Analisis Interaksi Pengguna Sosial Media Sekolah di Palembang Berdasarkan Topik dengan *hLDA* dan *SVM*

Felicia[a*], Muhammad Rizky Pribadi[b]
[a,b]*Informatika, Universitas Multi Data Palembang*
Corresponding Author:
[a]*felicia2912@mhs.mdp.ac.id*

ABSTRAK
Instagram merupakan media sosial yang dapat dimanfaatkan untuk mempromosikan sekolah dengan membagikan berbagai dokumentasi kegiatan sekolah, tetapi sekolah masih kesulitan menganalisis engagement untuk mengetahui minat audience. Pengembangan perangkat lunak ini bertujuan untuk mengidentifikasi topik dari caption dan menganalisis like engagement dari setiap topik. Digunakan 3.900 data *caption* yang dikumpulkan dari lima akun *Instagram* sekolah di Palembang dengan *Instaloader*. Algoritma *hLDA* diimplementasikan untuk mengidentifikasi topik dari data *caption*, dan menghasilkan *dataset* baru yang memberi informasi topik setiap *caption*. *Dataset* ini kemudian diklasifikasikan menggunakan *SVM* dan *SVM-SMOTE*. *SMOTE* digunakan untuk mengatasi ketidakseimbangan kelas agar dapat meningkatkan hasil klasifikasi. Pada proses klasifikasi dataset dibagi menjadi 70% untuk *training* dan 30% untuk *testing*, dengan evaluasi berdasarkan *F1-Score*. Hasil terbaik diperoleh oleh *SVM-SMOTE*, dengan nilai *F1-Score* terbaik dari *Dataset hLDA* 3 *Level* (13 label), mencapai 95.68% dan nilai terendah dari *Dataset hLDA* 5 *Level* (8 label), mencapai 79.43%. *Dataset* yang memiliki lebih banyak topik memberikan hasil klasifikasi yang lebih baik. Berdasarkan jumlah like setiap topik *Dataset hLDA* 3 *Level*, yang paling diminati adalah topik 11 yang meliputi fasilitas sekolah, seragam murid, dan *event* hiburan. Informasi ini dapat membantu sekolah untuk mengembangkan lebih lanjut topik yang paling diminati serta meningkatkan topik yang kurang diminati.

Kata Kunci : *Caption, hLDA, Instagram, SVM*

ABSTRACT
Instagram is a social media that can be used to promote schools by sharing various documentation of school activities, but schools still have difficulty analyzing engagement to find out the audience's interests. This software development aims to identify topics from captions and analyze the like engagement of each topic. 3,900 caption data were collected from five school Instagram accounts in Palembang with Instaloader. The hLDA algorithm is implemented to identify topics from the caption data, and generate a new dataset that gives the topic information of each caption. This dataset was then classified using SVM and SVM-SMOTE. SMOTE is used to overcome class imbalance in order to improve classification results. In the classification process, the dataset is divided into 70% for training and 30% for testing, with evaluation based on F1-Score. The best results were obtained by SVM-SMOTE, with the best F1-Score value from hLDA 3 Level Dataset (13 labels), reaching 95.68% and

the lowest value from hLDA 5 Level Dataset (8 labels), reaching 79.43%. Datasets that have more topics give better classification results. Based on the number of likes for each topic in the hLDA 3 Level Dataset, the most popular topic is topic 11, which includes school facilities, student uniforms, and entertainment events. This information can help schools further develop the most liked topics and improve the less liked topics.

Keywords : Caption, hLDA, Instagram, SVM

PENDAHULUAN

Media sosial merupakan bentuk dari berkembangnya teknologi dan internet. Melalui media sosial, para pengguna dapat berkomunikasi dengan pengguna lainnya, mendokumentasikan dan mengupload berbagai gambar atau video, serta dapat melakukan promosi produk tertentu. Berdasarkan data yang dirilis Napoleon Cat, hingga April 2023 jumlah pengguna *Instagram* di Indonesia mencapai angka 109.33 juta pengguna. Jumlah pengguna *Instagram* di Indonesia meningkat sebesar 3.15% jika dibandingkan dengan jumlah pengguna *Instagram* pada April 2022, yaitu sebanyak 105.99 Juta (Putri Aulia Mutiara Hatla, 2023). Penggunaan Instagram di Indonesia sempat mengalami penurunan dari Juni 2022, kemudian mengalami peningkatan kembali sejak Februari 2023. Pengguna berjenis kelamin perempuan mendominasi penggunaan *Instagram* dengan persentase sebesar 53.1 %, dan 46.9 % lainnya merupakan pengguna berjenis kelamin laki – laki. Pemakaian *Instagram* di Indonesia paling banyak digunakan oleh kelompok masyarakat yang berusia 18 – 24 tahun dengan persentase sebesar 38%, dan paling sedikit digunakan oleh kelompok masyarakat berusia 45 – 54 tahun dengan persentase sebesar 5%.

