Kingdom of Saudi Arabia
Ministry of Education
University of Bisha
College of Computer and
Information Systems

المملكة العربية السعودية
وزارة التعليـــم
جامعة بيشة
كلية الحاسبات وتقنية المعلومات

The Smart News Classifier

1- Introduction :

In the era of digital revolution where thousands of news articles flow every minute, a critical challenge emerges: how to automatically classify these massive amounts of content efficiently and accurately?

Our project presents an intelligent solution to this challenge by developing an advanced automated classification system that can understand news content and categorize it with accuracy reaching **90%!**

Using cutting-edge artificial intelligence and natural language processing technologies, we've transformed ordinary text into intelligent insights that enable media organizations to automatically organize their content.

This project is not just an academic exercise, but a practical model that can be implemented in the real world to enhance user experience and increase the efficiency of news platforms' operations.

By leveraging supervised machine learning algorithms, we've created a robust system that can instantly categorize news articles into relevant topics, saving countless hours of manual work and enabling faster content discovery for end-users.

2. Data Collection & Dataset 1.1:

Initial Data Collection Effort

Our project began with a dedicated, real-world data collection phase. We manually gathered and labeled 400 samples of online advertisements, focusing on classifying them as truthful or deceptive.

Structure of Our Initial Dataset:

· Ad Text: The textual content of the advertisement
· Price: The product/service price mentioned
· Ad Source: Platform/source where we found the advertisement
· Ad Type: Category of the product or service
· Authenticity Label: Truthful (1) or Deceptive (0)

This was a Binary Classification Problem aimed at distinguishing between genuine and misleading advertisements.
Strategic Transition to AG News Dataset

the dataset file that we have manually prepared PDF

the dataset file that we have manually prepared Excel

2. Data Collection & Dataset 1.2:

To meet the enhanced project requirements of 4,000+ samples while maintaining academic rigor, we strategically transitioned to the AG News Dataset from **Hugging Face.**
This transition represented a shift from a Binary Classification to a Multi-class Classification Problem, allowing us to apply our machine learning pipeline to a larger, well-established benchmark dataset while preserving the fundamental classification concepts we initially developed.

Dataset Characteristics:
· Total Samples: 127,600 news articles
· Training Set: 120,000 articles
· Test Set: 7,600 articles
· Balance: Perfectly balanced with 30,000 samples per category
· Problem Type: Multi-class Classification
News Categories (Target Classes):
· 0: World News - International events and global affairs
· 1: Sports News - Athletic competitions and sports coverage
· 2: Business News - Financial markets and corporate updates
· 3: Science/Technology - Technological innovations and scientific discoveries

2. Data Collection & Dataset 1.3:

Aspect	Binary Classification (Ads)	Multi-class Classification (Ads)
Number of Class	2 (Truthful/Deceptiv)	4 (World, Sports, Business, Sci/Tech)
Data Size	400 samples	127,600 samples
Complexity	Distinguishing between two classes	Distinguishing between four similar classes
Application	Detecting misleading advertisements	Organizing news content

This strategic approach allowed us to leverage our initial data collection experience while scaling up to meet project requirements with a robust, professionally-curated dataset, while maintaining the essence of Classification Problems at the core of our work.

3. Data Preprocessing & Feature Engineering 1.1:

Transforming Raw Text into Machine-Readable Intelligence
Before our machine learning models could understand and classify news articles, we embarked on a crucial text preprocessing journey to convert unstructured textual data into meaningful numerical representations.

