DOI: https://doi.org/10.26565/2304-6201-2024-64-01

УДК (UDC) 004.94

Тітаренко Тімур *магістр; Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022*
e-mail: tima.tytarenko.001@gmail.com
https://orcid.org/0000-0001-6417-151X

Толстолузька Олена Геннадіївна *д. т. н., с. н. с.; професор кафедри теоретичної та прикладної системотехніки; Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 4, Харків-22, Україна, 61022;*
e-mail: elena.tolstoluzka@karazin.ua;
https://orcid.org/0000-0003-1241- 7906.

Узлов Дмитро Юрійович *к.т.н., доцент закладу вищої освіти кафедри теоретичної та прикладної інформатики*
Харківський національний університет імені В .Н. Каразіна, майдан Свободи, 4, Харків, Україна, 61022
e-mail: dmytro.uzlov@karazin.ua
https://orcid.org/0000-0003-3308-424X

Модель нейронної мережі для цензурування текстових даних

Актуальність: в умовах стрімкого розвитку інтернет-комунікацій та зростання обсягів текстового контенту, актуальність роботи обумовлена необхідністю забезпечення ефективного цензурування текстових даних. Особливо в онлайн середовищі, де важливо забезпечити безпеку та етичність спілкування.
Мета: забезпечити більш якісний та безпечний контент для користувачів, які залежать від надійної та безпечної інформації в Інтернеті, за допомогою розробки та впровадження нейронної мережі, яка буде здатна визначати недопустимий контент у текстових дані в реальному часі.
Методи дослідження: в ході виконання досліджень були використані методи обробки та підготовки даних, методи глибокого навчання, теорія нейронних мереж, теорія штучного інтелекту, математичний аналіз, методи аналізу інформативності, методи оцінки якості класифікації, дослідження практичного застосування. Програмне забезпечення розроблено за допомогою мови Python. Також були використані наступні бібліотеки: keras, sklearn, pandas і інші.
Результати: головним результатом роботи була розробка моделі нейронної мережі, яка цензурує текстові дані в реальному часі, модель виявляється високо масштабованою і готовою до навчання на даних інших мов.
Висновки: розглянуто проблему цензурування текстових даних. Оскільки це завдання обробки природної мови, було запропоновано та розроблено модель нейронної мережі на основі RNN, а саме LSTM. Дослідження засвідчило важливість інноваційних підходів у вирішенні проблем цензури текстових даних, а використання нейронних мереж та технологій штучного інтелекту стає перспективним напрямком для подальших досліджень та впроваджень у цій області.
***Ключові слова**: нейронні мережі, цензурування тексту, LSTM, NLP, класифікація текстових даних.*

Як цитувати: Тітаренко Т., Толстолузька О. Г., Узлов Д. Ю. Модель нейронної мережі для цензурування текстових даних. *Вісник Харківського національного університету імені В.Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2023. вип. 60. С.52-58. https://doi.org/10.26565/2304-6201-2023-60-02

How to quote: Tytarenko T., Tolstoluzka O., Uzlov D., “Using anomaly detection method to detect network attack”, *Bulletin of V.N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 60, pp.52-58, 2023. https://doi.org/10.26565/2304-6201-2023-60-02 [In Ukrainian].

1. Вступ

У сучасному інформаційному суспільстві, де потоки текстової інформації невпинно зростають, важливість ефективної фільтрації та цензурування текстових даних визнається як ключова область для забезпечення безпеки, конфіденційності та етичності в онлайн середовищі. Непередбачуваність та динаміка змін у сучасному мовному просторі створюють виклик для

розробки інноваційних моделей, які здатні точно та автоматично визначати неприпустимий контент, забезпечуючи безпеку та етичність онлайн-спілкування.

