DOI 10.31891/2307-5732-2023-323-4-118-121
УДК 621

ДУПЛЯК СТЕПАН.
Національний університет «Львівська політехніка»
ORCID: 0000-0002-9240-404X
e-mail: stepan.dupliak.knm.2019@lpnu.ua
ШАХОВСЬКА НАТАЛІЯ.
Національний університет «Львівська політехніка»
ORCID: 0000-0002-6875-8534
e-mail: Nataliya.b.shakhovska@lpnu.ua

ОЦІНКА АДЕКВАТНОСТІ КОНТЕНТУ ЗА КОНТЕКСТОМ МЕТОДАМИ АНСАМБЛІВ МОДЕЛЕЙ BERT

Розроблений у роботі ансамбль моделей машинного навчання для аналізу емоційності новин натренований на різних контекстах та незалежних наборах даних. Здійснено голосування по кожній моделі з ансамблю за особистий варіант правди згідно з локальним контекстом тієї моделі. Розроблено бінарний класифікатор адекватності/нормальності повідомлень на базі технології ансамблів. Можна спостерігати вибраний нами набір методів, які є оптимальними за параметрами часу виконання, часу тренування, обсягом оперативної пам'яті та відповідно точністю. Зокрема це такі методи, як Catboost XGBoost для класифікації та екстракції особливостей та контексту було обрано BERT та його підвид RoBERTa. Аналіз результатів показав, що точність алгоритму коливається від 80% до 85 % в ансамблі та від 65% до 93% окремими методами за окремими наборами даних.

Ключові слова: нейронні мережі, BERT, RoBERTa, Catboost, XGBoost.

DUPLIAK STEPAN, SHAKHOVSKA NATALIA
Lviv Polytechnic National University

ASSESSMENT OF ADEQUACY OF CONTENT BY CONTEXT USING BERT MODEL ENSEMBLE METHODS

Over the past 20 years, text has dominated the Internet as a means of communicating information. Every day, new people are born and every day, new people sign up for social media. Due to lack of education and attention, people use social media to express their thoughts and virtualize themselves, sometimes forgetting that there is another person on the other side of the monitor. Such processes in human life lead to the reckless or sometimes intentional generation of content that may violate the rules of the communities where this content is produced. One of the primitive and non-scalable examples of dealing with the problem of uncontrolled generation of social content is physical moderation. This method makes sense in private, closed channels of communication with the audience, where the bandwidth of a person as a moderator is sufficient to effectively control the information space. Despite the reliability of humans in terms of information and moderation, humans are lifelong learners, and they are what they read or see. Therefore, there is a possibility that human moderation is biased from the point of view of all people who are in the same information space. The topic of assessing the adequacy and ethics of a text is gaining popularity even as the amount of information generated in social networks increases. The problem is that modern methods of text evaluation are not able to work with differently contextualized data, i.e. a model trained on one data set is tied to the context of the data environment in which this set was collected. The method developed in this paper allows a model to be trained on different contexts and independent datasets, and to directly vote each model in the ensemble for its own version of the truth according to the local context of that model. We will develop a binary message adequacy/normality classifier based on ensemble technology. You can observe the set of methods I have chosen that are optimal in terms of execution time, training time, RAM, and, accordingly, accuracy. In particular, these are methods such as Catboost XGBoost for classification and extraction of features and context, BERT and its sub-type RoBERTa were chosen. I will conduct a corresponding analysis and experiment on these methods to verify that this method is really effective.

Keywords: neural networks, BERT, RoBERTa, Catboost, XGBoost.

