СИСТЕМА АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ФІШИНГОВИХ ЕЛЕКТРОННИХ ЛИСТІВ

ПРИХОДЬКО БОГДАГ ІТ

Цілі та завдання проєкту

ЦІЛЬ ПРОЄКТУ

- Створити систему, яка автоматично класифікує електронні листи як фішингові або легітимні, на основі їхнього текстового вмісту.
- Застосувати методи машинного навчання та глибинних нейронних мереж для вирішення задачі.

ЗАВДАННЯ

- Підготувати датасет текстів електронних листів з відповідними мітками (фішинг / легітимні).
- Провести попередню обробку (очищення, нормалізація) текстів.
- Реалізувати три моделі класифікації різної складності:
  - базова модель на основі логістичної регресії;
  - згорткова нейромережа (Text CNN);
  - трансформерна модель (BERT) для глибокого аналізу тексту.
-
- Оцінити якість моделей за метриками accuracy, precision, recall, F1-score.
- Порівняти результати моделей і визначити найбільш ефективну для задачі.

Загальна архітектура системи

1. **Вхід:** текст електронного листа, опціонально — тема листа, посилання, інші метадані.
2. **Попередня обробка:**
   - видалення HTML-тегів і «сміттєвих» символів;
   - нормалізація пробілів, приведення тексту до стандартного вигляду;
   - переклад українського тексту → англійська (у GUI) для моделей CNN/BERT;
   - (для моделей CNN/BERT) розбиття довгих текстів на чанки.
3. Моделі класифікації (логістична регресія, Text CNN, BERT)
4. **Вихід:** клас (0 = легітимне, 1 = фішингове) + ймовірнісна оцінка (confidence score)
5. І**нтерфейс:** GUI-застосунок, дозволяє користувачу вводити текст і отримувати результат класифікації в реальному часі.

Pipeline (етапи обробки)

ЗБІР ДАНИХ

використано датасет з ~19 489 записами (10 590 фішингових, 8 899 легітимних)

ПОПЕРЕДНЯ ОБРОБКА ТЕКСТУ

ПОДІЛ НА НАВЧАЛЬНІ / ВАЛІДАЦІЙНІ / ТЕСТОВІ МНОЖИНИ, ЗБЕРЕЖЕННЯ ПРОПОРЦІЇ КЛАСІВ

ФОРМУВАННЯ ОЗНАК / ТОКЕНІЗАЦІЯ

- для логістичної регресії: TF-IDF (уніграм + біграми)
- для Text CNN: Tokenizer + Embedding + Conv1D + Pooling.
- для BERT: токенізація , розбиття листів на чанки, усереднення ймовірностей по чанках.

НАВЧАННЯ МОДЕЛЕЙ

ОЦІНКА МОДЕЛЕЙ

accuracy, precision, recall, F1-score

Архітектура моделей

ЛОГІСТИЧНА РЕГРЕСІЯ

- Вхід: TF-IDF вектор ознак (уніграм + біграми)
- Алгоритм: логістична регресія з регуляризацією, max_iter налаштовано.
- Перевага: швидка, проста у впровадженні, дає базову точку порівняння.

Архітектура моделей

TEXT CNN

- Вхід: послідовність токенів → Embedding шар → фільтри Conv1D для виявлення локальних шаблонів у тексті.
- Далі: GlobalMaxPooling1D → Dropout → Повнозв’язні шари → Вихідний клас.
- Перевага: краще захоплює контекст фраз у порівнянні з TF-IDF, середня складність

Архітектура моделей

BERT TRANSFORMER

- Використано попередньо натреновану трансформер-модель (наприклад, BERT) → fine-tuning на задачі класифікації.
- Токенізація за WordPiece, обробка контексту, розбиття довгих листів на чанки, результати чанків усереднюються.
- Перевага: глибоке розуміння семантики, найвища продуктивність серед трьох підходів.

Навчання моделей

- Логістична регресія: навчання на TF-IDF ознаках, налаштування параметрів (регуляризація, max_iter)
- Text CNN: підбір Tokenizer, Embedding-розмірності, кількість фільтрів Conv1D, batch size, кількість епох, Dropout для запобігання переобученню.
- BERT: fine-tuning на спеціалізованому датасеті, обробка довгих повідомлень через чанки, усереднення результатів чанків.
- Метрики оцінки на тестовій вибірці: accuracy, precision, recall, F1-score — для всіх трьох моделей.

Результати та порівняння

МОДЕЛЬ	ACCURACY	PRECISION	RECALL	F1-SCORE
Логістична регресія	0.9836	0.9750	0.9953	0.9850
Text CNN	0.9785	0.9652	0.9962	0.9805
BERT Transformer	0.9985	0.9991	0.9981	0.9986

- Модель BERT демонструє найвищу точність та F1-score, що підтверджує її здатність добре узагальнювати і мінімально помилятись при класифікації фішингових листів.
- Text CNN показала дуже хороші результати, але трохи поступається BERT за стабільністю прогнозів.
- Логістична регресія — проста й швидка для впровадження, підходить як базове рішення, проте має нижчу ефективність порівняно з більш складними моделями.

Висновки

Побудовано систему для автоматичного визначення фішингових електронних листів на основі тексту.
Проведено порівняння трьох підходів з різною складністю (логістична регресія → Text CNN → BERT).
Найкраще показав себе підхід із трансформером (BERT) — рекомендовано для впровадження в реальних системах.
Text CNN може бути компромісом між складністю і точністю, коли ресурси обмежені.
Логістична регресія може служити швидкою «першою лінією» аналізу, але не замінює глибші моделі у задачах підвищеної відповідальності.

ДЯКУЮ ЗА УВАГУ