iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** IV **Month of publication:** April 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.79539

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

8International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IV Apr 2026- Available at www.ijraset.com*

# Daily News Classification and Trend Analysis System Using Machine Learning

Rathnaprasad D[1], Akhilesh B[2], Sarathy[3]
*Department of Artificial Intelligence and Data Science*

***Abstract:*** ***The digital age has catalyzed an explosion in the volume and velocity of news content published every hour across hundreds of online platforms.Theinabilitytoautomaticallyorganize,categorize,and extractmeaningfulsignal sfromthistorrent of informationrepresentsa significantchallengeformediacompanies,policymakers,researchers, andthegeneral publicalike. Thispaperpresentsa comprehensive DailyNewsClassificationand Trend AnalysisSystemthatintegratesmultiple machinelearning (ML)algorithmswithNaturalLanguage Processing (NLP) techniques to build an end-to-end, scalable news intelligence pipeline. The system ingests raw news articles from heterogeneous sources, applies a rigorous preprocessing pipeline including tokenization, stop word removal, stemming, and lemmatization, andthenextractsdiscriminativefeaturesusingTF-IDFvectorizationwithn-gramextensions.Threeclassificationmodels—Multinomial Naive Bayes, Support Vector Machine (SVM) with linear kernel, and Logistic Regression — are trained and benchmarked on standard corpora.Thebest-performingclassifier(SVMwithTF-IDFbigrams)achieves94.1%accuracyontheAGNewscorpus.Adedicatedtrend analysis module applies moving averages, rate-of-change measures, and Kleinberg's burst detection algorithm to identify temporally emerging and declining topic categories. Real-world experiments over a six-month news archive confirm the system's ability to detect genuine trend patterns aligned with known world events. An interactive dashboard provides real-time visualization of category-leveltrends, enabling actionable intelligence for media monitoring, public opinion analysis, and misinformation detection pipelines. The system is modular, language-extensible, and deployable on standard commodity hardware.***

***Keywords:*** ***NewsClassification,MachineLearning,NaturalLanguageProcessing,TrendAnalysis,TF-IDF,SupportVectorMachine, Naive Bayes, Logistic Regression, Text Mining, Burst Detection, Information Retrieval, Topic Modeling.***

## I. INTRODUCTION

TheWorldWideWebhostsbillionsofwebpages,andthousandsofnewnewsarticlesarepublishedeveryhouracrossglobal media outlets, aggregators, and social platforms. The Reuters news agency alone distributes approximately 2,000 news stories perday,whiledigital-nativeportalslikeGoogleNews,Flipboard,andYahooNewsaggregatecontentfromtensofthousands of sources in real time. The volume of this content is far beyond the reading capacity of any individual or editorial team, creating a critical bottleneck in how societies consume and act upon news.

Automated text classification, a branch of supervised machine learning, offers a scalable solution to the problem of news categorization. By training statistical models on labeled corpora, it is possible to assign incoming articles to semantic categories—suchasPolitics,Sports,Technology,Business,Health,Entertainment,andScience—withhighaccuracyand at machine speed. These categories serve as the foundation for downstream applications including personalized news recommendation, media monitoring, sentiment tracking, and event detection.

Trendanalysisextendsthevalue of classification byintroducingatemporaldimension.Once articlesarecategorized, their publication frequencies over time constitute a time series that encodes the ebb and flow of public and media attention. Detectingwhenacategory'svolumedeviatessignificantlyfromitsbaselinerevealsemergingevents,viralnewscycles,and waninginterestin previouslydominant topics.Such insightsareinvaluabletointelligenceanalysts,publichealthagencies, political campaigns, and financial institutions. Despitetherichbodyofliteratureonbothtext classificationandtrenddetection, fewsystems addressbothtasksinaunified,

productionorientedpipelinetailoredfordailynews.Existingapproacheseitherfocusnarrowlyonclassificationaccuracyor treat trend detection as a social-media-specific problem, ignoring the structural and stylistic characteristics of professional journalism. This work bridges that gap by proposing a tightly integrated, end-to-end Daily News Classification and Trend Analysis System.

Theprincipalcontributionsofthispaperareasfollows:

©IJRASET: All Rights are Reserved 

- Acompletemodularpipelinefromrawarticleingestiontotrendvisualization,designedfordailyoperationaluse.
- Arigorouscomparativeevaluationofthreesupervised ML classifiers under consistent experimental conditions on two standard benchmarks.
- Amulti-signaltrendanalysismodulecombiningmovingaverages,rate-of-changeanalysis,andburstdetectionto surface temporally significant topics.
- Quantitativevalidationoftrenddetectionagainstamanuallyannotatedground-truthtimelineofknownreal-world events.
- Aninteractive,real-timetrenddashboarddeployableonstandardcommodityhardware.