Вступ

Інтелектуальний агент як метод фільтрації та модерації даних почав з'являтись рівномірно з тим, як зростала валова кількість продукованої інформації за момент часу [1–4]. Крім базових речей, таких як пошук неприйнятих слів, виразів, абревіатур. машинний інтелект почав розуміти інформацію в часі та сприймати інформацію не як незалежні змінні в просторі, але як потік інформації, який має чітко впорядкований зв'язок. Цим можна завдячувати машинному навчанню та рекурентним моделям. З часом машинне навчання вперлось в стелю контексту інформації. Інформація, написана людиною, на жаль не є сухими даними. Людям у соціальних мережах притаманно говорити сарказмом або надавати умовних позначень певним об'єктам, ознакам або навіть процесам, які називаються по різному, але означають одне й теж. Щоб класифікувати текстові дані на предмет негативного висловлювання чи образ, сучасні методи LSTM [5]чи RNN [6] та їх комбінації вже не витримували критики ефективності. Попитом на подібні алгоритми зростав, виникли алгоритми трансформери з них сімейство алгоритмів BERT [2] та його варіації. Це стало початком нових алгоритмів трансформерів. Особливістю цих методів є їхня будова, яка містить так зване маскування. При тренуванні мережі фрагменти речень замінюється на прогалини, та мережа намагається інтерполювати та екстраполювати слова в реченні без втрати контексту. Коли мережа натренована, вона може вільно класифікувати контекст повідомлення навіть якщо контекст є неочевидним.

Мета цієї роботи полягає в тому, щоб навчати мережі BERT окремо на різних наборах даних та скласти їх в один ансамбль, який буде містити оцінку різних контекстів та могти оцінювати адекватність повідомлень від вже більш глобального контексту. Оцінка здійснюватиметься голосуванням всіх мереж та обирання за певним правило середнього.

Аналіз літературних джерел

Статистичні алгоритми використовували методи машинного навчання, такі як наївний Байєс [1], машини опорних векторів (SVM [2]) і дерева рішень, для класифікації тексту. Статистичні алгоритми були більш гнучкими, ніж системи, засновані на правилах, і могли навчатися на даних, що робило їх більш точними.

Наступним важливим кроком в еволюції класифікації тексту стало впровадження алгоритмів глибокого навчання. Зокрема, згорткові нейронні мережі (CNN) та рекурентні нейронні мережі (RNN) зробили революцію в класифікації текстів [3]. Ці алгоритми можуть автоматично виокремлювати ознаки з тексту, усуваючи потребу в ручному створенні ознак. Вони також були більш точними, ніж попередні методи, і могли обробляти складні текстові дані. В останні роки трансферне навчання ще більше підвищило точність алгоритмів класифікації текстів. Трансферне навчання передбачає попереднє навчання моделі на великому масиві текстових даних, а потім її доопрацювання на меншому [4], специфічному для певної галузі наборі даних. Цей підхід покращує продуктивність моделей класифікації текстів, особливо коли набір даних, специфічний для конкретної галузі, невеликий.

Останньою розробкою в галузі класифікації текстів є моделі на основі трансформерів. Ці моделі, такі як BERT [5] (Bidirectional Encoder Representations from Transformers), стали найсучаснішими в класифікації текстів. Вони використовують механізми самонавчання для вивчення контекстних представлень слів, що забезпечує кращу точність і здатність справлятися з більш складними завданнями обробки природної мови. У публікації [7] порівняно різні мережі у задачах розпізнаванню настроїв у тексті і показано, що CNN RNN Bi-LSTM дає кращу точність порівняно з класичними моделями навчання. Робота [9] базується на аналізі даних із соціальної мережі twiter арабською мовою за допомогою ансамблів BERT ат комбінацій інших натренованих моделей. З недоліків цього підходу треба зазначити те, що результати отримані тренуванням великих мереж BERT на невеликих наборах даних.

Використання BERT дерев є доволі цікавою та свіжою ідеєю, яка була описана в [3]. У сфері комп'ютерної вірусології рідко можна знайти методи, які можуть ефективно шукати та класифікувати програми ворожого характеру, але цей метод надає можливість ефективно за мірками сучасності виконувати задачу класифікації вірусів. “Крім того, запропонована нами модель випадкового трансформаторного лісу (RTF) на основі пакетування, сукупність BERT або CANINE, досягає найсучасніших оцінювальних балів у трьох із чотирьох наборів даних, зокрема, вона фіксує стан -art F1-оцінка 0,6149 за одним зі стандартних наборів даних” [3]. З недоліків цієї роботи можна сказати, що цей метод є більш прикладним у сфері медіаконтенту, а не у сфері кібербезпеки. Також показники класифікації близько 60% не дозволяють говори хорошу універсальну ефективність методу.

Методи та моделі

Опис наборів даних

Було проаналізовано такі набори даних (таблиця 1).

