BÁO CÁO FAKE NEW DETECTIONS

I. EDA:

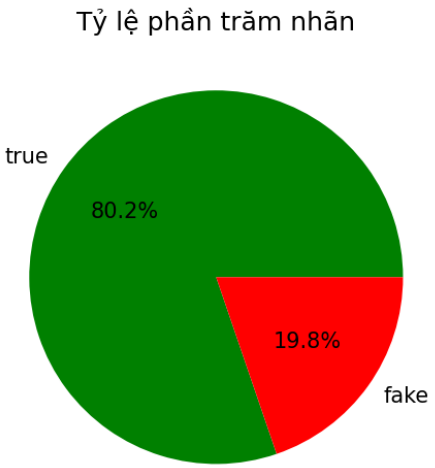
1. Thông tin cơ bản
    - Số hàng: 8378
    - Số cột: 4
    - Tên các cột: [‘title’, ‘text’, ‘year-month’, ‘labels’]
    - Không có giá trị NULL
    - 1 vài dữ liệu mẫu:

	title	...	labels
0	Disapproval rating for Japan PM Abe exceeds su...	...	true
1	Virginia officials postpone lottery drawing to...	...	true
2	Trump administration issues new rules on U.S. ...	...	true
3	Trump administration taps coal consultant for ...	...	true
4	Clashes in Rome as police evict refugee squatt...	...	true

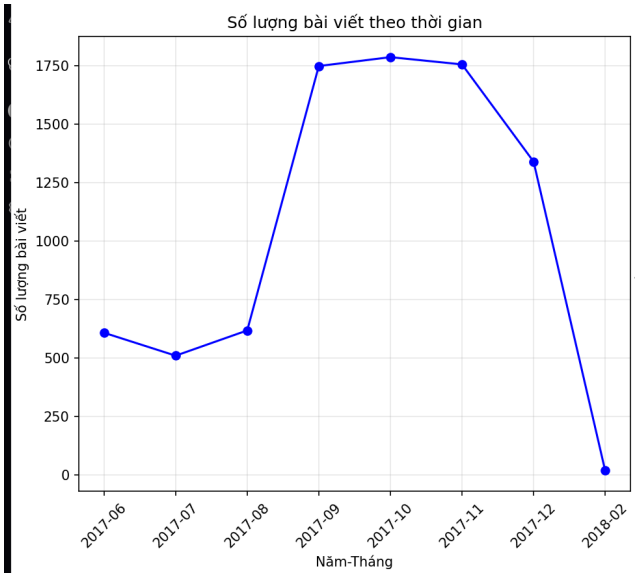
=> Dữ liệu sạch, không cần xử lý missing values

2. Phân tích labels:
    - Label ‘TRUE’: 6723
    - label ‘FAKE’: 1655

=> Dữ liệu nhãn bị lệch



3. Phân tích theo thời gian:



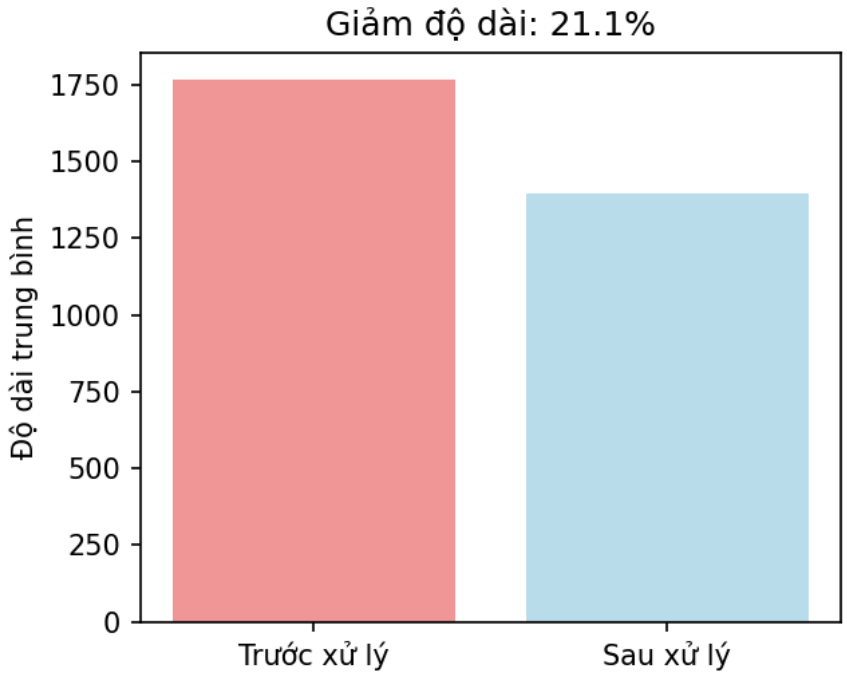
=> Số lượng news tăng mạnh từ tháng 9 đến tháng 11 năm 2017 và dường như chỉ còn một vài mẩu news vào tháng 2 năm 2018. Giai đoạn tăng mạnh đi kèm với sự tăng mạnh của tin giả nên cần xây dựng mô hình phân biệt.