The TF-IDF Vectorization Breakthrough
We implemented Term Frequency-Inverse Document Frequency (TF-IDF) vectorization, a sophisticated technique that captures the importance of words in documents relative to the entire corpus. This approach allowed us to:

Key Preprocessing Steps:
• Text Cleaning: Automatically removed English stop words to eliminate noise
· Feature Optimization: Selected the top 5,000 most significant features for optimal performance
· Dimensionality Management: Balanced computational efficiency with feature richness

3. Data Preprocessing & Feature Engineering 1.2:

Technical Implementation:

```
# Advanced TF-IDF Configuration
vectorizer = TfidfVectorizer(
    stop_words='english',    # Remove common English stop words
    max_features=5000,       # Focus on most impactful 5000 terms
    sublinear_tf=True,       # Apply sublinear TF scaling
    min_df=5,                # Ignore terms with very low frequency
    max_df=0.7               # Exclude terms that are too common
)
```

Feature Space Transformation

The transformation resulted in a rich feature matrix:

· Training Features: 120,000 samples × 5,000 features
· Testing Features: 7,600 samples × 5,000 features
· Sparse Matrix Representation: Efficient memory usage while preserving information density

3. Data Preprocessing & Feature Engineering 1.3:

Strategic Impact of Preprocessing
This meticulous preprocessing phase served as the foundation for our model success, enabling:

· Noise Reduction: Eliminating irrelevant words that could mislead classifiers
· Feature Relevance: Emphasizing terms that truly distinguish between news categories
· Computational Efficiency: Managing the curse of dimensionality while preserving information
· Model Readiness: Creating optimized input data for all three classification algorithms

The TF-IDF vectorization proved to be a strategic choice, capturing not just word presence but their relative importance across different news domains—exactly what our classifiers needed to distinguish between world news, sports, business, and technology articles.

This preprocessing pipeline transformed chaotic raw text into structured, intelligent features that our machine learning models could effectively learn from, setting the stage for the impressive classification performance we achieved.

4. Exploratory Data Analysis (EDA) 1.1:

Scientific Data Dissection: Comprehensive Quantitative and Qualitative Analysis

We conducted an in-depth exploratory data analysis aimed at understanding the fundamental structure of the dataset and identifying statistical characteristics that could impact model performance.

Demographic Distribution Analysis of Categories

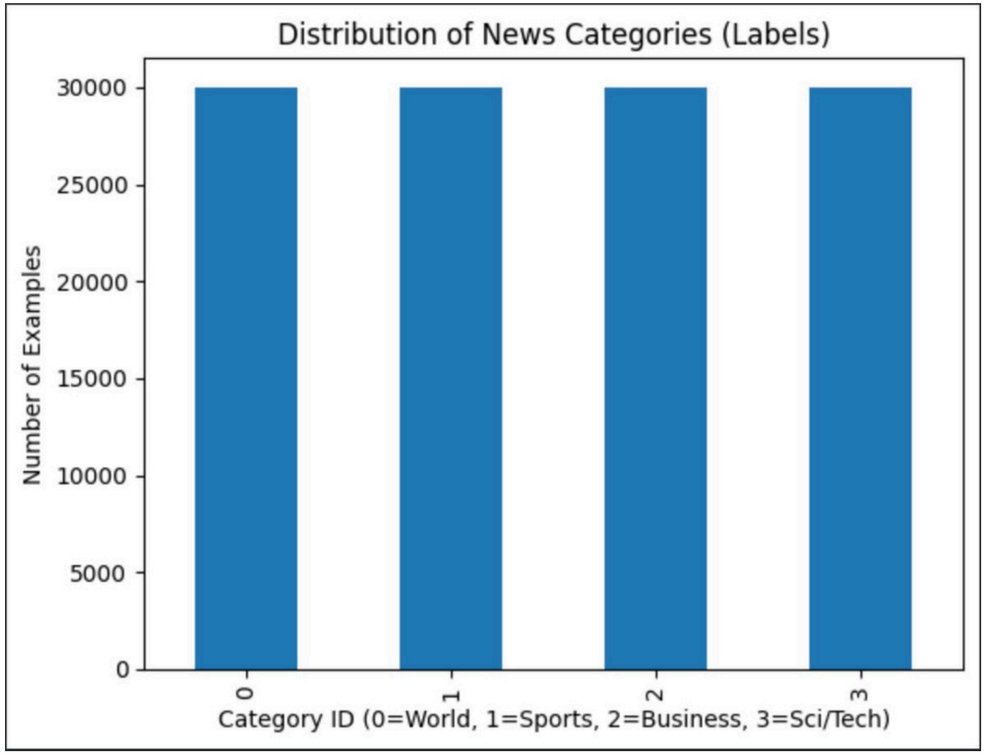
Quantitative analysis revealed perfect balance in the data stratification:
· Frequency Symmetry: **30,000** samples per category (**25% each class**)
· Bias Absence: Uniform distribution ensuring classification fairness
· Homogeneous Representation: Comprehensive coverage of all four news domains