Seiring semakin banyaknya pengguna *Instagram*, sekolah – sekolah di Palembang sudah mulai menggunakan *Instagram* sebagai media untuk melakukan promosi dan membantu sekolah untuk membangun citra yang baik. Melalui *Instagram* sekolah dapat membagikan berbagai foto dan video mengenai kegiatan atau *event* yang diselenggarakan, fasilitas sekolah, prestasi siswa/siswi, kegiatan ekstrakulikuler yang diadakan, seminar, informasi penerimaan murid baru, dan informasi mengenai pencapaian alumni sekolah sebagai testimoni kualitas pengajaran dari sekolah tersebut. Dengan demikian, konten – konten yang di unggah di akun *Instagram* sekolah tersebut dapat meningkatkan citra sekolah dan menarik minat calon siswa di kalangan masyarakat umum. Beberapa sekolah yang sudah menggunakan *Instagram* diantaranya adalah Sekolah Methodist 2 Palembang, SMA Xaverius 3 Palembang, Sekolah Maitreyawira Palembang, Sekolah Kusuma Bangsa, dan SMK Xaverius Palembang.

Akun *Instagram* Sekolah Maitreyawira memiliki pengikut dengan jumlah 9.040 pengikut dengan 988 postingan. Tetapi jumlah pengguna yang menyukai atau mengomentari postingan tidak sesuai dengan jumlah pengikut akun tersebut, bahkan terdapat beberapa postingan yang tidak mencapai 50 *like*. Dan jika melihat perkembangan jumlah murid Sekolah Maiteryawira Palembang dalam beberapa tahun terakhir, Sekolah Maitreyawira mengalami peningkatan jumlah murid yang tidak terlalu signifikan, dari 1.161 murid pada tahun 2019 naik menjadi 1.333 murid pada tahun 2023. Jika dibandingkan dengan sekolah kompetitor seperti Sekolah Kusuma Bangsa dan *Ignatius Global School*, peningkatan jumlah murid Sekolah

Maitreyawira terbilang rendah. Sekolah Kusuma Bangsa mengalami peningkatan jumlah murid dari 773 murid di tahun 2019 menjadi 1.409 murid di tahun 2023, sementara *Ignatius Global School* mengalami peningkatan jumlah murid dari 924 murid di tahun 2019 menjadi 2.775 murid di tahun 2023. Selain itu juga terdapat sekolah yang mengalami penurunan jumlah murid, yaitu Sekolah Methodist 2 Palembang dari jumlah murid 1.273 di tahun 2019 menjadi 1.142 di tahun 2023. Jika dilakukan perbandingan kondisi Instagram dengan sekolah swasta lainnya, postingan akun *Instagram* Sekolah Methodist 2 Palembang memiliki rata – rata *like* paling sedikit berkisar di antara 80 sampai 100 *like*, sementara untuk Sekolah Kusuma Bangsa dan *Ignatius Global School* memiliki rata – rata *like* paling sedikit berkisar di atas 100. Dari perbandingan perkembangan jumlah murid dan kondisi akun sekolah, dapat dilihat bahwa sekolah yang mengalami kenaikan jumlah murid yang signifikan cenderung memiliki interaksi pengikut yang lebih banyak.

Penelitian ini bertujuan untuk mengembangkan perangkat lunak yang dapat melakukan pemodelan topik dan melakukan analisis interaksi pengunjung *Instagram* sekolah berdasarkan jumlah like setiap topik. Pemodelan topik merupakan pendekatan *text mining* yang digunakan untuk melakukan penemuan data – data teks yang tersembunyi dan menemukan hubungan antara teks yang satu dengan yang lainnya dari korpus (Listari, 2019). Pada pengembangan perangkat lunak ini, pemodelan topik akan menggunakan algoritma *hLDA*, kemudian hasil pemodelan topik akan digunakan untuk proses klasifikasi dengan *SVM*. Pengembangan perangkat lunak ini menggunakan algoritma *hLDA* karena algoritma ini merupakan model topik hierarki yang menggunakan *nCRP* sebagai *prior*, algoritma ini memiliki kemampuan menghasilkan representasi tingkat abstraksi dan hubngan antar topik (Lindgren, 2020) yang berbeda dengan algoritma *LDA*. Dimana *LDA* sering mengabaikan korelasi kata yang dapat menyebabkan kehilangan pemodelan hubungan antar topik, sedangkan *hLDA* mampu menangkap sub-topik yang lebih spesifik yang memberikan pemahaman yang lebih baik mengenai bagaimana setiap topik berhubungan dengan topik lainnya.

Kontribusi utama penelitian ini adalah untuk mengembangkan perangkat lunak yang dapat membantu sekolah mengidentifikasi topik yang paling diminati dan kurang diminati oleh pengguna *Instagram*, sehingga sekolah dapat mengoptimalkan konten *Instagram* untuk meningkatkan *engagement* yang dapat menarik lebih banyak calon siswa.

LITERATUR REVIEW

Pada penelitian yang dilakukan oleh (Novarian Nathanael et al., 2023) mengenai *Topic Modelling* Tugas Akhir Mahasiswa Fakultas Informatika 8 Institut Teknologi Telkom Purwokerto Meggunakan Metode *Latent Dirichlet Allocation*. Penelitian ini bertujuan untuk mengetaui tren topik dengan menggunakan dataset berupa abstrak tugas akhir mahasiswa Fakultas Informatika pada perpustakan IT Telkom Purwokerto. Penelitian ini menggunakan metode *coherence value* untuk evaluasi jumlah topik dan menghasilkan delapan topik yang didukung dengan nilai *coherence* sebesar 0.446752.