2. Етичні, соціальні та технічні проблеми

- Цензура може порушити принцип свободи слова, який вважається ключовим для демократичних суспільств. Обмеження доступу до певної інформації може призвести до обмеження свободи вираження та обміну ідеями.
- Суб'єктивність виникає з того, що визначення того, що є неприйнятним або образливим, може значно варіюватися залежно від культур, групи людей чи індивідуальних переконань. Одна й та ж сама фраза чи вислів може бути сприйнятими різними людьми по-різному. Спроби визначити "неприйнятний" вміст можуть враховувати різні аспекти.
- Неспроможність визначити контекст: автоматизовані системи цензури, зокрема засновані на штучних нейронних мережах, можуть мати складнощі в розумінні контексту, що може призвести до помилкового блокування чи фільтрації текстів.
- Алгоритми цензури можуть стикатися із складнощами в розрізненні іронії, гумору, або того, як певний вислів може змінювати своє значення в залежності від контексту.
- Обхід цензури: цей аспект вказує на те, що користувачі можуть використовувати різні техніки для ухилення від систем цензури та отримання доступу до забороненого або обмеженого контенту. Обхід цензурних обмежень може стати проблемою для тих, хто намагається контролювати доступ до певного вмісту. Одним зі шляхів є використання альтернативних слів, користувачі можуть змінювати слова чи використовувати схожі терміни, щоб уникнути фільтрів. Маскування контенту, використання специфічних символів, реєстрації чи інших маскувальних технік для ухилення від фільтрів.
- Виклики в області техніки: обробка природної мови виявляється складним завданням, оскільки воно вимагає розуміння не лише синтаксичних та семантичних правил, але і врахування великого різноманіття мовних виразів та ідіом.

Використання штучного інтелекту, зокрема штучних нейронних мереж, виявляється найбільш ефективним у вирішенні викликів, пов'язаних із цензурою текстових даних. Штучні нейронні мережі володіють здатністю ефективно адаптуватися до складних структур текстів та контекстів, що робить їх особливо корисними у виявленні та фільтрації небажаного контенту. Вирішення цих викликів потребує впровадження передових методів обробки природної мови, застосування технологій машинного навчання, а також постійного навчання систем для адаптації до змін у мовному вживанні та соціальних стандартах. Безперечно, розробка та вдосконалення таких технологій повинні враховувати етичні аспекти, такі як конфіденційність даних та забезпечення свободи вираження.

3. Обґрунтування та вибір типу нейронної мережі: LSTM (Long Short-Term Memory)

- *Урахування контексту:*
  LSTM вміє зберігати та використовувати інформацію з попередніх кроків в послідовності. Це особливо важливо для аналізу тексту, де розуміння контексту та зв'язків між словами грає велику роль.
- *Уникнення проблеми зниклих градієнтів:*
  LSTM вирішує проблему зниклих градієнтів, яка може виникати при тренуванні глибоких нейронних мереж. Це робить його ефективним в роботі з великими обсягами текстових даних.
- *Робота з послідовностями різної довжини:*

LSTM може працювати з послідовностями різної довжини, що важливо для роботи текстовими даними, де речення можуть мати різну кількість слів.[1-2].

На рисунку 1 відображене схематичне представлення мережі з довгою короткостроковою пам’яттю.

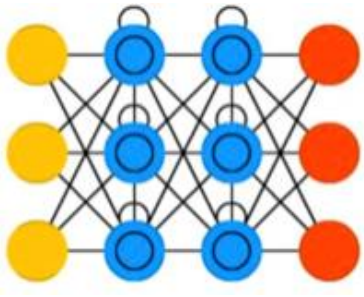


Рисунок 1 – Схематичне представлення мережі з довгою короткостроковою пам’яттю

4. Архітектура нейронної мережі

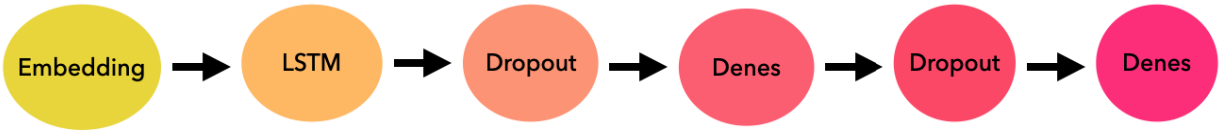
На рисунку 2 в схематичному вигляді представлена розроблена архітектура нейронної мережі для цензурування текстових даних. Дамо короткий опис її компонентів.

- *Embedding Layer:*
  Використовується для перетворення словесного тексту в числовий формат, що може бути використано нейронною мережею.
- *LSTM Layer:*
  LSTM шар використовується для аналізу послідовності векторів, які представляють вбудовані слова.
- *Dropout Layer (перший):*
  Dropout шари використовуються для уникнення перенавчання шляхом випадкового "вимикання" частини нейронів під час тренування.
- *Dense Layers (перший):*
  Відповідає за обробку та витягування ключових ознак з інформації, що була згенерована попереднім LSTM шаром[3].
- *Dropout Layer (другий):*
  Використовується як і попередній *Dropout* для запобігання перенавчанню.
- *Dense Layers (другий):*

Застосовується для класифікації тексту на кілька класів. Допомагає у виразному представленні вихідної інформації для рішення задачі цензурування.

