Bola Internasional yang dipublikasikan di situs Kompas.com. Proses pengumpulan data dilakukan secara manual dengan mengambil 28 berita pada kategori Internasional yang secara khusus membahas isu-isu terkait sepak bola. Data yang dikumpulkan kemudian diproses lebih lanjut dengan tahapan preprocessing untuk meningkatkan kualitas teks dan memudahkan proses ekstraksi topik.

Tahapan preprocessing yang digunakan dalam penelitian ini mencakup case folding (mengubah huruf kapital menjadi huruf kecil), cleaning data (menghapus tanda baca, angka, karakter khusus, serta duplikasi), normalisasi (mengubah kata tidak baku menjadi baku), tokenisasi (memisahkan kata per kata), stopword removal (menghilangkan kata-kata umum yang tidak memiliki makna penting), serta stemming (mengubah kata menjadi bentuk dasar). Setelah proses ini selesai, data teks disimpan dalam format CSV untuk digunakan dalam tahap selanjutnya.

Untuk mengidentifikasi topik utama dari kumpulan berita tersebut, penelitian ini menerapkan metode pemodelan topik menggunakan algoritma Latent Dirichlet Allocation (LDA) dengan bantuan library gensim. Pemodelan dilakukan dengan mencoba berbagai jumlah topik (range 2–10), dan evaluasi model dilakukan menggunakan metrik coherence score untuk mengukur konsistensi semantik antar kata dalam setiap topik. Berdasarkan evaluasi tersebut, diperoleh bahwa model dengan jumlah topik = 7 menghasilkan nilai coherence tertinggi yaitu sebesar 0.513965. Nilai ini menunjukkan bahwa model tersebut paling stabil dan memiliki kualitas distribusi kata yang paling relevan secara semantik.

Selanjutnya, interpretasi terhadap masing-masing topik dilakukan menggunakan pendekatan penilaian manusia, yaitu dengan meninjau kata-kata yang paling sering muncul dalam tiap topik dan mengkaitkannya dengan konteks isi berita. Hasil interpretasi menunjukkan bahwa beberapa topik mencerminkan isu-isu nyata seperti kemenangan tim nasional Indonesia di ajang internasional, performa tim dan pemain seperti Trinidad dan Agyemang, peran negara seperti Amerika Serikat dan Brasil, serta keterlibatan federasi seperti FIFA dan PSSI. Proses interpretasi ini mendukung temuan bahwa topik-topik yang dihasilkan memang mencerminkan realitas isi berita secara konsisten.

Visualisasi topik dengan menggunakan pyLDAvis semakin memperkuat analisis. Visualisasi tersebut menampilkan hubungan antar topik dalam bentuk lingkaran (untuk menunjukkan proporsi dan keterkaitan antar topik) dan bar chart (untuk menampilkan 30 kata yang paling mendominasi tiap topik). Dari visualisasi ini terlihat bahwa topik-topik yang terbentuk memiliki distribusi kata yang berbeda dan tidak saling tumpang tindih secara signifikan, yang memperkuat kualitas hasil pemodelan.

Dengan demikian, dapat disimpulkan bahwa penggunaan metode LDA dalam penelitian ini berhasil mengidentifikasi dan mengelompokkan tema-tema utama yang sering muncul dalam berita Bola Internasional di Kompas.com. Model dengan jumlah topik 7 terbukti paling optimal secara kuantitatif (coherence score tertinggi) dan kualitatif (mudah diinterpretasikan oleh manusia). Penelitian ini juga menunjukkan bahwa coherence score dan penilaian manusia saling melengkapi dalam menilai kualitas pemodelan topik. Secara keseluruhan, pendekatan ini dapat digunakan sebagai alat bantu analisis dalam memahami kecenderungan isi pemberitaan pada media daring, khususnya dalam konteks olahraga internasional.

**Referensi**

[1] R. Aulia, J. Muhammad, and Ifwandi, “Double Multiple Jump,” *J. Ilm. Mhs. Pendidik. Jasmani, Kesehat. Dan Rekreasi*, vol. 1, no. 2, pp. 71–78, 2015.

[2] A. Nasution, “Survei Teknik Dasar Berman Sepak Bola pada Siswa Smkt Somba Opu Kabupaten Gowa,” *Ilmu Keolahrgaan*, pp. 1–10, 2018.

[3] E. Effendy, Zakaria, Azlisa, and Anggarana, “Dasar Dasar Penulisan Berita,” *J. Pendidik. dan Konseling*, vol. 5, no. 2, pp. 4042–4044, 2023, [Online]. Available: https://journal.universitaspahlawan.ac.id/index.php/jpdk/article/view/13888

[4] M. A. H. Erwan Effendy, Forsaktinahot Hasugian, “Menulis Isi Berita Dan Feature,” *J. Pendidik. dan Konseling*, vol. 4, pp. 1349–1358, 2022.

[5] J. Beno, A. . Silen, and M. Yanti, “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title,” *Braz Dent J.*, vol. 33, no. 1, pp. 1–12, 2022.

[6] E. Puspita, D. F. Shiddieq, and F. F. Roji, “Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc),” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 2, pp. 481–489, 2024, doi: 10.57152/malcom.v4i2.1204.