УДК 004.8:004.91

ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ СИНТЕТИЧНИХ ТЕКСТІВ

Безродний В.В.
Науковий керівник – доцент каф. ПІ, к.т.н., доцент Турута О.П.
Харківський національний університет радіоелектроніки, каф. ШІ
м.Харків, Україна

The topic of synthetic text detection has become increasingly important in recent years due to the widespread use of text-generating models and the potential for malicious actors to use them for spreading misinformation and propaganda. This research aims to identify effective methods for detecting synthetic texts, which can include both fully synthetic texts generated by AI models and partially synthetic texts that have been manipulated or edited. It involves exploring various techniques, such as statistical analysis, machine learning, and natural language processing, to analyze and compare the characteristics of synthetic and human-generated texts.

Синтетичні тексти – це тексти, які були згенеровані комп'ютерними програмами та алгоритмами, а не написані людиною. Розглянемо основні методи визначення синтетичних текстів.

Bag-of-words класифікатор. Деякі детектори використовують класичні методи машинного навчання, такі як логістична регресія, щоб навчити модель з нуля розрізняти тексти. Наприклад, використовують просту базову модель, яка представляє документ із вектором tf-idf (уніграми та біграми) поверх моделі логістичної регресії. Покоління більших моделей важко виявити порівняно з меншими моделями, що вказує на те, що чим більша модель, тим ближчий стиль створеного тексту до тексту, написаного людиною.

Zero-shot класифікатор. У налаштуваннях zero-shot класифікації попередньо навчена модель (наприклад, GPT-2, GROVER) використовується для виявлення поколінь від себе чи подібних моделей. Детектор не потребує контрольованих прикладів виявлення для подальшого навчання. Загальна логарифмічна ймовірність представляють базову лінію, яка використовує модель для оцінки загальної логарифмічної ймовірності, а також порогові значення на основі цієї ймовірності, щоб зробити прогноз. Наприклад, текст передбачається як згенерований машиною, якщо загальна ймовірність тексту ближча до середньої ймовірності для всіх текстів, згенерованих машиною, ніж до середньої ймовірності текстів, написаних людиною. Однак, цей класифікатор може працювати погано порівняно з класифікатором на основі логістичної регресії.

Інструмент Giant Language Model Test Room (GLTR). Інструмент GLTR пропонує набір базових статистичних методів, які можуть

підкреслити відмінності в розподілі тексту, створеного моделлю, і тексту, написаного людиною. Зокрема, GLTR дає змогу вивчати фрагмент тексту шляхом візуалізації ймовірності моделі кожного маркера, рангу кожного маркера в прогнозованому розподілі наступного маркера та ентропії прогнозованого розподілу наступного маркера. На основі цих візуалізацій інструмент чітко показує, що TGM надмірно генеруються з обмеженої підмножини справжнього розподілу природної мови. Дійсно, рідкісне використання слів у тексті, згенерованому моделлю, помітно менше порівняно з текстом, написаним людиною. Інструмент дозволяє людям (включаючи неекспертів) вивчати фрагмент тексту, але може бути менш ефективним у майбутньому, коли моделі почнуть генерувати текст без статистичних аномалій.

GROVER детектор. пропонується детектор на основі лінійного класифікатора на основі моделі GROVER, який перевершує існуючі детектори і таким чином зробити висновок, що найкращі моделі для генерації нейронної дезінформації також є найкращими для виявлення власних поколінь. Стандартний детектор GROVER погано працює при виявленні тексту, створеного моделями, відмінними від оригінальної моделі GROVER.

RoBERTa, BERT та інші трасформери. Детектор RoBERTa, навчений на прикладах top-p, добре переносить приклади з усіх інших методів декодування. Детектор добре працює під час навчання на прикладах із більшої моделі GPT і добре переносить приклади, згенеровані меншою моделлю GPT. З іншого боку, навчання на менших виходах моделі GPT призводить до поганої продуктивності при класифікації більших виходів моделі. Найцікавішим висновком цієї є те, що точне налаштування за допомогою моделі RoBERTa досягає вищої точності, ніж точне налаштування моделі GPT з еквівалентною потужністю.

На сьогоднішній день, RoBERTa є одним з найбільш ефективних і точних моделей для детекції синтетичних текстів. Вона використовує масштабовану архітектуру та глибоке навчання, щоб виявляти найдрібніші відмінності між синтетичним та натуральним текстом.

Список використаних джерел:

1. Ganesh Jawahar. Automatic Detection of Machine Generated Text: A Critical Survey / Ganesh Jawahar, Muhammad Abdul-Mageed, Laks V.S. Lakshmanan. – 2020. – https://aclanthology.org/2020.coling-main.208.pdf

2. Andres Garcia-Silva. Understanding Transformers for Bot Detection in Twitter / Andres Garcia-Silva, Cristian Berrio, Jose Manuel Gomez-Perez. – 2021. – https://arxiv.org/pdf/2104.06182.pdf

3. Sebastian Gehrmann. GLTR: Statistical Detection and Visualization of Generated Text / Sebastian Gehrmann, Hendrik Strobelt, Alexander M. Rush. – 2019. – https://arxiv.org/pdf/1906.04043.pdf