Враховуючи вищезазначені фактори, обрана архітектура LSTM забезпечує ефективне виявлення та цензурування неприпустимого контенту у текстових даних, здатність адаптуватися до різних контекстів та уникнення проблем зниклих градієнтів під час тренування.[4-5]

Рисунок 2 – Схематичне представлення архітектури нейронної мережі



5. Навчання нейронної мережі, та тестування

Процес навчання та тестування нейронної мережі передбачає виконання декількох кроків.

- Попередня обробка:
  Очищення даних: Вилучення непотрібних символів, HTML-тегів чи спеціальних символів.
- Розділення даних:
  Розподіл даних на тренувальний, валідаційний та тестовий набори для ефективного тренування та оцінки моделі.

- Токенізація та доповнення:

  Токенізація - це етап обробки тексту, під час якого вхідний текст розділяється на окремі одиниці, так звані токени. Токени можуть бути словами, фразами або іншими одиницями тексту. В контексті цензурування токенізація допомагає розділити текстові фрази на окремі слова, що стає важливим етапом перед подальшим перетворенням тексту для використання в нейронній мережі. Токенізація може бути виконана різними способами. Зазвичай вона включає в себе видалення зайвих символів, поділ тексту на окремі слова та створення токенів. Наприклад, речення «Це речення для цензурування» може бути розділене на токени: ["Це", "речення", "для", «цензурування"].[6].

  Після токенізації, текстові токени потрібно перетворити в числові значення, які можуть бути подані нейронною мережею. Цей етап називається трансформацією тексту в числові послідовності. Цей етап дозволяє побудувати числовий вектор для кожного текстового фрагменту, що є необхідним для подальшого навчання нейронної мережі.

- Імплементація моделі та навчання:

  Імплементація моделі: Імплементуємо розроблену архітектуру (рис. 2) до програмної реалізації. Після створення архітектури моделі, наступний крок - це її компіляція, тобто конфігурація процесу навчання. Компіляція включає в себе вибір оптимізатора, функції втрат та метрик, які використовуються для оцінки продуктивності моделі під час навчання. Після відібрання та підготовки наборів даних, модель переходить до етапу навчання. У цьому етапі використовується тренувальний набір даних, і модель намагається оптимізувати свої параметри за допомогою методу градієнтного спуску та зворотнього поширення помилки. Тренування включає в себе подачу текстових даних в мережу, оцінювання виходів, порівняння їх з очікуваними значеннями та коригування ваг моделі для покращення її точності.

- Оцінка:

  Валідація та тестування: Оцінка моделі на валідаційному наборі для налаштування параметрів та на тестовому наборі для оцінки загальної ефективності, та оцінка продуктивності навченої моделі. Для цього використовуються різні метрики, щоб отримати об'єктивне уявлення про те, наскільки добре модель справляється з поставленим завданням цензурування тексту[7]. На рисунку 3 наведено приклад даних на яких навчається модель.

	count	hate_speech	offensive language	neither	class	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. & as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya 
5	3	1	2	0	1	!!!!!!!!!!!!!!!!"@T_Madison_x: The shit just blows me..claim you so faithful and down for somebody but still fucking with hoes! 😂😂😂"
6	3	0	3	0	1	!!!!!!"@__BrighterDays: I can not just sit up and HATE on another bitch .. I got too much shit going on!"
7	3	0	3	0	1	!!!!“@selfiequeenbri: cause I'm tired of you big bitches coming for us skinny girls!!”
8	3	0	3	0	1	" & you might not get ya bitch back & thats that "
9	3	1	2	0	1	" @rhythmixx_ :hobbies include: fighting Mariam" bitch
10	3	0	3	0	1	" Keeks is a bitch she curves everyone " lol I walked into a conversation like this. Smh

Рисунок 3 – Приклад даних на яких навчається модель(твіти)

Після відібрання та підготовки наборів даних, модель переходить до етапу навчання. У цьому етапі використовується тренувальний набір даних, і модель намагається оптимізувати свої параметри за допомогою методу градієнтного спуску та зворотнього поширення помилки.

Тренування включає в себе подачу текстових даних в мережу, оцінювання виходів, порівняння їх з очікуваними значеннями та коригування ваг моделі для покращення її точності.

Окремий валідаційний набір даних в розробленій моделі використовується для перевірки, наскільки добре модель генералізує свої знання на нових, раніше не бачених даних. Валідація дозволяє визначити ступінь, до якої модель уникнула перенавчання (overfitting) або, навпаки, недонавчання (underfitting).