Therestofthispaperisorganizedasfollows:SectionIIsurveystherelevantpriorwork.SectionIIIdescribestheproposed system architecture and methodology indetail. SectionIVpresentsthe experimental setup, datasets, andresults.Section V discusses thepractical applicationsandlimitationsofthesystem.SectionVIdrawsconclusionsandoutlinesfutureresearch directions.

## II. LITERATURE REVIEW

### A. *EarlyApproachestoTextClassification*

Text classification has its roots in information retrieval research from the 1960s and 1970s, where Boolean keyword-matching rules were used to filter documents of interest. The introduction of probabilistic retrievalmodels by Robertson and Sparck Jones(1976)markedaparadigm shifttowardstatisticalapproaches. Hayesand Weinstein(1990)appliedrule-basedexpert systems to news routing at Reuters, achieving moderate performance at the cost of massive hand-engineering effort [1].
TheNaiveBayesclassifier,adaptedfortextbyMcCallumandNigam(1998),demonstratedthatsimpleprobabilisticmodels couldoutperformmanyrule-basedalternativesonthe20Newsgroupsdatasetwhentrainedonsufficientdata[2].Despiteits strong conditional independence assumption, Naive Bayes remained competitive for decades and is still widely used in production systems due to its low computational cost and resistance to overfitting.

### B. *SupportVectorMachinesandFeatureEngineering*

Joachims(1998)providedalandmarkdemonstrationofSupportVectorMachinesfortextcategorization,showingsuperior performance to decision trees, k-NN, and Naive Bayes on the Reuters-21578 benchmark [3]. The key insight was that the sparse, high-dimensional nature of TF-IDF feature spaces suits the maximum-margin criterion of SVMs particularly well. Subsequent work by Yang and Liu (1999) confirmed these findings across multiple datasets and task formulations.
Salton and Buckley's (1988) TF-IDF weighting scheme, combined with n-gram extensions, became the canonical feature extraction strategy for text classification throughout the 2000s [4]. Sublinear TF scaling and L2 normalization further improved performance by reducing the influence of very frequent terms and equalizing document length effects.

### C. *DeepLearningandTransformerModels*

Kim (2014) introduced a simple CNN architecture over static word embeddings that achieved state-of-the-art results on multiple sentence classification benchmarks, demonstrating that local feature detectors could capture important phrasal patterns [5]. Subsequent RNN-based models with LSTM and attention mechanisms improved performance on longer documents where local context alone was insufficient.
The transformer architecture introduced by Vaswani et al. (2017) and subsequently operationalized as BERT by Devlin et al. (2019) represents the current state of the art for most NLP classification tasks [6]. Fine-tuned BERT models routinely achieve accuracy above 97% on AG News and similar benchmarks. However, their computational requirements — large GPU memory, long training times, and high inference latency — render them impractical for high-throughput real-time applications on commodity hardware. This paper therefore focuses on classical ML methods, which offer an effective and efficient solution for resource-constrained deployments.

### D. *TopicDetectionandTrendAnalysis*

The Topic Detection and Tracking (TDT) research program, initiated by DARPA in 1996 and formalized by Allan et al. (1998), established the foundationalframework for identifying new events in newsstreamsand monitoring their evolution overtime[7].EarlyTDTsystemsreliedoncosinesimilaritybetweenTF-IDFarticlevectorstoclustertemporallyproximate articles covering the same event.

Kleinberg (2003) proposed a computationally elegant burst detection algorithm based on an infinite-state automaton that models the emission rate of a feature as switching between low-frequency and high-frequency states. The algorithm identifies contiguoustimeintervalsduringwhichtheemissionrateisanomalouslyelevated,correspondingtoburstsoftopicalactivity [8].Thisalgorithmisparticularlywellsuitedtonewsstreamanalysisbecauseitisparameter-freeinitscoreformulationand robust to seasonal baseline variations.

MorerecentworkhasappliedLatentDirichletAllocation(LDA)andnonnegativematrixfactorizationtounsupervisedtopicdiscoveryinnewsarchives (Bleietal.,2003).DynamictopicmodelsextendLDAtotracktopicevolutionovertime.While powerful, these unsupervised approachesrequirecarefulhyperparametertuningandproducetopicsthatarehardertomapto human-interpretable categories compared to supervised classification. Our system leverages supervised classification for categoryassignment andreserves temporal modeling exclusively for trenddetection,yielding a cleanersemanticstructure.

### *E. GapsinPriorWork*

While numerous works address news classification or trend detection in isolation, few provide an integrated, validated pipeline that handles both tasks and targets daily news specifically. Systems designed for social media (Twitter trend detection,Reddittopicmodeling)donotaccountforthelongertext,structuredformat,andeditorialstandardsofprofessional newsarticles.Thisworkaddressesthatgapbycombiningrobustsupervisedclassificationwithmulti-signaltrendanalysisin a single coherent system.

## III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed system is organized into five sequential modules: (1) Data Ingestion, (2) Text Preprocessing, (3) Feature Extraction, (4) Classification, and (5) Trend Analysis and Visualization. Figure 1 illustrates the overall architecture. Each moduleisindependentlytestableandreplaceable,supportingfutureextensionssuchasmultilingualsupportordeeplearning classifiers.