4. Phân tích theo độ dài:
    - Độ dài trung bình text của label ‘TRUE’: 1633 ký tự
    - Độ dài trung bình text của label ‘FAKE’: 1949 ký tự

=> Trung bình tin fake dài hơn do thường cố gắng kể lể dài dòng, thêm chi tiết không cần thiết làm nội dung trông đáng tin hơn. Chênh lệch không thực sự lớn nên ta có thể kết hợp với các đặc trưng ngôn ngữ khác để đánh giá.

II. Data preprocessing

1. Làm sạch dữ liệu
    - Xử lý cột *title và text* và *combined:* chuyển toàn bộ về chữ thường, loại bỏ các ký tự đặc biệt, khoảng trắng thừa đồng thời ghép 2 cột *title và text* làm 1.
    - Loại bỏ stopwords và stemming cơ bản.
    - Top 5 từ trước remove stopwords: ['said', 'trump', 'u.', 'would', 's.']
    - Top 5 từ sau remove stopwords: ['trump', 'state', 'president', 'govern', 'year']



=> Dữ liệu văn bản được chuẩn hoá, giữ lại các từ khoá mang tính nội dung và dễ dàng đưa vào bước trích xuất đặc trưng

2. Xử lý nhãn và thời gian
    - Cột year-month được tách thành 2 cột riêng biệt: year và month
    - Chuyển nhãn từ chữ (true,fake) sang dạng nhị phân (1,0)

=> Dữ liệu bao phủ giai đoạn từ 2017 đến 2018, giúp thuận tiện cho việc phân tích theo thời gian. Nhãn chuyển đổi sang dạng số để làm việc cùng các mô hình và các hàm chức năng.

3. Mẫu dữ liệu sau khi được xử lý

   Dữ liệu từ raw.csv được lưu thành preprocessed_data.csv

	combined_text	year	month	label_binary
0	disapproval rating japan abe exceed kyodo poll...	2017	10	1
1	virginia official postpone lottery draw decide...	2017	12	1
2	trump administra issue rule visa waiver washin...	2017	12	1

III. Trained models

1. Mục tiêu
    - So sánh hiệu năng giữa một số mô hình cơ bản để chọn ra baseline tốt nhất cho bài toán *Fake News Detection*. Các tiêu chí so sánh bao gồm: Accuracy,

Precision, Recall, F1-score, thời gian huấn luyện và độ ổn định (standard deviation) qua cross-validation.

2. Dữ liệu & chuẩn bị
   - Dữ liệu sử dụng: preprocessed_data.csv
   - Kích thước dataset: 8378 bản ghi
   - Vectorization: TF-IDF với:
     + max_features = 10000 (giảm chiều dữ liệu)
     + ngram_range = (1,2) (lấy từ đơn hoặc 1 cặp từ)
     + stop_words = 'english' ('the','is','at'...)
     + max_df, min_df = 0.95, 2 (loại bỏ từ xuất hiện quá nhiều và từ quá hiếm)
   - Scikit-learn (sử dụng cross_validate, StratifiedKFold)
   - Chiến lược tìm kiếm: RandomizedSearchCV (30 candidates) và GridSearchCV.
3. Mô hình được so sánh

   Logistic Regression (solver = 'liblinear', max_iter = 1000)
   - Thư viện: sklearn.linear_model.LogisticRegression
   - Kết quả:Kết quả (5-fold CV):
     - Accuracy: **0.9619** (±0.0026)
     - Precision: **0.9584** (±0.0072)
     - Recall: **0.9958** (±0.0038)
     - F1-score: **0.9767** (±0.0015)
     - Thời gian train (tổng CV): **~4.39s**

   Random forest (n_estimators = 100)
   - Thư viện: sklearn.ensemble.RandomForestClassifier
   - Kết quả (5-fold CV):
     - Accuracy: **0.9685** (±0.0033)
     - Precision: **0.9660** (±0.0090)
     - Recall: **0.9958** (±0.0041)
     - F1-score: **0.9807** (±0.0020)
     - Thời gian train: **~6.24s**

   Naive Bayes (MultinominalNB)
   - Thư viện: sklearn.naive_bayes.MultinomialNB
   - Kết quả (5-fold CV):