Після кожного циклу тренування, коли модель пройшла через всі дані тренувального набору, проводиться валідація на валідаційному наборі.

Результати валідації служать орієнтиром для покращення параметрів моделі та вдосконалення її продуктивності.

Цей етап є важливим для створення моделі, яка не лише ефективно працює на тренувальних даних, але і може адекватно застосовувати свої знання до нових ситуацій. Метрики моделі нейронної мережі після навчання представлено в таблиці 1.

Таблиця 1 - Метрики моделі нейронної мережі після навчання

Набір даних/метрики	Precision	AUC-ROC	Loss
Початкова точність	0.8359	0.8088	0.4790
Кінцева точність	0.8902	0.8995	0.0390

Аналізуючи результати експериментів з різними наборами гіперпараметрів для моделі, можна зробити декілька важливих висновків. Найефективнішою виявилася конфігурація з LSTM = 128, Dense = 256, Batch_size = 64, Epochs = 5 та швидкістю навчання Adam = 0.01. Цей набір параметрів дозволив досягти високої точності (Precision 0.8802) та високого показника AUC-ROC (0.8895), що свідчить про високу якість моделі в класифікації та роботу з різними класами (рис.4).

Також важливо зазначити, що збільшення кількості LSTM-шарів не завжди призводить до поліпшення результатів, а зменшення кількості може вплинути на ефективність моделі. Збільшення кількості епох не завжди призводить до покращення, а швидкість навчання може виявитися ключовим фактором: збільшення швидкості може призвести до зростання точності та AUC-ROC.

Додатково, важливо відмітити стабільність моделі при зміні швидкості навчання. Модель показала стійкість результатів при зменшенні швидкості навчання (Adam = 0.0001), що може бути корисним при роботі з великими та складними наборами даних.

У кінці, експерименти з гіперпараметрами вказують на необхідність систематичного та пристосованого підходу до вибору параметрів для кожної конкретної задачі та датасету.

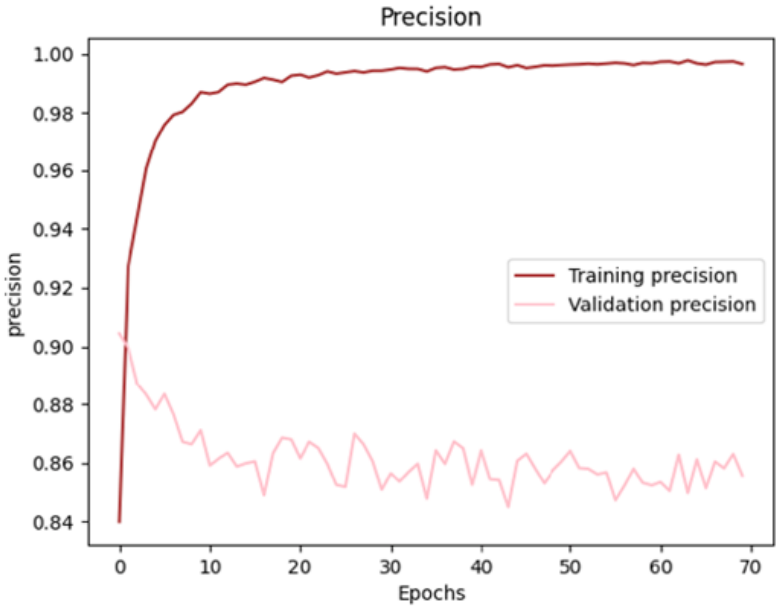


Рисунок 4 – Графік класової точності моделі

6. Висновки

Головним результатом даної роботи була розробка моделі нейронні мережі, яка цензурує текстові дані в реальному часі, модель виявляється високо масштабованою і готовою до навчання на даних інших мов. Її архітектура, з використанням LSTM-шарів та ембедінгів, дозволяє ефективно враховувати контекст та взаємодію між словами у текстах. Оскільки використовується токенізація тексту та побудова словника, модель може адаптуватися до різних мовних варіацій. Змінюючи навчальний датасет на дані іншої мови, можна досягти гарної адаптації, оскільки модель вивчає структуру мови та зв'язки між словами. Такий підхід робить модель гнучкою та застосовною до різних завдань та мовних середовищ.