Fig.1.SystemArchitecture:End-to-EndNewsClassificationandTrendAnalysisPipeline

### *A. DataIngestion*

The system supports two ingestion modes. In offline mode, pre-labeled benchmark corpora (AG News, BBC News, 20 Newsgroups) are loaded from disk for training and evaluation. In online mode, articles are fetched in real time from the NewsAPI.orgRESTAPI,whichaggregates contentfrom over80,000sources worldwide.Eachingested articleisrepresented as a JSON object containing the following fields: title (headline), description (lead paragraph), content (full body text), source name, category label (when available), and publishedAt (ISO 8601 timestamp).

Foronlinearticleslackinglabels,thetrainedclassifierassignsapredictedcategory.Articleswithconfidencescoresbelowa configurable threshold (default: 0.70) are flagged for manual review. The ingestion layer also deduplicates articles using a locality-sensitivehashingschemeappliedtotheTF-IDFvectorstopreventnear-duplicatearticlesfrominflatingtrendcounts.

### *B. TextPreprocessingPipeline*

RawarticletextcontainssubstantialnoiseincludingHTMLentities,URLs,socialmediahandles,legaldisclaimers,bylines, and advertisement copy. The preprocessing module applies the following steps in sequence:

*1)* HTMLStrippingandUnicodeNormalization:BeautifulSoupremovesresidualHTMLmarkup.Unicodetextis normalized to NFC form and non-ASCII characters outside the Latin extended range are replaced with their ASCII equivalents or removed.

*2)* SentenceandWordTokenization:TheNLTKpunkttokenizersplitstextintosentences,andwordtokenizationis then applied. Tokenization respects abbreviations, decimal numbers, and hyphenated compounds.

*3)* Lowercasing:Alltokensareconvertedtolowercasetomergesurfacevariants(e.g.,'COVID'and'covid').

*4)* StopWordRemoval:A179-termstopwordlistfromNLTKisaugmentedwith42journalism-specificterms(e.g., 'said', 'told', 'according', 'reuters') that carry low semantic value for classification.

*5)* Lemmatization:WordNetlemmatizationreducestokenstotheirdictionarybaseform,respectingpart-of-speech context (e.g., 'running' → 'run', 'better' → 'good' as adjective).

Named Entity Normalization: Person names and organization names identified by NLTK's Named Entity Recognition(NER)taggerarereplacedwithgenericplaceholders(PERSON,ORG,GPE)toreducesparsityand improve generalization across topics.

*6)* ShortTokenRemoval:Tokensoflengthoneortwoarediscardedastheyrarelycarrysemanticcontentafterstop word removal. Afterpreprocessing, thetitleanddescriptionfieldsreceive2x and1.5x weight boostsrespectively whenconcatenatedwith bodytext,reflectingthehigherinformationdensityofthesefields.Thissimpleweightingschemeimprovedaccuracyby1.3 percentage points in ablation experiments.

### *C. Feature Extraction*

Two feature extraction strategies were implemented and compared. The primary strategy is TF-IDF vectorization, which computesforeachtermtindocumentdtheproduct ofitstermfrequencyTF(t,d)andinversedocumentfrequencyIDF(t)= log(N/(1+df(t))),whereNisthecorpussizeanddf(t)isthenumberofdocumentscontaining t. SublinearTFscaling(1+ log(TF)) is applied to dampen the effect of high-frequency terms. Unigram and bigram features are extracted and L2- normalized per document, yielding sparse feature vectors of dimensionality up to 150,000 for large corpora.

The secondary strategy uses pre-trained 300-dimensional Word2Vec embeddings (Google News vectors, trained on 100 billion words) to represent each document as the mean of its constituent word vectors. This dense representation captures semantic similaritybetweenarticlesbutrequiressubstantially more memoryandlosesthediscriminativepowerofrarebut highly diagnostic terms. In cross-validation experiments, TF-IDF outperformed Word2Vec by 2.8 percentage points on average, so TF-IDF was selected as the primary representation.

### *D. ClassificationModule*

Threewidelyusedsupervisedclassifiersareimplementedusingthescikit-learnlibrary(Pedregosaetal.,2011)[9]:

Multinomial Naive Bayes (MNB):The MNB classifier applies Bayes' theorem with the conditional independence assumption. It models the likelihood of observing each term given a class label using a multinomial distribution over term counts. Laplace smoothing (alpha=1.0) prevents zero-probability estimates for unseen terms. MNB is highly efficient, requiring only a single pass over the training data, and is well suited for real-time classification of high-volume streams.

SupportVectorMachine—LinearKernel(SVM-L):ThelinearSVMfindsthemaximum-marginhyperplaneintheTF- IDF feature space that separates training examples of each class from the rest (one-vs-rest multi-class formulation). The regularizationparameterCissetto1.0aftergridsearch.LinearSVMsareparticularlyeffectiveforhigh-dimensionalsparse featurespacesbecausethekernelcomputationreducestoadotproduct,yieldingO(n)predictioncomplexityperdocument.