- Accuracy: **0.9563** (±0.0052)
- Precision: **0.9733** (±0.0102)
- Recall: **0.9723** (±0.0153)
- F1-score: **0.9728** (±0.0033)
- Thời gian train: **~2.49s**

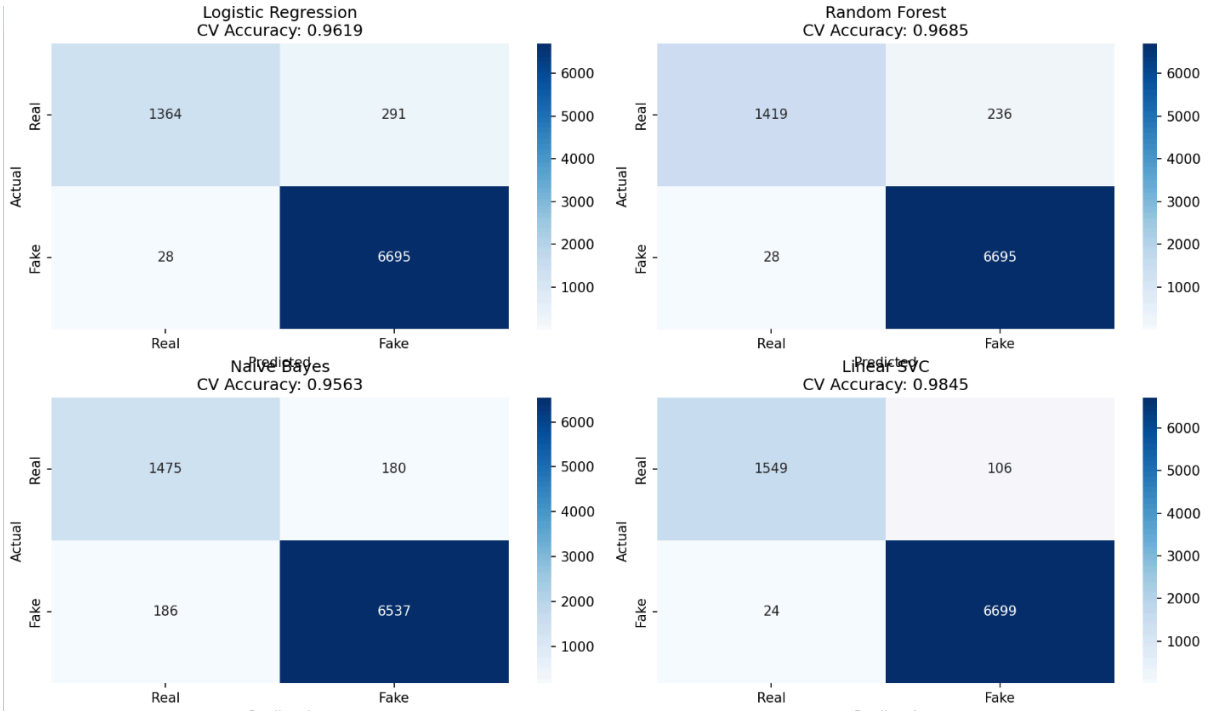
Linear SVC (dual = false, max_iter = 5000)

- Thư viện: sklearn.svm.LinearSVC
- Kết quả (5-fold CV):
  - Accuracy: **0.9845** (±0.0024)
  - Precision: **0.9844** (±0.0067)
  - Recall: **0.9964** (±0.0047)
  - F1-score: **0.9904** (±0.0014)
  - Thời gian train: **~1.26s**

Bảng so sánh chi tiết:

Model	Accuracy_Mean	Accuracy_Std	F1_Mean	F1_Std	Precision_Mean	Recall_Mean	Train_Time
Logistic Regression	0.9619	0.0026	0.9767	0.0015	0.9584	0.9958	4.4535
Random Forest	0.9685	0.0033	0.9807	0.0020	0.9660	0.9958	6.1775
Naive Bayes	0.9563	0.0052	0.9728	0.0033	0.9733	0.9723	2.3661
Linear SVC	0.9845	0.0024	0.9904	0.0014	0.9844	0.9964	1.2861

=> Overall score: 0.9796 cao nhất nên chọn Linear SVC



=> Qua confusion matrix từ cross-validation, có thể thấy Linear SVC cho kết quả vượt trội: số lượng tin giả bị bỏ sót (FN) rất thấp (24), đồng thời giảm thiểu nhầm lẫn tin thật thành giả (FP = 106). Trong khi đó, Logistic Regression và Random Forest vẫn còn khá nhiều FP, còn Naive Bayes thì bỏ sót nhiều fake news. Do đó Linear SVC là lựa chọn tối ưu cho bài toán Fake News Detection