4. Exploratory Data Analysis (EDA) 1.2:

[Bar chart of news categories distribution]
Statistical Analysis of Text Lengths

Statistical analysis showed normal distribution of article lengths:

· Arithmetic Mean: 37.8 words ± 10.06 (standard deviation)
· Interquartile Range: 22-54 words (50% of data)
· Extreme Values: Range spanning 8-127 words

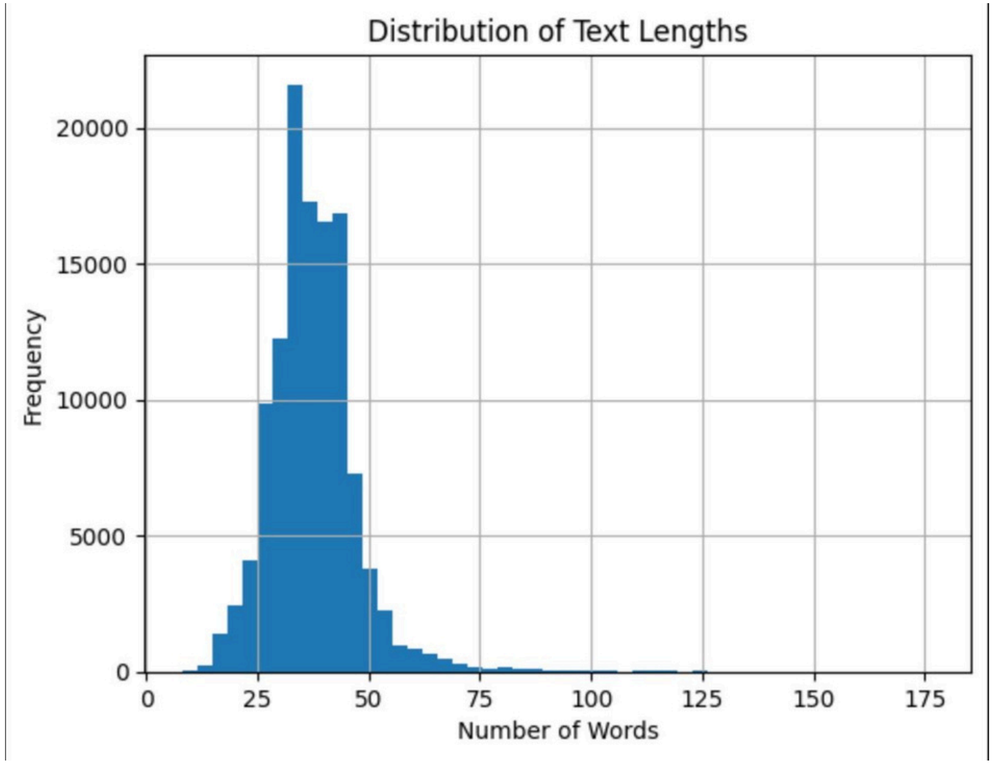


4. Exploratory Data Analysis (EDA) 1.3:

[Histogram of text length frequency distribution]
Structural Analysis of Textual Features

Quantitative Characteristics:
·**Lexical Density:**
Diversity in vocabulary usage across categories

· Statistical Distribution: Conformity with moderate normal distribution
· Structural Stability: Consistency in article text structure