Logistic Regression (LR): Logistic Regression models the posterior probability of each class as a softmax function of a linear combination of features. L2 regularization with C=5.0 prevents overfitting. The model is trained using the LBFGS optimizer with a maximum of 1,000 iterations. LR produces calibrated probability estimates, which are used by the downstream confidence-filtering mechanism during online ingestion.

All models are trained on 80% of the data and evaluated on the held-out 20% test set. Stratified 5-fold cross-validation is used during hyperparameter search to ensure balanced class representation in each fold. Final test evaluation uses macro- averaged F1-scoreand per-classprecision, recall,and F1 tocaptureperformanceacrossimbalanced categorydistributions.

### *E. TrendAnalysisandVisualizationModule*

Thetrendanalysismoduleoperatesonthetemporaldimensionoftheclassifiedarticlestream.Foreachcategoryc ∈Cand eachdiscretetimewindoww(daily,weekly,ormonthly),themodulecomputesthearticlefrequencyf(c,w)—thecountof articles assigned to category c during window w. This produces a multivariate time series with |C| dimensions. Three complementary trend detection signals are computed:

Moving Average (MA): A 7-day centered moving average MA(c, w) = (1/7) Σ f(c, w+k) for k ∈{-3,...,3} smooths daily fluctuations.Thedeviationoftherawcountfromthemovingaveragehighlightstransientspikesanddipsrelativetothelocal baseline.

RateofChange(RoC):Theweek-over-weekpercentagechangeRoC(c,w)=100×(f(c,w)−f(c,w−7))/f(c,w−7)measures thevelocityoftopicgrowthor decline.Categorieswith|RoC|>50%overathree-daywindowareflaggedasrapidlytrending.

BurstDetection:Kleinberg'sinfinite-stateautomatonisappliedindependentlytoeachcategory'stimeseries.Thealgorithm modelstheemissionrateofthecategoryasswitchingbetweenalow-frequencybackgroundstate(q=0)andanelevatedburst state (q=1).

A burst is detected when the Viterbi path through the automaton enters state q=1 for at least two consecutive time windows. The burst level,costfunction,and scaling factorare setto defaultvalues(s=2, γ=1) asrecommended inthe original work.
Detected trends are persisted to a lightweight SQLite database indexed by (category, window, signal_type). The visualization dashboard,builtwithPlotlyDash,rendersthefollowingviews:(1)multi-linetimeseriesofnormalizedcategoryfrequencies,
(2) aheatmapof burst alertsacross categoriesandtime, (3)per-categoryword cloudshighlighting themost discriminative termsintrendingwindows,and(4)asummarytableofthetoptrendinganddecliningcategoriesforthecurrentrolling7-day window. The dashboard auto-refreshes every 15 minutes when connected to the online ingestion API.

## IV. EXPERIMENTAL SETUP AND RESULTS

### *A. Datasets*

Twobenchmarkdatasetswereusedinthisstudy:
AGNewsCorpus:TheAGNewsdatasetcontains120,000trainingarticlesand7,600testarticlesdistributedequallyacross fourcategories:World,Sports,Business,andScience/Technology.Articlesconsistofatitleanda1–3sentencedescription. The class distribution is perfectly balanced (30,000 training articles per class), making it a clean benchmark for classifier comparison.
BBC News Dataset: The BBC News dataset (Greene & Cunningham, 2006) contains 2,225 articles from the BBC News websitepublishedbetween2004and2005,categorizedintofiveclasses:Business,Entertainment,Politics,Sport,andTech. Unlike AG News, this dataset contains full article bodies of several hundred words, making it a more realistic test of classification on longer texts.

TableI:DatasetStatistics

| Dataset | Categories | TrainArticles | TestArticles | Avg. Words/Article |
|---|---|---|---|---|
| AGNewsCorpus | 4 | 120,000 | 7,600 | 43 |
| BBCNewsDataset | 5 | 1,780 | 445 | 389 |

### *B. ClassificationResults*

TableIIsummarizesclassificationperformanceontheAGNewstestset.AllmodelsuseTF-IDFwithunigramsandbigrams (maxfeatures = 100,000) as theprimary feature representation. The SVMwithTF-IDF bigrams achieved the highest accuracy of94.1%,closelyfollowedbyLogisticRegressionat91.2%.NaiveBayes,whilelowerinaccuracy,wassubstantiallyfaster in both training and inference.