Таблиця 1

Набори даних

назва набору даних	**ознака набору даних**
Toxic Tweets Dataset [7]	Токсичність
Suspicious Communication on Social Platforms [8]	Знущання
Fake News [9]	Брехня
HateXplain [9]	Образа

Набір даних токсичних твітів: Toxic Tweets Dataset – це колекція твітів, які були позначені як токсичні або нетоксичні.

Підозрілі повідомлення на соціальних платформах: Набір даних "Підозрілі повідомлення на соціальних платформах" – це колекція повідомлень, які були позначені як потенційно підозрілі або зловмисні на платформах соціальних мереж.

Фейкові новини: Набір даних "Фейкові новини" – це колекція новинних статей, які були позначені як справжні або фейкові.

HateXplain: Набір даних HateXplain – це колекція твітів з мовою ворожнечі, які були позначені поясненнями, чому вони вважаються ненависницькими.

Значення всіх даних формуються на основі загальнодоступної інформації тексту та категорії до якої вона належить.

Аналіз даних експерименту

На рис. 1 подано аналіз розподілу новини в залежності від кількості слів. Можна зазначити мультимодальний розподіл у даних з оригінальних текстів, натомість у даних негативних спостерігається пік, та ці публікації описують більш нормальний розподіл, хоча ним не може називатись. Також слід вказати на залежність довжини статті від того, чи є вона правдоподібною.

Кластерний аналіз текстових даний

Для видобутку сутностей та покращення розуміння даних, здійснено кластерний аналіз методом K-means. Для вибору оптимальної кількості кластерів використано метод ліктя та визначено, що найкращим вибором є розподіл з 3 кластерів. Кластерний аналіз вказує на чіткий розподіл всіх новин за ознакою довжини: короткі новини - до 150 слів, 2 середні новини - від 150 до 1000, 3 довгі новини - 1000+ слів. Із Рис 2 можна стверджувати що кластерний аналіз не зміг виокремити ознаку правдоподібності новини, тому потрібно провести глибокий аналіз , а саме аналіз нейромережами.

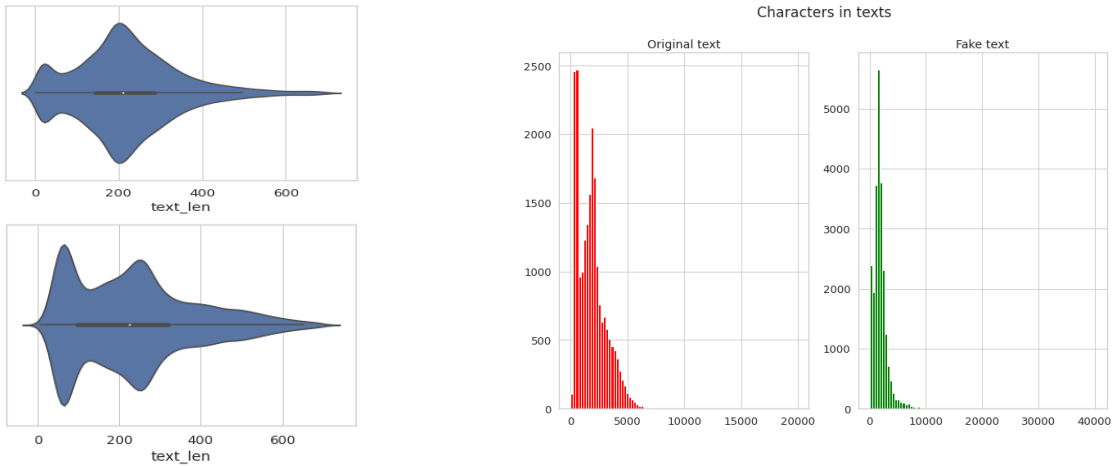

Рис. 1. Аналіз розподілу новини в залежності від кількості слів

Рис. 2. Простір текстових даних представлений градієнтом що маркує зміну довжини статті

Ідея методу та алгоритм проведеного експерименту

Основна ідея методу подана на рис. 3 і полягає у таких кроках:

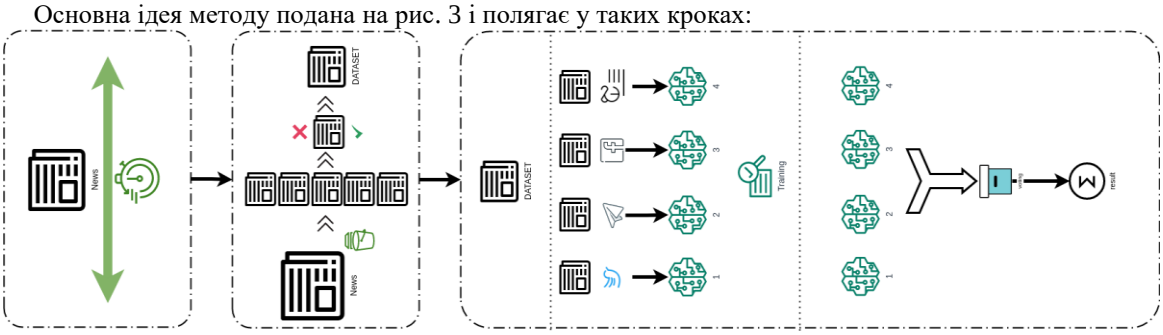

Рис. 3. Ілюстрація потоку даних від початку реалізації до кінця

1. Вибрати дані за контекстом та певними часовими рамками.
2. Підготовлені дані згрупувати та токенізувати за допомогою претренованих енкодерів BERT та RoBERTa,
3. Натренувати окремо кожну модель на своєму наборі даних
4. Скласти ансамбль моделей та провести голосування
5. Оцінити результати.

Натреновані моделі далі об’єднуються в ансамбль, в якому вони повинні проголосувати за ту чи іншу характеристику контексту. Результати експерименту показали, що створення ансамблів із моделей, які створені та натреновані на різних та незалежних даних, є ефективним. Ефективність полягає в, тому що збільшується загальний позитивний тренд оцінювання текстів та є можливість оцінити результат ансамблю в залежності від результатів голосувань. Точність алгоритму коливається від 80% до 85 % в ансамблі та від 65% до 93% окремими методами по окремих наборах даних. Тобто дослідження доцільно вважати успішним, адже можна сказати, що вектор моделей із різними контекстними характеристиками може давати хороші узагальнювані оцінки, зокрема визначити ймовірності належності тексту до тієї чи іншої підкатегорії.

На рис. 4 бачимо, що ще окрім самих натренованих моделей, у цьому алгоритмі тренується функція відсікання класифікатора. Якщо у звичайному алгоритмі голосування відбувається усереднення показників і йде відсічка по класу зі сталим коефіцієнтом 0.5, то ефективність алгоритму різко падає до ~79%. Запропонований мною спосіб відсікання збільшує загальну ефективність ансамблю на кілька відсотків.

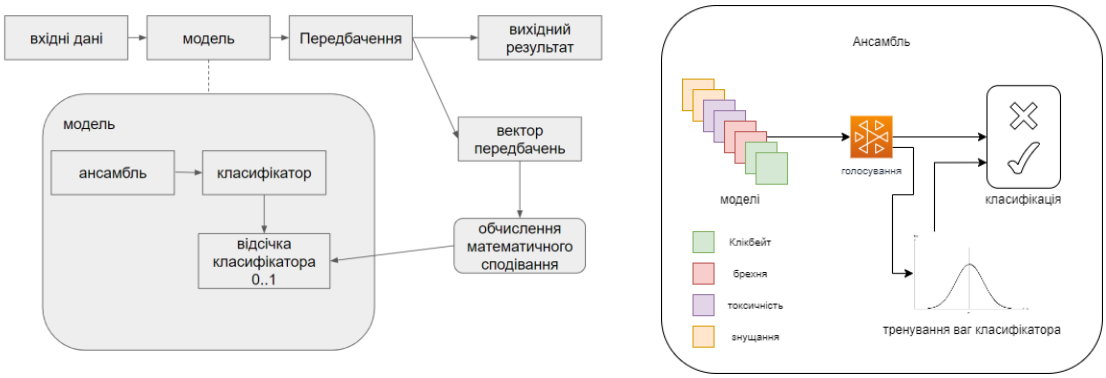

Рис. 4. Опис роботи моделі тренування коефіцієнта відсічки класифікатора

Суть алгоритму полягає в виборі коефіцієнта відсічки: на вході голосування є вектор оцінок незалежних класифікаторів, який в залежності від контексту буде голосувати по різному. Ідея методу при кожному голосуванні визначати середнє значення оцінки та додавати його до списку середніх оцінок. Список середніх оцінок являє собою набір незалежних оцінок, по якому застосовується функція математичного сподівання. Таким чином, на виході отримано коефіцієнт відсічки, який наближається до оптимально коефіцієнта поділу класів.