4. Tuning Linear SVC
    - GridSearch (thử 432 tổ hợp tham số)
    - RandomizedSearch (thử random 30 tổ hợp tham số)
    - Cả 2 đều cho ra 1 best parameter:

```
Best parameters:
  classifier__C: 10
  classifier__dual: False
  classifier__max_iter: 1000
  tfidf__max_df: 0.9
  tfidf__max_features: 10000
  tfidf__min_df: 2
  tfidf__ngram_range: (1, 2)
```

- Nhưng do sự khác biệt giữa số lượng tổ hợp tham số mà GridSearch cần 1623.48s trong khi RandomizedSearch chỉ cần 116.27s tuning
- Ta có được best CV F1 score: **0.9913**

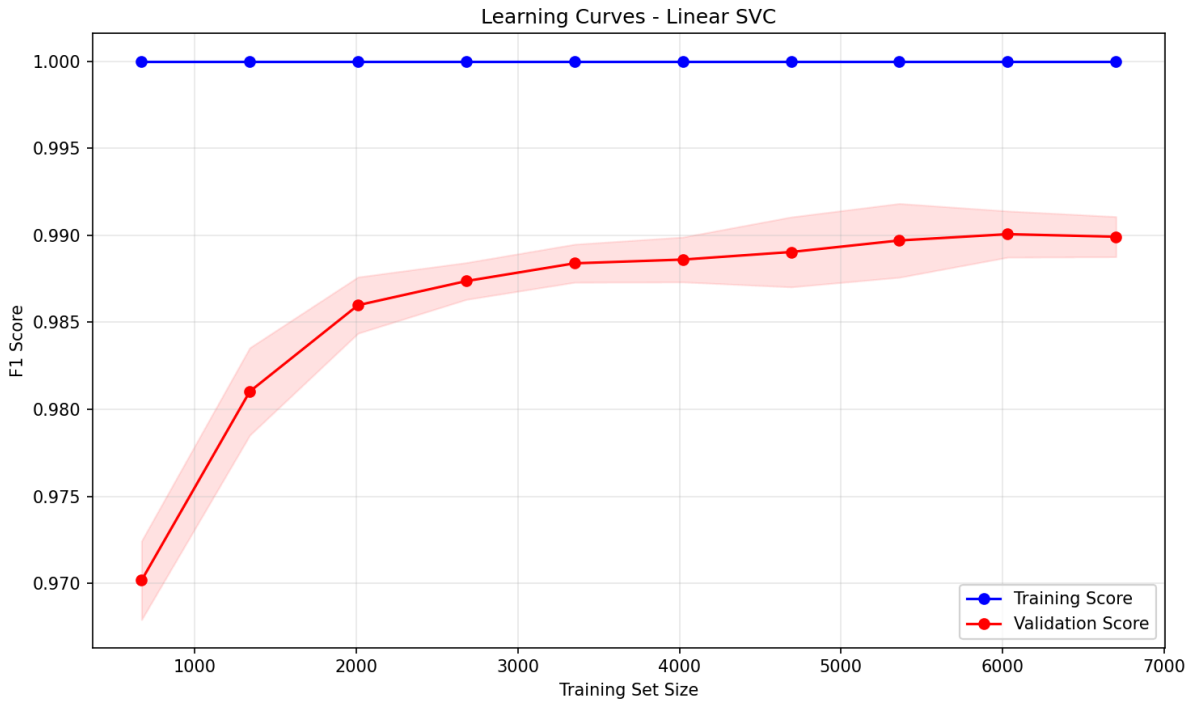
IV. Evaluate Final Model

Linear SVC:

```
Accuracy: 0.9860 (+/- 0.0044)
Precision: 0.9869 (+/- 0.0063)
Recall: 0.9958 (+/- 0.0028)
F1: 0.9913 (+/- 0.0027)
```

.

- Mức độ chuẩn xác cao: tất cả các metrics đều >0.98
- Cân bằng precision và recall: mô hình ổn định
- Sai số rất nhỏ: mô hình ổn định qua các folds của cross-validation
- Kiểm tra overfitting:
  Learning curve (Train vs Val):



- Ta thấy Validation score tăng từ từ và Val score cuối: 0.9899 khá gần Trainning score

=> Model không bị overfitting - Được generalize tốt

V. Conclusion

- Linear SVC tuned tốt, F1 ~ 0.99, không overfitting
- Naive Bayes:

```
Accuracy: 0.9685 (+/- 0.0066)
Precision: 0.9660 (+/- 0.0082)
Recall: 0.9958 (+/- 0.0024)
F1: 0.9807 (+/- 0.0039)
```

- Ta thử tuning với Naive Bayes, một mô hình có điểm số default khá tốt nhưng kết quả vẫn kém so với Linear SVC