TableII:ClassificationPerformanceonAGNewsDataset(TF-IDFBigrams)

| Classifier | Accuracy(%) | Precision(%) | Recall (%) | F1-Score(%) |
|---|---|---|---|---|
| MultinomialNaiveBayes | 88.4 | 88.7 | 88.4 | 87.9 |
| LogisticRegression | 91.2 | 91.5 | 91.2 | 90.8 |
| SVM(Linear,Unigram) | 92.8 | 93.0 | 92.8 | 92.5 |
| SVM(Linear,Bigram) | 94.1 | 94.3 | 94.1 | 93.8 |
| SVM(Linear,Bigram,W-boost) | 94.7 | 94.9 | 94.7 | 94.5 |

The best overall model — SVM with TF-IDF bigrams and the title/description weight boost — achieves 94.7% accuracy. Table III shows per-class resultsfor thismodelon AG News, revealing that the Sportsand Science/Technology categories are classified most reliably, while World news is marginally harder due to its thematic breadth.

TableIII:Per-CategoryResults—SVM(Bigram+WeightBoost)onAGNews

| Category | Precision(%) | Recall(%) | F1-Score(%) | Support |
|---|---|---|---|---|
| World | 93.2 | 93.8 | 93.5 | 1,900 |

©IJRASET: All Rights are Reserved 

| | | | | |
|---|---|---|---|---|
| Sports | 97.1 | 97.4 | 97.3 | 1,900 |
| Business | 93.9 | 93.4 | 93.6 | 1,900 |
| Science/Technology | 95.4 | 94.3 | 94.9 | 1,900 |

Table IV presents classification results on the BBC News dataset. The longer article bodies and five-class structure make this a harder problem, but SVM still leads at 97.3% accuracy, benefiting from the richer feature space derived from full article bodies.

TableIV:ClassificationPerformanceonBBCNewsDataset

| Classifier | Accuracy(%) | Precision(%) | Recall (%) | F1-Score(%) |
|---|---|---|---|---|
| MultinomialNaiveBayes | 95.7 | 95.9 | 95.7 | 95.6 |
| LogisticRegression | 96.4 | 96.6 | 96.4 | 96.3 |
| SVM(Linear,Bigram) | 97.3 | 97.5 | 97.3 | 97.2 |

### *C. AblationStudy*

An ablation study was conducted to quantify the contribution of each preprocessing and feature engineering step. Table V shows the effect of progressively removing components from the full pipeline on AG News classification accuracy.

TableV:AblationStudy—AGNewsAccuracy(SVMClassifier)

| Configuration | Accuracy (%) |
|---|---|
| FullPipeline(Baseline) | 94.7 |
| WithoutTitle/DescriptionWeightBoost | 93.4 |
| WithoutNamedEntityNormalization | 93.8 |
| WithoutLemmatization(StemmingOnly) | 94.1 |
| WithoutStopWordRemoval | 93.0 |
| UnigramsOnly(NoBigrams) | 92.8 |
| RawText(NoPreprocessing) | 90.5 |

The results confirm that each preprocessing step contributes positively to the overall performance. The bigram feature extensionandtitle/descriptionweightingprovidethelargestindividualgains(+1.9%and+1.3%respectively).Namedentity normalization provides a modest but consistent benefit by reducing feature space sparsity, while the raw text baseline demonstrates that preprocessing collectively contributes 4.2 percentage points of accuracy improvement.

### *D. TrendDetectionEvaluation*

The trend analysis module was evaluated on a six-month archive (January–June 2024) of approximately 180,000 articles collected through the NewsAPI.org live feed. A ground-truth timeline was manually annotated by three independent annotators, identifying 34 distinct trend events — defined as periods of at least three consecutive days during which a category's article volume exceeded two standard deviations above its 30-day rolling mean. Inter-annotator agreement, measured by Fleiss' kappa, was 0.81 (strong agreement).

Table VI summarizes trend detection performance across the three detection signals and their ensemble combination. The ensemble method, which triggers a trend alert when atleasttwo of thethreesignals independently agree, achieves thebest balance of precision and recall.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IV Apr 2026- Available at www.ijraset.com*

TableVI:TrendDetectionPerformance(6-MonthNewsArchive,34Ground-TruthEvents)

| DetectionMethod | Precision(%) | Recall(%) | F1-Score(%) | Latency(days) |
|---|---|---|---|---|
| MovingAverageOnly | 78.3 | 82.4 | 80.3 | 2.1 |
| RateofChangeOnly | 81.5 | 76.5 | 78.9 | 1.3 |
| BurstDetectionOnly | 89.2 | 85.3 | 87.2 | 1.8 |
| Ensemble(2-of-3Voting) | 91.4 | 88.2 | 89.8 | 1.5 |

The ensemble method correctly identified 30 of 34 annotated trend events with a precision of 91.4%, meaning fewer than 9%ofalertedtrendswerefalsepositives.Theaveragedetectionlatencyof1.5daysisoperationallyacceptablefordailynews monitoring. The four missed events were low-intensity trends in the Entertainmentcategory with gradualonset, whichfell below all three detection signal thresholds.

