УДК 65.011.56

В. В. Маліцький
С. Д. Штовба

АВТОМАТИЗАЦІЯ ВИЯВЛЕННЯ ПРИХОВАНОГО ТОКСИЧНОГО КОНТЕНТУ В ТЕКСТОВИХ ПОВІДОМЛЕННЯХ

Вінницький національний технічний університет

Анотація

Досліджено питання детекції прихованого токсичного контенту в текстових повідомленях. Розроблено дві моделі глибинного навчання на основі архітектури BERT, що здатні визначати загальний рівень токсичності повідомлення та визначати замасковані лайливі слова у повідомленнях.

Ключові слова: глибинне навчання, обробка природної мови, нейронна мережа, BERT.

Abstract

The task of hidden toxic content detection in text messages has been discovered. Two in-depth learning models based on the BERT architecture have been developed, which are able to determine the overall level of toxicity of the message and identify masked swear words in messages..

Keywords: Deep Learning, Natural Language Processing, Neural Network, BERT.

Вступ

Текстові повідомлення відіграють важливу роль у комунікації людей, і з початком цифрової епохи ця роль стає дедалі більшою. Повідомлення стали основною формою спілкування між малознайомими людьми у величезній кількості інтернет-майданчиків, що породжує питання модерації цього спілкування. Кожного року соціальні платформи на кшталт Facebook, Twitter, Instagram та інші витрачають мільйони доларів на підтримку належної модерації та фільтрування агресивних повідомлень. Кожна платформа зобов’язана забезпечувати належний рівень модерації опублікованих повідомлень та контенту загалом, що регулюється державними актами, недотримання яких загрожує великими штрафами та навіть блокуванням. Автоматизована система, що здатна своєчасно визначати та не допускати до публікації токсичний контент є гарною альтернативою людям, які здатні стомлюватись, помилятись та отримувати психологічну шкоду від цієї роботи. Наведені вище факти переконують у актуальності даної бакалаврської дипломної роботи на тему «автоматизація виявлення прихованого токсичного контенту в текстових повідомленнях».

Об’єктом роботи є визначення та фільтрація повідомлень, що явно чи приховано є токсичними.

Метою дослідження є розробка двох моделей машинного навчання, перша з яких здатна визначати загальну токсичність повідомлення, а друга – визначати замасковані лайливі слова у повідомленнях.

Для вирішення поставлених в роботі задач використовувалися методи глибинного навчання. Для розробки програмної частини системи вихідного контролю використовувалися методи алгоритмізації та програмування.

Результати дослідження

Для вирішення задачі було створено дві моделі глибинного навчання на основі архітектури BERT (Bidirectional Encoder Representations from Transformers), що була представленна Google у 2019 році [1], архітектуру якої схематично наведено на рисунку 1. Головною революційною перевагою цієї архітектури було те, що вона знаходить своє застосування у величезній кількості задач, пов’язаних з обробкою природної мови, а саме задачі класифікації тексту, пошуку іменованих сутностей у тексті, перекладі з однієї мови на іншу та генерації відповідей на питання. Її універсальність дозволяє використати цю модель і для пошуку токсичності у текстових повідомленнях.

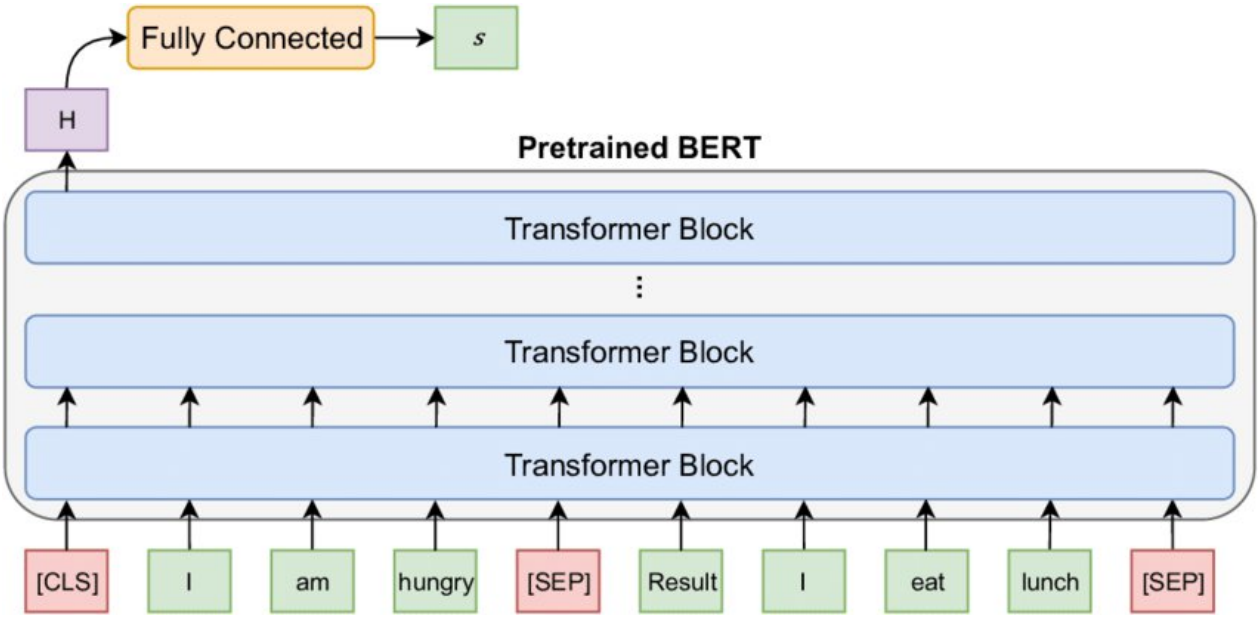

Рисунок 1 – Архітектура моделі BERT

Загалом було розроблено дві моделі, одна з яких, отримуючи на вхід повідомлення, оцінює його загальну токсичність від 0 до 1, а друга модель визначає слова, що є замаскованими лайливими словами. Вона видає ймовірність від 0 до 1 кожному слову у реченні, де 1 означає, що модель впевнена, що слово є лайливим.

