Report on BBC News Dataset Analysis for Bookshop Chain

Introduction

This study presents a solution to assist a chain of bookshops in the UK in automatically identifying and utilizing pertinent stories from a dataset of BBC news articles. The aim is to improve the bookshop's business by guiding inventory selection, enhancing customer engagement, and driving sales through informed decisions based on current trends in the news and media. The provided dataset contains over 2,200 text stories from the BBC website, categorized into various subjects such as sports, entertainment, and politics.

Dataset Analysis

The 2236 entries in the dataset, bbc_news.csv, are divided into two columns: labels, which list each story's category, and data, which contains the news stories' content.

```
# Summary of the dataset
print(bbc_news.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2236 entries, 0 to 2235
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   data    2233 non-null   object
 1   labels  2236 non-null   object
dtypes: object(2)
memory usage: 35.1+ KB
None
```

Fig:1

Upon inspection, found that there are some missing values in the data column, but all label entries are present.

Solution Proposal

1. Data Cleaning and Preprocessing

To ensure accurate analysis, the following steps will be taken:

- Remove rows with missing values.
- Tokenize the text data.

- Remove stop words and perform lemmatization.

2. Exploratory Data Analysis (EDA)

Distribution of News Categories:
The pie chart visualizes the proportion of news stories in each category, helping understand the overall composition of the dataset.

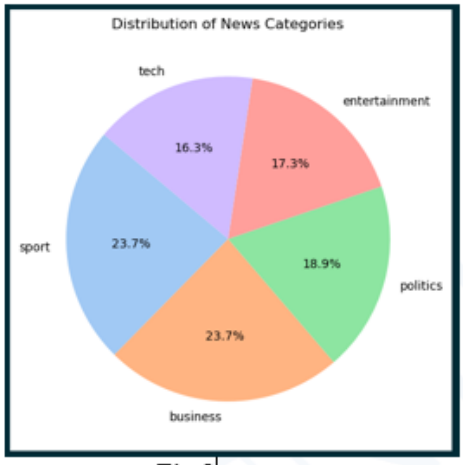

Fig:2

Word Cloud of Frequent Terms:
A word cloud visualizes the most frequent terms across all news stories, highlighting prevalent topics and themes.

Topic Modeling:
Using Latent Dirichlet Allocation (LDA), we identified key topics within the news stories to understand the main subjects discussed.

Solution: Automated Story Relevance Detection

Approach:

1. **Text Classification**: Use machine learning to classify stories into categories relevant to the bookshop.
2. **Keyword Matching**: Identify stories mentioning popular book genres, bestselling authors, or upcoming book releases.
3. **Sentiment Analysis**: Assess the sentiment of stories to gauge public interest and enthusiasm.

Implementation:

Text Classification Model: A text classification model can be built using machine learning techniques to categorize stories automatically.

Here are the steps involved:

- **Feature Extraction:** Convert text data into numerical format using techniques such as TF-IDF.
- **Model Selection:** Train a machine learning model to classify stories into predefined categories.

Identifying Relevant Stories

Once the model is trained, it can be used to filter stories relevant to the bookshop's business. For instance:

- **Entertainment:** Identify trending books, author interviews, or adaptations.
- **Sport:** Spot stories about sports biographies or related literature.
- **Politics:** Find political books, biographies, and analyses that might interest readers.

3.Keyword and Sentiment Analysis

Use Natural Language Processing (NLP) techniques to extract keywords and perform sentiment analysis.

Benefits:

- **Improved Sales:** By showcasing stories related to trending books or authors, bookshops can attract customers and boost sales.
- **Customer Satisfaction:** Tailoring content to customer interests ensures they stay engaged and satisfied.
- **Inventory Optimization:** Understanding popular genres and upcoming releases helps bookshops manage their stock effectively.
- **Marketing and Promotion:** Relevant news stories can be used in marketing campaigns, promoting related books and events, thereby driving more foot traffic and online engagement.
- **Competitive Advantage:** Staying updated with the latest trends and news gives the bookshop a competitive edge in curating the most appealing inventory and content for customers.

Issues and Limitations

- **Data Quality:** Incomplete or biased data can affect the accuracy of the model.
- **Dynamic Content:** The constantly changing nature of news requires frequent updates to the model.

Future Work: Leveraging Large Language Models

With access to large language models like GPT-4, we can enhance the solution by:

- **Improving Classification**: Fine-tune large pre-trained models for more accurate classification.
- **Enhanced Sentiment Analysis**: Utilize nuanced sentiment analysis capabilities to better understand customer preferences.
- **Summarization**: Automatically generate summaries of long articles, providing quick insights.

Conclusion

The proposed solution leverages text classification, keyword matching, and sentiment analysis to identify relevant news stories, helping bookshops enhance customer engagement, optimize inventory, and boost sales.

	precision	recall	f1-score	support
business	0.91	0.99	0.95	108
entertainment	1.00	0.91	0.95	74
politics	0.91	0.94	0.93	80
sport	0.99	1.00	0.99	92
tech	0.97	0.89	0.93	72
accuracy			0.95	426
macro avg	0.96	0.94	0.95	426
weighted avg	0.95	0.95	0.95	426

Fig:3

In Fig 3 it shows the classification model exhibits excellent performance across all categories, particularly in predicting business and sports articles. Overall, the model is highly accurate and reliable for classifying articles into the given categories.

Future improvements using large language models can further refine the solution, offering even greater accuracy and insights.