Sebagai perbandingan dengan algoritma *LDA*, terdapat algoritma *DistilBERT* yang merupakan turunan dari *BERT*, yang telah menjadi s*tate-of-the-art* dalam model bahasa karena kinerja nya yang unggul dalam berbagai kasus pemrosesan bahasa alami. Penggunaan

algoritma *DistilBERT* ini dapat dilihat dalam penelitian yang dilakukan oleh (Mahesastra I Made Anditya & Darmawan I Dewa Made Bayu Atmaja, 2022), mengenai Pemodelan Topik Teks Berita Menggunakan *DistilBERT*. Penelitian ini menggunakan algoritma *DistilBERT* sebagai perbandingan pemodelan topik dengan algoritma *LDA* untuk melakukan pengelompokkan berita pada 26 portal berita berbahasa Indonesia dengan total berita sebesar 6.598 berita. Hasil *clustering* dengan metode *HDBSCAN* menghasilkan 23 kelompok berita, dan penggunaan model *embedding DistilBERT* memberikan hasil perhitungan *topic coherence* klaster dengan nilai rata – rata sebesar 0.8402955343126479. Hal ini menunjukkan bahwa penggunaan model *embedding DistilBERT* dapat memberikan hasil yang lebih unggul dibandingkan menggunakan algoritma *LDA* yang memperoleh nilai rata – rata *topic coherence* sebesar 0.8389966817692823. Melalui penelitian ini memberikan wawasan bahwa algortima *DistilBERT* dengan kemampuannya untuk menganalisis representasi kata dengan lebih kontekstual dapat menghasilkan pemodelan topik yang lebih presisi dalam konteks pengelompokkan berita berbahasa Indonesia.

Terdapat penelitian yang melakukan perbandingan anatara algoritma *BERTopic* terhadap algoritma *LDA* dalam sebuah penelitian berjudul *Comparison of Topic Modelling Approaches in the Banking Context* yang dilakukan oleh (Ogunleye et al., 2023). Penelitian ini mencermati bahwa algoritma *LDA* sebagai pendekatan pemodelan topik tradisional ini mengalami pengkritikan karena keterbatasannya dalam menangani ketersebaran data, terutama pada teks pendek. Sehingga pada penelitian ini dilakukan perbandingan hasil pemodelan topik antara algoritma *LDA* dengan algoritma *BERTopic* untuk melakukan pemodelan topik dalam konteks perbankan Nigeria. Hasil penelitian menunjukkan skor koherensi yang dihasilkan berkisar antara 0.3 – 0.65 dengan jumlah topik yang ditetapkan berkisar antara 5 – 250. Sedangkan penggunaan algoritma *BERTopic* menghasilkan koherensi sebesar 0.76 dengan 10 topik. Hal ini menunjukkan algoritma *BERTopic* memberikan kinerja yang cukup tinggi dan dapat membatasi jumlah topik, sehingga topik yang dihasilkan dapat diinterpretasikan dengan lebih baik dan memiliki istilah yang berkualitas tinggi.

Terdapat penelitian yang menggunakan algoritma *Top2Vec* yaitu penelitian yang dilakukan oleh (Ulinnuha Zahra, 2023) mengenai Penggunaan *Top2Vec* untuk Pemodelan Topik pada *Headline News* Portal Berita Berbahasa Indonesia. Penelitian ini menggunakan pemodelan topik untuk melakukan identifikasi topik berita yang paling banyak dibicarakan pada portal berita Tribunnews dan *CNN* Indonesia dari bulan Mei sampai Oktober 2022. Terdapat empat kategori berita yang digunakan, yaitu ekonomi, sosial, kesehatan, dan politik. Data yang digunakan untuk penelitian berjumlah 613 topik dari Tribunnews, dan 363 topik dari *CNN* Indonesia. Proses pemodelan topik ini mendapatkan nilai *coherence* tertinggi pada iterasi ke – 10, dimana setiap kategori memperoleh topik yang paling sering dibahas dengan nilai *coherence* sebagai berikut, kategori ekonomi dengan topik "Ancaman Resesi Ekonomi" dengan nilai *coherence* 0.772, pada kateogri sosial mengenai "Kenaikan BBM Bersubsidi" dengan nilai *coherence* 0.713, kategori kesehatan dengan topik "Gagal Ginjal Akut" dengan nilai *coherence* 0.668, dan kategori politik dengan topik "Nasdem Usung Anies" sebagai Capres dengan nilai *coherence* 0.616. Selain itu didapatkan juga hasil *Topic Information Gain* sebagai akurasi kata pada setiap topik, antara lain yaitu pada portal berita Tribunnews, kategori sosial berjumlah 4.209, ekonomi berjumlah 4.549, politik berjumlah 5.975, dan

kesehatan berjumlah 5.349. Sedangkan pada *CNN* Indonesia menghasilkan kategori sosial berjumlah 4.792, ekonomi berjumlah 5.673, politik berjumlah 4.987, dan kesehatan berjumlah 5.530.