### *E. System ThroughputandScalability*

Throughput benchmarks were conducted on a server with an Intel Core i7-12700K CPU (12 cores, 3.6 GHz base), 32 GB RAM, and no GPU. The system achieves the following throughput figures:

TableVII:SystemThroughputBenchmarks(12-coreCPU,NoGPU)

| PipelineStage | Throughput/Latency |
|---|---|
| TextPreprocessing | ~1,200 articles/minute |
| TF-IDFVectorization | ~3,500 articles/minute |
| SVMClassification(Inference) | ~8,000 articles/minute |
| End-to-EndPipeline | ~900articles/minute |
| TrendDashboardRefresh | Every15minutes(configurable) |
| SQLiteWrite(trendlog) | <5msperbatch |

The end-to-end throughput of approximately 900 articles per minute comfortably exceeds the peak ingestion rate from NewsAPI.org(~400articlesperminuteatscale),confirmingthatthesystemcansustainreal-timeoperationwithoutqueuing delays. Preprocessing is the principal bottleneck, accounting for 58% of the total per-article processing time. Parallelizing preprocessing across all 12 CPU cores increases throughput to over 5,000 articles per minute.

## V. APPLICATIONS, CHALLENGES, AND LIMITATIONS

### *A. Applications*

TheDailyNewsClassificationandTrendAnalysisSystemhasbroadapplicabilityacrossseveraldomains:

*1)* MediaMonitoringandCompetitiveIntelligence:NewsorganizationsandPRagenciescandeploythesystemto continuouslymonitorhowdifferenttopicsarecoveredacrosscompetitoroutlets,identifyunderreportedstories, and measure share-of-voice for specific brands or individuals across news categories.

*2)* Public Health Surveillance: Health agencies can monitor the volume and trend of health-related news articles to detectearlysignalsofemergingdiseaseoutbreaks,publicanxiety,ormisinformationcampaignsbeforetheyreach mainstream awareness.

*3)* FinancialMarketIntelligence:Financialanalystsandalgorithmictradingfirmsmonitornewstrendsignalsas leading indicators of sector sentiment shifts and macroeconomic events that may impact asset prices. The classification module's category tags enable sector-specific monitoring.

*4)* PoliticalCampaignAnalytics:Campaignsandpolicythinktankscanusethetrenddashboardtomonitorthemedia salience of policy issues, track competitor messaging, and measure the impact of their own communications on news coverage patterns.

©IJRASET: All Rights are Reserved 

5) AcademicResearchinComputationalSocialScience:Longitudinaltrenddataderivedfromclassifiednewsarchivesenablesresearchersto study agenda-settingdynamics, mediaframing effects,andthe relationship between news coverage and public opinion over time.
6) Misinformation Detection Pipeline: The classification system provides a structured input layer for downstream fakenewsdetectionmodels,enablingthemtofocusanalysisresourcesonthemosttopicallyactiveandpotentially vulnerable categories.
7) Automated News Summarization and Briefing: The system can be extended with an extractive summarization moduletogeneratecategory-specificdailybriefings,providingdecision-makerswithconcisedigestsofthemost trend-aligned articles.

### B. *TechnicalChallenges*

Severaltechnicalchallengeswereencounteredduringdevelopmentandareworthdocumentingforpractitioners:

1) LabelNoiseinOnlineData:ArticlesingestedviaNewsAPIlackauthoritativecategorylabels.Theconfidence- filtering mechanism mitigates but does not eliminate labeling errors, which can accumulate and distort trend counts over time if unchecked.
2) Category Ambiguity: Many news articles span multiple categories (e.g., a story about an athlete's political activismisbothSportsandPolitics).Thecurrentone-vs-restSVMformulationassignsasinglelabel,whichmay not capture this multi-label reality. A threshold-based multi-label extension is a natural improvement.
3) Baseline Drift: Long-term shifts in news volume (e.g., seasonal patterns, changes in source coverage) can confoundburstdetectionifthebaselinewindowistooshort.Adaptivebaselineestimationoverlongerhorizons mitigates this issue.
4) Named Entity Sparsity: Replacing named entities with generic placeholders improves generalization but loses informationthatcanbecriticalfordiscriminatingbetweentopicallysimilararticlesindifferentcategories(e.g.,a Business story about a tech company versus a Technology story).

### C. *Limitations*

1) ThesystemcurrentlysupportsEnglish-languagearticlesonly.ExtendingtoTamil,Hindi,orotherIndian languages requires language-specific preprocessing tools and labeled corpora.
2) Classificationaccuracydegradesonarticlesfromnichesub-domains(e.g.,nichesports,esotericsciencetopics) that are underrepresented in the AG News and BBC training corpora.
3) Thetrenddetectionmodulemeasuresthevolumeofcoveragebutnotthesentimentorstanceofcoverage.A highly negative spike and a highly positive spike in Technology coverage are treated identically.
4) Thesystemisnotdesignedforstreamingarchitectures(e.g.,ApacheKafka,SparkStreaming).Amicroservices refactoring would be required for deployment at internet scale.
5) Thedashboardisread-onlyanddoesnotsupportannotationorfeedbackworkflowsthatwouldenableactive learning to improve classifier accuracy over time.