Для виконання цієї задачі було модифіковано датасет повідомлень з соціальної мережі Civil Comments, яка після припинення своєї роботи видала у вільний доступ близько 2 000 000 коментарів, що залишили їхні користувачі [2]. Цей датасет вручну промаркували анотатори, оцінивши токсичність кожного повідомлення. Однак вони не промаркували лайливі слова. Тому цим я зайнявся у рамках цієї роботи, знайшовши слова, що виникали у словниках лайливої лексики. Також я замінив 50% найпопулярніших лайливих слів їхніми замаскованими формами з рисунку 2, адже користувачі нечасто маскували лайку у своїх повідомленнях.

	text	canonical_form_1	canonical_form_2	canonical_form_3	category_1	category_2	category_3	severity_rating	severity_descri
2	69	69			sexual anatomy / sexual acts			1	Mild
3	@55	ass			sexual anatomy / sexual acts			1	Mild
4	@ssfcker	fuck	ass		sexual anatomy / sexual acts	sexual orientation / gender		2.8	Severe
5	@ssfucker	fuck	ass		sexual anatomy / sexual acts	sexual orientation / gender		2.8	Severe
6	@ssfvcker	fuck	ass		sexual anatomy / sexual acts	sexual orientation / gender		2.4	Strong
7	@sshole	ass			sexual anatomy / sexual acts			1.6	Strong
8	0ral seks	sex			sexual anatomy / sexual acts			1	Mild
9	0ral sex	sex			sexual anatomy / sexual acts			1.8	Strong
10	0rg@sm	orgasm			sexual anatomy / sexual acts			1	Mild
11	0rgasms	orgasm			sexual anatomy / sexual acts			1	Mild
12	3jakulating	ejaculation			sexual anatomy / sexual acts	bodily fluids / excrement		1.6	Strong
13	4r5e	arse			sexual anatomy / sexual acts			1.4	Mild
14	4r5ed	arse			sexual anatomy / sexual acts			1.4	Mild
15	4r5es	arse			sexual anatomy / sexual acts			1.4	Mild
16	4skin	foreskin			sexual anatomy / sexual acts			1	Mild
17	5h17	shit			bodily fluids / excrement			1	Mild
18	5h1t	shit			bodily fluids / excrement			1	Mild
19	5kank	skank			sexual orientation / gender			2	Strong

Рисунок 2 – Список замаскованих форм лайливих слів

Знаючи усі лайливі слова у датасеті, я зміг промаркувати кожне слово нулями та одиницями, де одиниця означала лайливе слово. Це було необхідно, щоб надати нейромережі інформацію про те, яке слово є лайливим, а яке ні. Також я видалив з набору данних більшість повідомлення, щоб подолати проблему дизбалансу лайливих слів, яких набагато менше за звичайні. В залишку, у

датасеті залишилося 80 000 повідомлень, де у половині були лайливі слова. Повідомлення, що залишилися, я розділив на тренувальну, валідаційну та тестову вибірки у пропорціях 70/15/15%

Для тренування моделей глибинного навчання я використав фреймворк глибинного навчання PyTorch [3], який дозволяє будувати дуже гнучкі моделі машинного навчання і контролювати кожен їх аспект. Кожну модель я будував на основі вже претренованої версії BERT, що була навчена на даних англомовної Вікіпедії та корпусу класичних книг. Дотренування моделей відбувалось протягом однієї епохи (одного проходження через тренувальну вибірку). Обидві моделі показали гарні результати за точністю та F1 мірою.

Висновки

Під час написання бакалаврської дипломної роботи було проведено аналіз проблеми визначення токсичності у текстових повідомленнях у контексті соціальних мереж. Було розроблено дві моделі глибинного навчання з архітектурою BERT, що здатні оцінювати рівень токсичності текстового повідомлення та детектувати лайливі слова, в тому числі і замасковані за допомогою замін одного чи кількох символів. Також було модифіковано датасет коментарів користувачів соціальної мережі задля виконання поставлених задач.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding;
2. Kaggle - Jigsaw Unintended Bias in Toxicity Classification [Електронний ресурс] – Режим доступу до ресурсу: https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data/;
3. PyTorch – From Research to Production [Електронний ресурс] – Режим доступу до ресурсу: https://pytorch.org/;

Маліцький Вадим Валентинович *– студент групи КІВ-16б, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця. e-mail: vadym.malitskyi357@gmail.com*

Штовба Сергій Дмитрович *- к.т.н., доцент кафедри комп'ютерних систем управління, Факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, м. Вінниця. e-mail: shtovba@vntu.edu.ua*