Pada penelitian yang dilakukan oleh (Wang et al., 2021) dengan judul *Measuring Technology Complementarity Between Enterprises With an hLDA Topic Model*. Penelitian ini bertujuan untuk menyediakan *framework* untuk mengeksplorasi teknologi komplementer antar perusahaan secara kuantitatif berdasarkan data *text mining*. *Hierarchical Latent Dirichlet Allocation* digunakan mengidentifikasi topik yang tersembunyi dalam dokumen dan struktur hierarki dari topik dokumen tersebut. Penelitian ini berhasil mengidentifikasi topik – topik teknologi yang tersembunyi dalam dokumen laten dan mengukur tingkat komplementaritas antar kelas teknologi secara lebih luas dan pada subkelas di antara perusahaan – perusahaan. Meskipun demikian penelitian ini memiliki beberapa kekurangan, yaitu terdapat tingkatan topik yang sulit untuk dijelaskan karena banyak nya hal yang harus di analisis dan diinterpretasikan dan kualitas hasil klustering bergantung pada *hyperparameter*. Walaupun terdapat kekurangan tersebut, penelitian ini memberikan pemahaman teknologi komplementer di antara perusahaan melalui pendekatan *hLDA* yang memungkinkan dilakukan pengidentifikasian topik secara hierarkis.

METHODS

Dataset

Dalam penelitian ini, *dataset* yang digunakan diperoleh dengan mengumpulkan data *Caption* dari beberapa akun *Instagram* Sekolah di Palembang, yaitu Sekolah Methodist 2 Palembang, SMA Xaverius 3 Palembang, Sekolah Kusuma Bangsa, Sekolah Maitreyawira Palembang, dan SMK Xaverius Palembang. Pengumpulan data akan dilakukan dengan proses *scraping* menggunakan *library Instaloader*. Data yang diperoleh dari proses *scraping* pada periode data di tahun 2020 – 2023 adalah sebanyak 3.900 data. Setelah melalui proses *pre-processing*, data yang tersisa adalah sebanyak 3.728 data.

Algoritma *hLDA* (*Hierarchical Latent Dirichlet Allocation*)

Algoritma *Hierarchical Latent Dirichlet Allocation (hLDA)* merupakan hasil perluasan dari algoritma *LDA* yang menggunakan *nested Chinese Restaurant Process (nCRP)* untuk menghasilkan distribusi *prior* untuk mempelajari struktur pohon berdasarkan teori *Bayesian* (Wang et al., 2021). Pendekatan ini memungkinkan sturktur pohon terus menambah cabang seiring bertambahnya data. Dalam susunan hierarki topik *hLDA*, topik yang umum akan ditempatkan dibagian akar, sedangkan topik yang lebih spesifik akan ditempatkan disimpul daun. Proses generative dari algoritma *hLDA* adalah sebagai berikut :

1. Untuk setiap node $t \in T$ pada pohon :
   a. Buat sebuah topik $\beta_t \sim Dirichlet(\eta)$
2. Untuk setiap dokumen $d \in D = \{d_1, d_2, \dots, d_n\}$:
   a. Buat jalur $c_d \sim nCRP(\gamma)$
   b. Buat distribusi disetiap level pada pohon, $\theta_d \sim Dirichlet(\alpha)$
3. Untuk setiap kata
   a. Pilih level dari $Z_{d,n} \mid \theta_d \sim Discrete\ (\theta_d)$

b. Pilih kata dari $W_{d,n} \mid \{z_{d,n}, c_d, \beta\} \sim Discrete(\beta_{c_d}[z_{d,n}])$, yang diparameterisasi berdasarkan posisi $z_{d,n}$ pada jalur c_d

Dari proses *generative hLDA*, terdapat beberapa parameter yang digunakan untuk mengatur ukuran dan bentuk pohon, yaitu γ yang mempengaruhi kemungkinan dokumen memilih jalur yang baru, η yang mengatur ketersebaran topik, α merupakan parameter tingkat korpus yang mempengaruhi keragaman dalam pemilihan proporsi topik, jika α besar akan menghasilkan proporsi topik yang lebih seimbang, jika nilainya kecil akan menyebabkan proporsi yang lebih berat di topik tertentu, *L* mewakili jumlah level pada struktur pohon, dan θ merupakan vektor yang menunjukkan proporsi dari topik dalam dokumen, β berperan sebagai parameter distribusi data disetiap komponen yang menentukan bagaimana data pada topik terdistribusi dan η sebagai parameter yang digunakan untuk menentukan distribusi prior dari β sebelum melihat data.

Pemodelan topik pada perangkat lunak ini menggunakan algoritma *hLDA* karena, *hLDA* menggunakan *nCRP* sebagai *prior* yang memungkinkan pemodelan tingkat abstraksi dan hubungan hierarki antar topik, serta dapat menangkap korelasi kata yang sering diabaikan pada algoritma *LDA*. Keunggulan lain dari *hLDA* adalah dapat beradaptasi dengan data tanpa asumsi topik laten, dapat menciptakan topik baru, menetapkan kata umum ke simpul akar, dan mengurangi beban *pre-processing*.