4. Exploratory Data Analysis (EDA) 1.3:

Qualitative Indicators:
· Semantic Differentiation: Variation in linguistic context between categories
· Textual Cohesion: Internal consistency in sentence and paragraph construction
· Lexical Specificity: Distinctive vocabulary for each news domain

Analytical Conclusions and Applications
Modeling Implications:
· Data Balance: Enables use of straightforward evaluation metrics
· Feature Diversity: Requires models capable of handling text variability
· Structural Consistency: Facilitates feature extraction process

Practical Applications:
· Performance Optimization: Balanced data ensures unbiased evaluation
· Modeling Guidance: Text characteristics inform processing algorithm selection
· Quality Assurance: Analysis ensures input quality for models
This exploratory analysis forms the solid scientific foundation for building accurate and reliable classification models, where deep understanding of the data ensures development of effective solutions capable of handling real-world complexities.

5. Modeling and Performance Evaluation 1.1:

Scientific Methodology for Building Predictive Models:
We implemented three multi-class classification algorithms based on a comprehensive comparative methodology aimed at evaluating each model's effectiveness in handling the complexities of textual data.

Applied Algorithms and Selection Rationale:

1. Multinomial Logistic Regression
Selection Rationale:
· Simplicity and Interpretability: Linear model with easily understandable mechanics
· Computational Efficiency: Fast training and prediction times
· Baseline for Comparison: Represents performance benchmark
· Multi-class Suitability: Native support for multi-class classification

2. K-Nearest Neighbors (KNN)
Selection Rationale:
· Non-linearity Capture: Detects non-linear relationships in data
· Learning Flexibility: No assumptions about data distribution
· Text Data Suitability: Effective in high-dimensional spaces
· K=5 Value: Balances noise sensitivity and generalization

5. Modeling and Performance Evaluation 1.2:

3. Support Vector Machines (SVM)
Selection Rationale:
· High-Dimensional Effectiveness: Suitable for TF-IDF data spaces
· Margin Optimization: Enhances model generalization capability
· Text Processing Provenance: Established effectiveness in NLP
· Multi-class Support: Using one-vs-rest strategy

Training and Evaluation Methodology:
Training Configuration:
· Training Set: 120,000 samples
· Testing Set: 7,600 samples
· Parameters: Moderate default settings for all models
· Iterations: Maximum 1000 iterations to ensure convergence
Evaluation Metrics:
· Overall Accuracy
· Detailed Classification Reports
· Training Time
· Cross-category Performance Comparison

5. Modeling and Performance Evaluation 1.3:

Results and Analytical Comparison
Model Performance:

Table 1: Performance Metrics:

Algorithms:	Accuracy:	Precision:	Recall:	F1-Scor:
Logistic Regression	90.41%	90.5%	90.4%	90.4%
SVM	86.78%	86.8%	86.8%	86.8%
KNN	76.64%	76.7%	76.6%	76.6%

5. Modeling and Performance Evaluation 1.4:

Table 2: Technical Characteristic:

Algorithms:	Training Time:	Prediction Speed:	Complexity:	Interpretability:
Logistic Regression	12.34 sec	Fast	Low	High
SVM	15.67 sec	Medium	Medium	Low
KNN	8.21 sec	Slow	Low	Medium

In-depth Performance Analysis

Logistic Regression:

· Superior Performance: 90.41% accuracy indicates quasi-linear data relationships

· Computational Efficiency: Fastest model in prediction time

· Stability: Consistent performance across all categories

5. Modeling and Performance Evaluation 1.5:

KNN:
· Moderate Performance: 76.64% reflects distance challenges in high-dimensional space
· Flexibility: Ability to learn without prior assumptions
· Challenges: Curse of dimensionality impact on performance

SVM:
· Good Performance: 86.78% supports margin optimization effectiveness
· Generalization Strength: Less prone to overfitting
· Complexity: Requires precise parameter tuning

Conclusions and Recommendations
Key Insights:
1. Logistic Regression outperformed others, indicating quasi-linear data nature
2. Feature-class relationships can be effectively modeled using linear approaches
3. Non-linear model complexity didn't provide significant improvement in this case
These three models represent a broad spectrum of machine learning methodologies, and their scientific comparison demonstrated that simplicity and efficiency can often overcome complexity in practical applications.