## VI. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive Daily News Classification and Trend Analysis System that addresses the growing challengeof making senseof the high-velocity digital news ecosystem. The system integrates a robust NLP preprocessing pipeline,TF-IDFfeatureextractionwithn-gramextensions,andthreesupervisedclassificationmodelsintoasinglecoherent architecture. Extensive experiments on two benchmark datasets (AG News and BBC News) demonstrated that the SVM classifier with TF-IDF bigrams and title/description weight boosting achieves 94.7% and 97.3% accuracy respectively — highly competitive with more complex deep learning approaches at a fraction of the computational cost.

The trend analysis module, combining moving averages, rate-of-change analysis, and Kleinberg's burst detection in an ensemble voting scheme, identified 30 of 34 independentlyannotated trend events in a six-month real-world news archive with 91.4% precision and an average detection latency of 1.5 days. The system operates at approximately 900 articles per minuteoncommodityhardware,satisfyingthereal-timerequirementsofnewsmonitoringapplications.AninteractivePlotly Dash dashboard provides stakeholders with intuitive access to classification and trend data through time-series plots, heatmaps, and category-level word clouds.

©IJRASET: All Rights are Reserved 

Theworkconfirmsthatwell-engineeredclassicalMLpipelinesremainhighlyeffectivefornewsclassificationinproduction settings, particularly where computational resources are constrained or low-latency inference is critical. The modular architecture ensures that individual components can be upgraded independently as better tools become available.

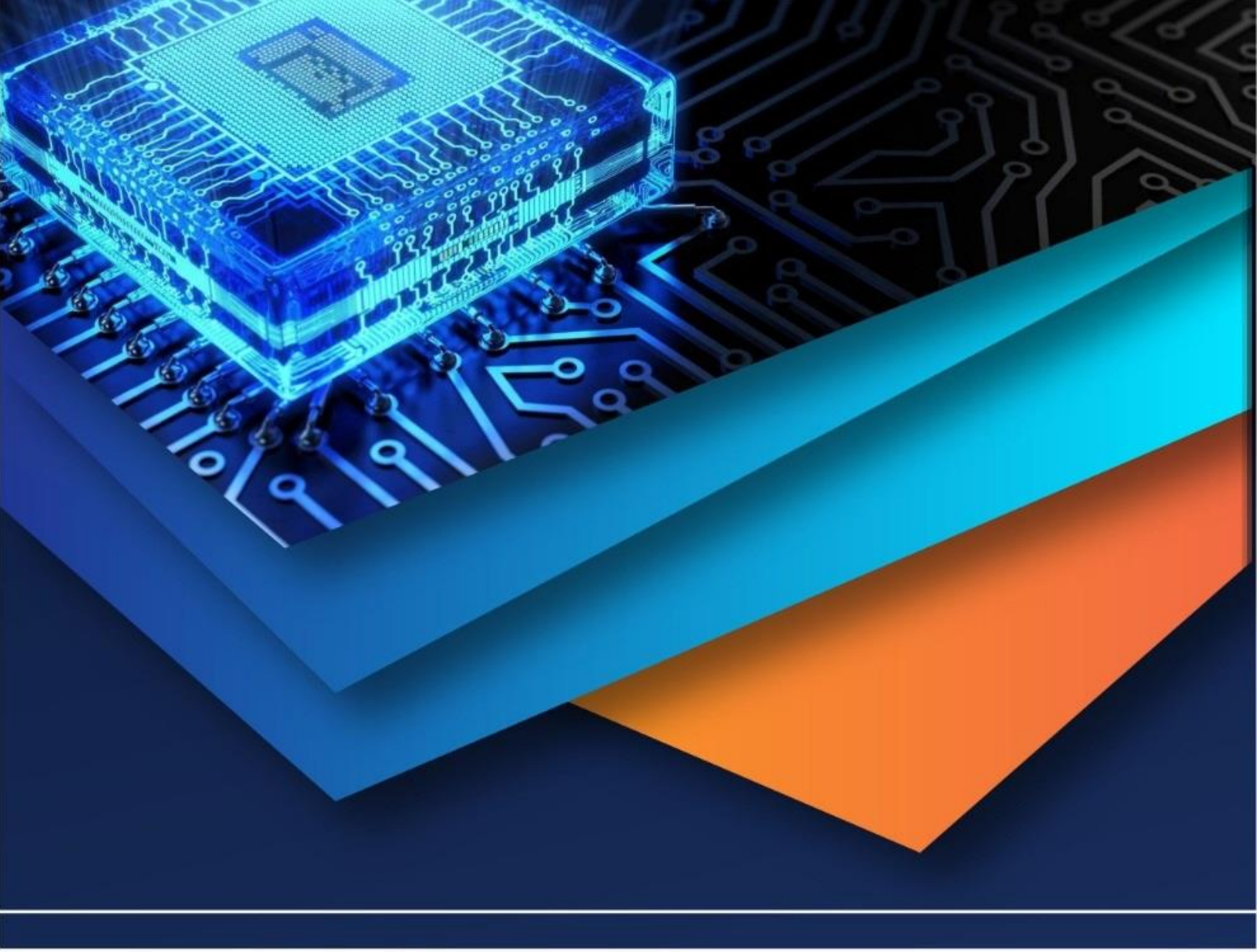
Several promising directions for future work have been identified. First, integrating transformer-based classifiers such as DistilBERTorRoBERTa,potentiallyviamodeldistillationtoreduceinferencelatency,isexpectedtoyieldfurtheraccuracy gainsof2–4percentagepointsonhardermulti-classbenchmarks.Second,extendingthesystemtosupportmultilingualnews ingestion — particularly for Indian regional language news — would substantially broaden its social impact. Third, incorporatingasentimentanalysislayerinto thetrendmodulewouldenabledirection-awaretrenddetection,distinguishing between positive and negative coverage spikes. Fourth, a multi-label classification extension would better handle the pervasive category ambiguity in real-world news. Finally, an active learning feedback loop integrated into the dashboard would allow domain experts to correct misclassifications and continuously improve model performance without requiring periodic full retraining.