Algoritma *SVM* (*Support Vector Machine*)

Algoritma *SVM* diperkenalkan oleh seorang Professor dari Columbia, Amerika Serikat yaitu Vladimir N Vapnik pada tahun 1992. *SVM* merupakan model *Machine Learning* kategori *Supervised Learning* yang dapat digunakan untuk menyelesaikan permasalahan klasifikasi, regresi, dan pendeteksian *outlier* (W Dadan Dahman, 2021). Tujuan dari algoritma ini adalah untuk menemukan *hyperplane* terbaik dalam ruang berdimensi *n* yang berfungsi sebagai pemisah untuk titik – titik data input, dimana ruang berdimensi *n* merupakan ruang dengan jumlah *n* fitur. Algoritma *SVM* dibagi menjadi dua jenis, yaitu *SVM Linear* digunakan pada data yang dapat dipisahkan secara *linear* dimana data dapat diklasifikasi menjadi dua kelas dengan menggunakan satu garis lurus, dan *SVM Non-Linear* digunakan pada data yang tidak dapat diklasifikasi dengan menggunakan garis lurus (Trivusi, 2022a).

Pada pengembangan perangkat lunak pemodelan topik ini, klasifikasi dengan algoritma *SVM* akan menggunakan dua *kernel*, yaitu *Linear* dan *RBF (Radial Basis Function)*. *Kernel Linear* merupakan fungsi *kernel* paling sederhana yang digunakan ketika data yang dianalisis sudah terpisah secara *linear*, sedangkan *kernel RBF* merupakan *kernel* yang paling banyak digunakan untuk klasifikasi data yang tidak dapat dipisahkan secara *linear* (Trivusi, 2022b).

Klasifikasi dilakukan dengan menggunakan algoritma *SVM*, karena *SVM* dapat bekerja secara efektif pada data yang berdimensi tinggi (memiliki banyak fitur atau label) dan dapat menggunakan subset poin pelatihan dalam *support vector* yang dapat menghemat penggunaan memori.

***SMOTE* (*Synthetic Minority Oversampling Techinique*)**

SMOTE merupakan teknik yang digunakan untuk mengatasi ketidakseimbangan kelas dalam kumpulan data dengan cara menambahkan sampel sintetik pada kelas minoritas (Octavianus

Kevin, 2023). Dengan melakukan *oversampling* data dengan metode *SMOTE*, dapat meningkatkan hasil akurasi dari proses klasifikasi pada dataset yang mengalami ketidakseimbangan kelas. Cara kerja *SMOTE* adalah sebagai berikut :

1. Mengidentifikasi kelas minoritas
2. Secara acak memilih satu titik data minoritas
3. Memilih *K-NN* dari kelas yang sama, dimana nilai default *k* adalah lima
4. Memilih satu tetangga secara acak kemudian menghitung jarak perbedaannya, lalu dikalikan dengan bilangan antara 0 – 1, hasil ini akan menjadi sampel sintetis
5. Langkah 1 – 4 akan berulang hingga jumlah data seimbang

SMOTE digunakan pada pengembangan perangkat lunak ini karena dataset yang digunakan mengalami ketidakseimbangan kelas, dimana hal ini dapat mempengaruhi hasil klasifikasi. Sehingga diperlukan penggunaan *SMOTE* yang dapat menambahkan sampel sintetik ke kelas minoritas untuk meningkatkan hasil dari proses klasifikasi.

HASIL DAN PEMBAHASAN

Pada pengembangan perangkat lunak ini, digunakan empat *dataset* yaitu *dataset labeling manual*, dan *dataset* hasil pemodelan topik *hLDA* dengan menggunakan 3, 4, dan 5 *level*. Pada semua *dataset* hasil *hLDA*, pengelompokkan akan dilakukan berdasarkan topik pada *level* kedua. Karena data yang digunakan mengalami ketidakseimbangan jumlah data di setiap topiknya, model terbaik diukur dengan melihat nilai *F1 – Score* dari hasil klasifikasi. Hal ini dikarenakan *F1 – Score* akan memberikan rata – rata harmonik dari *precision* dan *recall* yang dapat memberikan gambaran yang lebih baik mengenai kinerja model *SVM* dan *SVM-SMOTE* dengan *kernel Linear* dan *RBF*.

Dataset Labeling Manual

Pengujian ini dilakukan dengan menggunakan *dataset* yang di labeling secara manual, dengan jumlah label sebanyak 20. Topik yang didapatkan dari labeling manual adalah akademik, alumni, berita duka, dokumentasi, edukasi, ekstrakulikuler, *event*, fasilitas, *graduation*, hiburan, informasi, karya murid, kunjungan edukasi, pengumuman, peringatan, prestasi, *project*, promosi, seminar dan ucapan. Hasil evaluasi dari klasifikasi *SVM* dan *SVM-SMOTE* dengan *dataset labeling* manual dapat dilihat pada Tabel 1.