Результати експерименту

Кожну із натренованих моделей я протестував на вибірках відносно їхнього набору даних із таблиці. Бачимо, що моделі доволі добре розуміють контекст та по категоріях Знущання, Токсичність, Брехня показують результати в середньому 90% якості показано в таблиці 2.

Таблиця 2

Результати класифікації окремих моделей

Категорія	метод	f1-score	precision	recall
клікбейт	catboosting	0.686	0.727	0.683
	XGBosting	0.691	0.751	0.687
Знущання	catboosting	0.932	0.930	0.935
	XGBosting	0.944	0.942	0.945
Токсичність	catboosting	0.908	0.907	0.910
	XGBosting	0.911	0.910	0.913
брехня	catboosting	0.882	0.883	0.882
	XGBosting	0.880	0.881	0.880

Таблиця 3

Результат роботи ансамблю

Категорія	Ансамбль	Оптимізована відсічка	Звичайна відсічка
клікбейт	catboosting	acuracy = 0.812 recall = 0.801 precision = 0.817 f1 = 0.809	acuracy = 0.79 recall = 0.717 precision = 0.951 f1 = 0.817
	XGBosting		
знущання	catboosting		
	XGBosting		
токсичність	catboosting		
	XGBosting		
брехня	catboosting		
	XGBosting		

З таблиці 3 можна зробити висновок, що мій метод допомагає збалансувати позитивний та негативний тренд оцінок мережі цим самим покращити та зменшити шуми роботи ансамблю. Якщо проаналізувати роботу по тренуваннях та тестувань моделей, то результати тренування моделі із таблиці 2 показали хороший результат по більшості категоріям і тим самим прогнозовано, що ансамбль теж мав позитивну поведінку в оцінках тексту. Хоча й оцінка була й нижча, якщо судити тільки по окремих

спеціалізованих методах. ансамбль тестувався на об'єднані тестових даних по всіх категоріях; тому він є максимально незалежний в оцінці щодо класу.

Висновок

Мережу BERT навчено окремо на різних наборах даних та складено в один ансамбль, що дає змогу здійснити оцінку різних контекстів та визначити адекватність повідомлень від вже більш глобального контексту. Оцінка здійснена на основі голосування всіх мереж та обиранням за правилом середнього. Розроблений спосіб відсікання класифікатора, який також є частиною ансамблю, збільшує загальну ефективність ансамблю на кілька відсотків. Точність алгоритму коливається від 80% до 85 % в ансамблі та від 65% до 93% окремими методами по окремих наборах даних. Тобто дослідження, подані у таблиці 3, доцільно вважати успішним, адже можна сказати, що вектор моделей із різними контекстними характеристиками може давати хороші узагальнювані оцінки, зокрема дати ймовірності належності тексту до тієї чи іншої підкатегорії.

References

1. Dharani V., Hegde D. Mohana Spam sms (or) email detection and classification using machine learning. 2023.

2. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv, 2019.

3. Liu Y., Ott M., Goyal N. RoBERTa: a robustly optimized bert pretraining approach. arXiv, 2019.

4. Batra H., Punn N. S., Sonbhadra S. K., Agarwal S. BERT-based sentiment analysis: a software engineering perspective. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021, Vol. 12923 LNCS, P. 138–148.

5. Kang E. Long short-term memory (lstm): concept. 2017.

6. Indra G., Duraipandian N. Modeling of optimal deep learning based flood forecasting model using twitter data. Intelligent Automation and Soft Computing. 2023. Vol. 35, No. 2. P. 1455–1470.

7. Toxic tweets dataset \ textbar kaggle.

8. Suspicious communication on social platforms. 2023.

9. Fake news / hatexplain: a benchmark dataset for explainable hate speech detection. Hate-ALERT, 2023.