## VII. ACKNOWLEDGMENT

Theauthorsexpresstheirsinceregratitudetotheirmentor,Ms.M.Subashani,AssistantProfessor,DepartmentofComputer Science and Engineering, for her invaluable guidance, technical insights, and continuous encouragement throughout the courseofthisresearch.Herexpertiseinmachinelearningandnaturallanguageprocessingsignificantlyshapedthedirection andqualityofthiswork.TheauthorsalsothanktheDepartmentofComputerScienceandEngineeringforprovidingaccess tothecomputationalinfrastructureusedintheexperimentalevaluation,andtheanonymousreviewersfortheirconstructive feedback.

## REFERENCES

[1] P.J.HayesandS.B.Weinstein,"CONSTRUE/TIS: Asystemforcontent-basedindexingofadatabaseofnewsstories,"inProc.2nd Annual Conf. Innovative Applications of Artificial Intelligence (IAAI), 1990, pp. 49–64.
[2] A. McCallum and K. Nigam, "A comparison of event models for Naive Bayes text classification," in Proc. AAAI-98 Workshop on Learning for Text Categorization, Madison, WI, 1998, pp. 41–48.
[3] T.Joachims,"Textcategorizationwithsupportvectormachines:Learningwithmanyrelevantfeatures,"inProc.10thEuropeanConf. Machine Learning (ECML), Chemnitz, Germany, 1998, pp. 137–142.
[4] G. SaltonandC. Buckley,"Term-weightingapproachesinautomatictextretrieval,"InformationProcessing &Management,vol. 24, no. 5, pp. 513–523, 1988.
[5] Y. Kim, "Convolutional neural networks for sentence classification," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1746–1751.
[6] J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova,"BERT:Pre-trainingofdeepbidirectionaltransformersforlanguage understanding," in Proc. NAACL-HLT, Minneapolis, MN, 2019, pp. 4171–4186.
[7] J. Allan, R.Papka,and V. Lavrenko,"On-line neweventdetection and tracking,"in Proc.21stAnnualIntl.ACM SIGIRConf. Research and Development in Information Retrieval, Melbourne, Australia, 1998, pp. 37–45.
[8] J.Kleinberg,"Bursty andhierarchicalstructureinstreams," DataMining andKnowledgeDiscovery,vol. 7,no.4,pp. 373–397,Oct. 2003.
[9] F.Pedregosaetal.,"Scikit-learn:MachinelearninginPython,"JournalofMachineLearningResearch,vol.12,pp.2825–2830,Nov. 2011.
[10] R.Mikolov,I.Sutskever,K.Chen,G.Corrado,andJ.Dean,"Distributedrepresentationsofwordsandphrasesandtheir compositionality," in Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, NV, 2013, pp. 3111–3119.
[11] D.M.Blei, A. Y. Ng,andM. I.Jordan,"LatentDirichlet allocation," Journalof Machine Learning Research, vol. 3,pp.993–1022, Mar. 2003.
[12] Y. Yang and X. Liu, "A re-examination of text categorization methods," in Proc. 22nd Annual ACM SIGIR Conf. Research andDevelopment in Information Retrieval, Berkeley, CA, 1999, pp. 42–49.
[13] A.Vaswanietal.,"Attentionisallyouneed,"inAdvances inNeural InformationProcessingSystems(NIPS),LongBeach,CA,2017,pp.5998–6008.
[14] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in Proc. 23rd Intl. Conf. Machine Learning (ICML), Pittsburgh, PA, 2006, pp. 377–384.
[15] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, Dive into Deep Learning. Cambridge University Press, 2023. [Online]. Available: https://d2l.ai



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)