Tabel 1. Hasil Evaluasi dengan *Dataset Labeling Manual*

Model	*Kernel*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	*Linear*	65.50%	87.79%	38.16%	44.40%
SVM-SMOTE	*Linear*	97.21%	97.27%	97.11%	97.15%
SVM	*RBF*	65.50%	87.79%	38.16%	44.40%
SVM-SMOTE	*RBF*	97.21%	97.27%	37.11%	97.15%

Dataset hLDA 3 *Level*

Pengujian ini akan dilakukan dengan menggunakan *dataset* hasil pemodelan topik *hLDA* dengan 3 *level*, dimana data dikelompokkan berdasarkan *level* ke – 2 dan memiliki label sebanyak 13 label. Hasil evaluasi dari klasifikasi *SVM* dan *SVM-SMOTE* dengan *dataset hLDA* 3 *Level* dapat dilihat pada Tabel 2.

Tabel 2. Hasil Evaluasi dengan *Dataset hLDA* 3 *Level*

Model	*Kernel*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	*Linear*	55.92%	85.77%	30.00%	33.02%
SVM-SMOTE	*Linear*	95.06%	95.42%	95.18%	95.24%
SVM	*RBF*	55.92%	85.77%	30.00%	33.02%
SVM-SMOTE	*RBF*	95.06%	95.42%	95.18%	95.24%

Dataset hLDA* 4 *Level

Pengujian ini akan dilakukan dengan menggunakan *dataset* hasil pemodelan topik *hLDA* dengan 4 *level*, dimana data dikelompokkan berdasarkan *level* ke – 2 dan memiliki label sebanyak 10 label. Hasil evaluasi dari klasifikasi *SVM* dan *SVM-SMOTE* dengan *dataset hLDA* 4 *Level* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Evaluasi dengan *Dataset hLDA* 4 *Level*

Model	*Kernel*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	*Linear*	44.93%	79.27%	28.27%	30.83%
SVM-SMOTE	*Linear*	87.04%	89.97%	87.04%	87.79%
SVM	*RBF*	44.93%	79.27%	28.27%	30.83%
SVM-SMOTE	*RBF*	87.04%	89.97%	87.04%	87.79%

Dataset hLDA* 5 *Level

Pengujian ini akan dilakukan dengan menggunakan *dataset* hasil pemodelan topik *hLDA* dengan 5 *level*, dimana data dikelompokkan berdasarkan *level* ke – 2 dan memiliki label sebanyak 8 label. Hasil evaluasi dari klasifikasi *SVM* dan *SVM-SMOTE* dengan *dataset hLDA* 5 *Level* dapat dilihat pada Tabel 4.

Tabel 4. Hasil Evaluasi dengan *Dataset hLDA* 5 *Level*

Model	*Kernel*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
SVM	*Linear*	44.93%	79.27%	28.27%	30.83%
SVM-SMOTE	*Linear*	87.04%	89.97%	87.04%	87.79%
SVM	*RBF*	44.93%	79.27%	28.27%	30.83%
SVM-SMOTE	*RBF*	87.04%	89.97%	87.04%	87.79%

Analisis Hasil Klasifikasi

Dari hasil klasifikasi dapat diambil kesimpulan bahwa klasifikasi *SVM-SMOTE* memberikan hasil yang lebih baik dibandingkan hasil klasifikasi *SVM*. Dengan menggunakan *dataset* hasil pemodelan topik *hLDA*, hasil klasifikasi *SVM-SMOTE* menunjukkan nilai *F1 – Score* yang paling baik adalah hasil klasifikasi dengan *dataset hLDA* 3 *Level* dengan nilai 95.24%, kemudian hasil terbaik kedua adalah hasil klasifikasi dengan *dataset hLDA* 4 *Level* dengan nilai 87.79%, dan hasil terkecil adalah hasil klasifikasi dengan *dataset hLDA* 5 *Level* dengan niali 79.54%. Dari nilai *F1 – Score* yang diperoleh dapat dilihat bahwa semakin sedikit jumlah label maka semakin kecil nilai *F1 – Score* yang diperoleh, hal ini dapat dilihat dari nilai *F1 – Score* terbesar yaitu *hLDA* 3 *Level* memiliki label terbanyak yaitu 13 label, dan nilai *F1 – Score* terkecil yaitu *hLDA* 5 *Level* memiliki label tersedikit yaitu 8 label.

Dari proses klasifikasi dengan menggunakan empat dataset yang berbeda didapatkan nilai *F-1 Score* dari *dataset labeling manual* mendapatkan hasil paling besar, dengan selisih sekitar

1.91% dari nilai *F1 – Score dataset hLDA* 3 *Level*. Tetapi proses *labeling manual* memiliki kekurangan, yaitu membutuhkan waktu yang lebih lama dibandingka dengan menggunakan algoritma *hLDA*. Dan membutuhkan biaya yang mahal, berdasarkan informasi yang didapatkan dari *fastwork.id*, biaya jasa pelabelan data untuk 1.000 data adalah Rp 600.000. Sehingga jika menggunakan pelabelan data secara manual untuk *dataset* yang digunakan pada pengembangan perangkat lunak ini membutuhkan biaya lebih dari Rp 1.800.000, karena data yang digunakan adalah sebanyak 3.728 data.