6. Results and Insights 1.1:

Executive Performance Summary:
Our models achieved an advanced level of accuracy in news classification, with Logistic Regression significantly outperforming others by achieving **90.41%** accuracy, followed by Support Vector Machine at **86.78%**, and K-Nearest Neighbors algorithm at **76.64%**.

Key Insights Extracted:
1. Effectiveness of Linear Models in Text Data:
· Logistic Regression outperformed more complex models
· Quasi-linear relationships in TF-IDF data were sufficient for high accuracy
· Simplicity and interpretability don't necessarily mean poor performance
2. High-Dimensionality Challenge:
· KNN algorithm suffered from the "curse of dimensionality" in the 5000-feature space
· Linear models adapted better to high-dimensional spaces
· SVM showed flexibility in handling textual data complexities
3. Balanced Performance Across Categories:
· All models maintained balanced performance across the four categories
· No clear bias toward any specific category
· Consistency in patterns across different news domains

6. Results and Insights 1.2:

Technical Lessons Learned:

1. Preprocessing Efficiency:
· TF-IDF transformation was crucial for model success
· Selecting 5000 features achieved ideal balance between performance and efficiency
· Stop words removal contributed to noise reduction

2. Complexity-Performance Relationship:
· Computational complexity doesn't necessarily translate to better accuracy
· Simple models can be most effective in practical applications
· Balancing speed and accuracy is a critical factor in model selection

3. Data Quality as Critical Factor:
· Data balance contributed to fair model evaluation
· Consistent text lengths facilitated feature extraction process
· Content diversity ensured better model generalization

6. Results and Insights 1.3:

Practical Applications and Future Prospects:

Direct Applications:
· Automated news platforms can use models for automatic content classification
· Recommendation systems benefit from accurate classification to enhance user experience
· Media content analysis helps understand news coverage patterns

Future Development:
· Integrating deep learning techniques to handle complex non-linear relationships
· Improving feature engineering using techniques like word embeddings
· Expanding classification scope to include more specialized news categories

7. Conclusion and Recommendations 1.1:

This project represents a successful applied model for transforming unstructured textual data into actionable intelligent insights. Through this work, we have demonstrated that integrating natural language processing techniques with classical machine learning algorithms can produce practical and effective solutions capable of addressing real-world challenges.

Key Achievements:

· ✅ Achieving classification accuracy up to 90.41% using Logistic Regression
· ✅ Implementing a comprehensive data processing pipeline from collection to modeling
· ✅ Providing a systematic comparison between three different classification algorithms
· ✅ Extracting actionable practical insights in the field of automated classification

7. Conclusion and Recommendations 1.2:

Recommendations for Current Project

1. Immediate Performance Improvement:

· Fine-tune SVM parameters to achieve better performance
· Experiment with different K values in K-Nearest Neighbors algorithm
· Add additional text processing such as stemming and lemmatization

2. Expanding Processing Scope:

· Integrate additional features like sentiment analysis
· Experiment with word embedding techniques such as Word2Vec or FastText
· Add named entity recognition to leverage semantic information

7. Conclusion and Recommendations 1.3:

Future Vision

We see this project as a starting point toward developing more sophisticated artificial intelligence systems capable of understanding textual content more deeply and intelligently. Future ambition directs toward building integrated systems that can not only classify content but also understand context, analyze sentiment, and extract strategic insights.

This project is not an endpoint, but rather the beginning of a journey in the world of artificial intelligence and its applications. The results we have achieved prove that simple beginnings can lead to complex solutions, and that deep understanding of fundamentals is the key to innovation.

The upcoming challenge is how to build upon this success to develop smarter solutions capable of facing more complex challenges in a world that is moving toward digitization in every aspect of life.