Головним напрямком використання моделі є цензурування текстових даних в онлайн середовищі, де потік текстової інформації росте експоненційно, і ефективна фільтрація та цензура текстових даних стають визначальними аспектами для забезпечення безпеки, конфіденційності та етичності.

Модель демонструє високий рівень ефективності та продуктивності за рахунок розробленої архітектури, різноманіття та якість даних для навчання, та обраних гіперпараметрів. Архітектура складається з LSTM-шарів, ембедінгів та дропаут-шарів які ефективно працюють з текстовими даними, про це свідчить висока точність 88% та високий показник AUC-ROC, значення якого складає 0.8895. Дані для навчання, являють собою 25 тисяч текстових повідомлень з мережі Twitter, що попередньо оброблені, а саме видалені HTML-сутності, URL-адреса та інші непотрібні символи. Проведені експерименти та аналіз результатів дозволили визначити оптимальні гіперпараметри моделі.

Також модульна структура та архітектура програмної реалізації дозволяють легко інтегрувати та адаптувати модель для використання її на різних платформах. Як практичний приклад, модель була успішно імплементована в середовищі Telegram. Це дозволяє в реальному часі аналізувати та цензурувати текстовий контент, що надходить через цю платформу.

На етапі вибору типу моделі нейронної мережі були розглянуті та порівняні різні архітектури, такі як персептрон, багатошаровий персептрон, згорткова нейронна мережа, рекурентні нейронні мережі, зокрема, LSTM, CNN та RNN.

СПИСОК ЛІТЕРАТУРИ

1. Zhang, X. A Mathematical Model of a Neuron with Synapses based on Physiology. Nat Prec (2008). https://doi.org/10.1038/npre.2008.1703.1

2. Sima J. “Introduction to Neural Networks,” Technical Report No. V 755, Institute of Computer Science, Academy of Sciences of the Czech Republic, 1998.
3. Kröse B., and van der Smagt P. An Introduction to Neural Networks. (8th ed.) University of Amsterdam Press, University of Amsterdam, 1996.
4. Harshali M. Deep learning — 2015.
5. Yingci L. Zhonghua C. Hongkai W. Peiou L. Zongwei Z. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT — 2017
6. Aurélien Géron. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 3rd Edition
7. Oliver Theobald, Best Machine Learning Books for Absolute Beginners, 2021 - 50 c.

Tytarenko Tymur *Master student; V.N. Karazin Kharkiv National University, Svobody Square, 4, Kharkiv-22, Ukraine, 61022.*
e-mail: tima.tytarenko.001@gmail.com
https://orcid.org/0000-0001-6417-151X

Tolstoluzka Olena *Doctor of Engineering Sciences; Professor of Theoretical and Applied Systems Engineering Department; V.N. Karazin Kharkiv National University, Svobody Square, 4, Kharkiv-22, Ukraine, 61022.;*
e-mail: elena.tolstoluzka@karazin.ua
https://orcid.org/0000000312417906

Uzlov Dmitro *Associate Professor of the Department of Theoretical and Applied Informatics, Faculty of Mathematics and Informatics, Ph.D. V.N. Karazin Kharkiv National University, Svobody Square, 4, Kharkiv-22, Ukraine, 61022.*
https://orcid.org/0000-0003-3308-424X

The model of a neural network for text data censoring

Relevance: Given the rapid development of Internet communications and the increasing amount of textual content, an urgent need to ensure effective censorship of textual data necessitates the relevance of this research. This is especially true for the online community, where it is essential to ensure the security and ethics of communication.
Purpose: to provide better and safer content for users who depend on reliable and secure Internet information by means of developing and implementing a neural network that will be able to identify inappropriate textual content in real time.
Research methods: methods of data processing and preparation, deep learning methods, neural network theory, artificial intelligence theory, mathematical analysis, methods of information content analysis, methods of classification quality assessment, and practical application research have been used in the course of the research. The software has been developed by using the Python language.
Results: the main achievement of the work is the development of a neural network model that censors textual information in real time, the model is highly scalable and can be trained on data from other languages.
Conclusions: The problem of text data censoring has been considered. Since this is a natural language processing task, an RNN-based neural network model, namely LSTM, has been proposed and developed. The study has shown the importance of innovative approaches in solving the problems of text data censorship, and the use of neural networks and artificial intelligence technologies is becoming a promising area for further research and implementation in this area.

Keywords*: neural networks, text censorship, LSTM, NLP, text data classification.*

Надійшла у першій редакції 14.10.2023, в останній – 22.11.2023.
The first version has been received on 14.10.2023, the final version – on 22.11.2023.