Engagement Rate

Dataset hLDA 3 *Level* akan digunakan untuk pengembangan perangkat lunak. *Dataset* ini akan digunakan untuk mencari *engagement rate* berdasarkan jumlah *like* dari setiap topik. Dimana *engagement rate* adalah metrik dasar dalam pemasaran media sosial untuk mengukur kinerja sebuah konten pada media sosial. *Engagement rate* dapat digunakan sebagai riset untuk mengetahui kebutuhan *audience* berdasarkan jumlah interaksi dengan konten tertentu. Untuk menghitung *engagement rate* dapat menggunakan persamaan (3) (Priadana et al., 2021)

$$ER = \frac{jumlah\ like + jumlah\ komen}{jumlah\ follower} \times 1000 \quad (3)$$

Pada pengembangan perangkat lunak ini, perhitungan *engagement rate* hanya menggunakan jumlah *like*, karena pada *dataset* yang digunakan terdapat sebanyak 2.825 data atau 75.77% dari data keseluruhan postingan yang tidak memiliki komen. Setelah menghitung *engagement rate* setiap post berdasarkan jumlah *like*, dilakukan perhitungan rata – rata *engagement rate* setiap topik. Hasil perhitungan ini akan digunakan untuk mengelompokkan topik menjadi tiga bagian dengan menggunakan kuartil. Kuartil adalah konsep statistika untuk membagi data menjadi tiga kuartil, yaitu kuartil pertama (Q1) yang merupakan nilai dari 25 dan distribusi terbawah dengan 75% distribusi teratas, kuartil kedua (Q2) berada dibagian tengah yang mewakili 50% data, dan kuartil ketiga (Q3) merupakan nilai 25% dari distribusi teratas dan 75% dari distribusi terbawah (Haris Moch & Feriyanti Nentin, 2023). *High Engagement* akan diambil dari kuartil pertama, *Low Engagement* akan diambil dari nilai kuartil ketiga, dan *Medium Engagement* diambil dari kuartil kedua. Dari pengelompokkan ini akan didapatkan informasi mengenai topik yang paling diminati dan kurang diminati. Hasil pengelompokkan topik berdasarkan nilai rata – rata *engagement rate* setiap topik dapat dilihat pada Tabel 5.

Tabel 5. Pengelompokkan Topik Berdasarkan Rata – Rata *Engagement Rate*

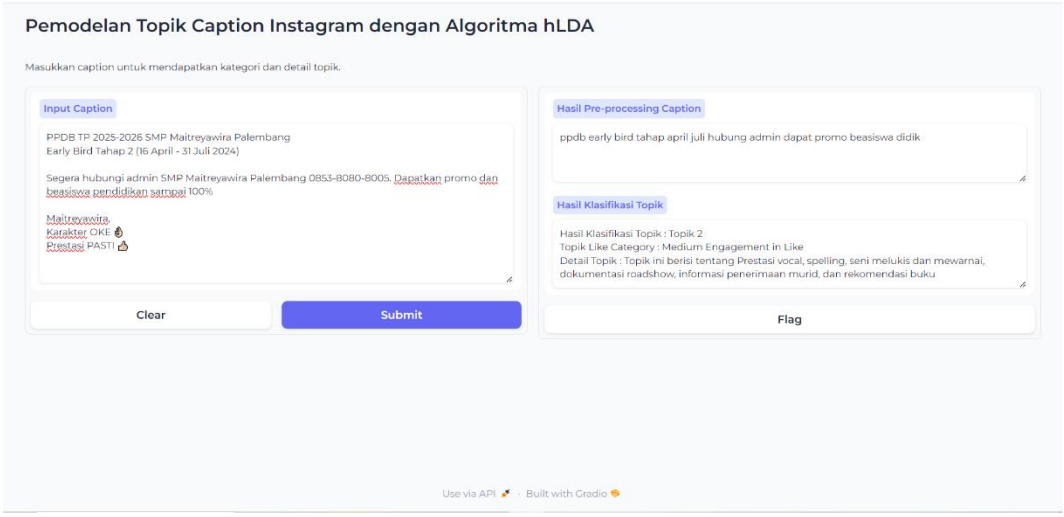
Nama Topik	Rata – Rata *Engagement Rate*	Kategori Topik
Topik 11	137.654	*High Engagement in Like*
Topik 12	36.951	*High Engagement in Like*
Topik 13	27.784	*High Engagement in Like*
Topik 1	25.583	*Medium Engagement in Like*
Topik 9	25.436	*Medium Engagement in Like*
Topik 5	25.383	*Medium Engagement in Like*
Topik 3	24.636	*Medium Engagement in Like*
Topik 10	23.579	*Medium Engagement in Like*

Topik 6	23.204	*Medium Engagement in Like*
Topik 8	22.939	*Low Engagement in Like*
Topik 7	21.230	*Low Engagement in Like*
Topik 2	21.055	*Low Engagement in Like*
Topik 4	20.460	*Low Engagement in Like*

Berdasarkan informasi yang didapatkan dari Tabel 5, dapat dilihat topik yang paling diminati merupakan Topik 11 yang berisi *caption* postingan tentang konten sekolah (fasilitas dan seragam) dan *event* hiburan sekolah. Sedangkan topik yang kurang diminati adalah Topik 4 berisi *caption* postingan tentang informasi lowongan kerja, perayaan keagamaan, dan HUT RI, informasi kegiatan *online* dan *project video*, *graduation*, dan edukasi kesehatan.

Perangkat Lunak Pemodelan Topik

Tampilan perangkat lunak pemodelan topik dengan algoritma *hLDA* dapat dilihat pada Gambar 1.



Gambar 1. Tampilan Perangkat Lunak Pemodelan Topik

Pada perangkat lunak pemodelan topik ini, *user* dapat memasukkan *caption* yang ingin diketahui topik dan *like engagement* nya pada kolom *input caption*. Setelah memasukkan *caption*, *user* dapat mengklik *button submit* untuk memproses *caption* yang di *input*. Perangkat lunak akan menampilkan informasi berupa hasil *pre – processing caption*, topik dari *caption*, topik *like category*, dan detail topik.

SIMPULAN

Kesimpulan dari pengembangan perangkat lunak ini menunjukkan bahwa *hLDA* dapat memodelkan topik dengan baik. Penggunaan *SVM-SMOTE* pada klasifikasi memberikan performa yang baik dengan nilai *F1-Score* tertinggi 95.24% dengan menggunakan *dataset hLDA* 3 *Level*, dan nilai terendah 79.54% dengan *dataset hLDA* 5 *Level*. Pada proses klasifikasi, dataset dengan label yang lebih sedikit menghasilkan nilai yang lebih rendah dibandingkan dataset dengan label yang lebih banyak. Dan dari pengelompokkan topik berdasarkan nilai rata – rata *engagement rate* didapatkan topik yang memiliki nilai terbesar,

yaitu Topik 11 yang berisi postingan yang memperkenalkan sekolah, seperti fasilitas dan seragam sekolah serta *event* hiburan yang diadakan di sekolah.

DAFTAR PUSTAKA

Haris Moch, & Feriyanti Nentin. (2023). Mengoptimalkan Belanja Operasional di Badan Perencanaan Eselon I terhadap Total Pagu Belanja Guna Mewujudkan Organisasi yang Kuat dan Profesional.

Lindgren, J. (2020). Evaluating Hierarchical LDA Topic Models for Article Categorization. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1447656&dswid=3417

Listari. (2019, June 21). Topic Modeling Menggunakan Latent Dirchlect Allocation (Part 1): Pre-processing Data dengan Python. https://medium.com/@listari.tari/topic-modelling-menggunakan-latent-dirchlect-allocation-part-1-pre-processing-data-dengan-python-87bf5c580923

Mahesastra I Made Anditya, & Darmawan I Dewa Made Bayu Atmaja. (2022). Pemodelan Topik Teks Berita Menggunakan DistilBERT. https://jurnal.harianregional.com/jnatia/id-92840

Novarian Nathanael, Khomsah Siti, & Arifa Amalia Beladinna. (2023). Topic Modelling Tugas Akhir Mahasiswa Fakultas Informatika Institut Teknologi Telkom Purwokerto Meggunakan Metode Latent Dirichlet Allocation. https://journal.ittelkom-pwt.ac.id/index.php/ledger/article/view/991

Octavianus Kevin. (2023, June 18). Oversampling Method SMOTE for Imbalanced Data.

Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. Applied Sciences, 13(2), 797. https://doi.org/10.3390/app13020797

Priadana, A., Saputra, A. B., Cahyo, P. W., & Habibi, M. (2021). Health in Digital Era 4.0: Analyzing Reader Engagement Rate on Instagram Account of Government Health Agencies. Proceedings of the International Conference on Health and Medical Sciences (AHMS 2020). https://doi.org/10.2991/ahsr.k.210127.057

Putri Aulia Mutiara Hatla. (2023, May 22). Instagram Down, 10 Warga Negara Ini Jadi Gak Bisa Eksis. Instagram Down, 10 Warga Negara Ini Jadi Gak Bisa Eksis

Trivusi. (2022a, July 3). Penjelasan Lengkap Algoritma Support Vector Machine (SVM).

Trivusi. (2022b, July 4). Apa itu Kernel Trick? Pengertian dan Jenis-jenis Fungsi Kernel SVM.

Ulinnuha Zahra. (2023). Penggunaan Top2vec Untuk Pemodelan Topik Pada Headline News Portal Berita Berbahasa Indonesia. https://repository.uinjkt.ac.id/dspace/handle/123456789/73325

W Dadan Dahman. (2021, July 12). Support Vector Machine (SVM). https://medium.com/sysinfo/support-vector-machine-svm-5d95a7d7a547

Wang, X., Qiao, Y., Hou, Y., Zhang, S., & Han, X. (2021). Measuring technology complementarity between enterprises with an hlda topic model. IEEE Transactions on Engineering Management, 68(5), 1309–1320. https://doi.org/10.1109/TEM.2019.2958113