https://doi.org/10.31649/1997-9266-2026-186-3-17-26

УДК 004.9

В. А. Висоцька[1,2]
Л. В. Чирун[1,3]
І. О. Бичков[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ СУЇЦИДАЛЬНИХ І ДЕПРЕСИВНИХ НАМІРІВ У СОЦІАЛЬНИХ МЕРЕЖАХ НА ОСНОВІ LOGISTIC REGRESSION, MULTINOMIAL NAIVE BAYES, LINEAR SVM TA CNN

[1]Національний університет «Львівська політехніка»;
[2]Харківський національний університет внутрішніх справ;
[3]Львівський національний університет імені Івана Франка

Розглянуто актуальну проблему розробки та впровадження автоматизованих систем для моніторингу ментального здоров'я користувачів у цифровому середовищі. З огляду на стрімке зростання випадків депресивних розладів та намірів суїциду, що фіксується Всесвітньою організацією охорони здоров'я, виникає нагальна потреба у створенні інструментів раннього виявлення психологічних ризиків на основі аналізу текстового контенту соціальних мереж та месенджерів. Особлива увага в роботі приділена специфіці українського контексту в умовах соціальної нестабільності, де ресурси медичних установ є обмеженими. Метою дослідження є розробка інтелектуальної інформаційно-аналітичної платформи MindGuard, яка дозволяє здійснювати глибокий аналіз текстів для виявлення ознак депресії, тривожності та суїцидальної поведінки. Для досягнення цієї мети сформовано та розмічено масштабний навчальний корпус, що включає понад 340 000 прикладів повідомлень із платформи Reddit, розподілених за категоріями: «non-psycho» (тексти без ознак розладів), «depression» (депресивні стани) та «suicide» (суїцидальні наміри). У межах дослідження реалізовано та порівняно декілька підходів до класифікації текстів: класичні алгоритми машинного навчання (Logistic Regression, Multinomial Naive Bayes, Linear SVM) з використанням статистичного зважування TF-IDF, а також методи глибинного навчання, зокрема архітектуру TextCNN із застосуванням попередньо навчених ембеддингів GloVe. Результати експериментальної оцінки показали, що модель CNN забезпечує найкращий баланс між точністю та обчислювальною ефективністю. Досягнуті показники F1-score становлять 0,92 для класу «non-psycho», 0,74 для «depression» та 0,76 для «suicide». Високі значення метрики ROC AUC (0,910...0,990) підтверджують здатність системи ефективно ранжувати тексти за ступенем ризику. Проведено детальний аналіз матриць неточностей, який виявив лексичну близькість між категоріями депресії та суїцидальних думок, що зумовлює основні складнощі в диференціації проміжних станів. Практична реалізація платформи включає розробку REST API на базі FastAPI та клієнтської частини на React. Тестування продуктивності підтвердило високу швидкодію системи: середня затримка (latency) становить близько 15 мс на запит, що дозволяє обробляти до 66 текстів на секунду на одному CPU-потоці. Це відкриває можливості для інтеграції MindGuard у роботу модераторських команд соціальних мереж, психологічних служб та освітніх закладів для автоматизації первинного скринінгу та своєчасного прийняття превентивних заходів. Наукова новизна роботи полягає у створенні гібридної архітектури, адаптованої до локального контексту, що забезпечує високу точність виявлення психологічно ризикованих повідомлень у реальному часі.

Ключові слова: обробка природної мови (NLP), машинне навчання, глибинне навчання, ментальне здоров'я, депресія, суїцидальні наміри, інформаційно-аналітична платформа, CNN, TF-IDF, соціальні мережі, психолінгвістика, цифрова психогігієна.

Вступ

У сучасному цифровому середовищі соціальні мережі та месенджери стали основними каналами комунікації, де щодня генерується мільйони текстових повідомлень. Значна частина з них ві-

дображає емоційний стан, настрій та навіть ментальне здоров'я користувачів. У той час як Всесвітня організація охорони здоров'я фіксує стрімке зростання випадків депресії, тривожних розладів і намірів суїциду, виникає нагальна потреба у створенні інструментів раннього виявлення таких ризиків на основі цифрового контенту. Особливої гостроти ця проблема набуває в Україні в умовах соціальної нестабільності та обмежених ресурсів медичних установ. З огляду на розвиток технологій обробки природної мови (NLP) та машинного навчання, зокрема трансформерних архітектур і глибинних нейромереж, відкриваються нові можливості для автоматизованого аналізу психологічного стану за текстовими повідомленнями. Проте наявні сервіси мають низку обмежень: відсутність спеціалізації на ментальному здоров'ї, недостатня інтерпретованість результатів, відсутність мультимовної підтримки та локалізації. У зв'язку з цим, *метою цього дослідження* є розробка інтелектуальної інформаційно-аналітичної платформи MindGuard, яка дозволяє здійснювати глибокий аналіз текстів користувачів соціальних мереж з метою виявлення ознак депресії, тривожних станів або суїцидальних намірів. Система має забезпечити високу точність класифікації, швидкодію, а також зручне API для інтеграції у зовнішні сервіси. Для досягнення поставленої мети в роботі вирішуються такі завдання:

– здійснити аналіз наукових і прикладних джерел у галузі психолінгвістики, NLP та систем діагностики ментального здоров'я;

– сформувати навчальний корпус текстів з розміткою категорій: non-psycho, depression, suicide;

– реалізувати та порівняти класичні та глибинні підходи до класифікації (TF-IDF з ML, CNN, трансформери);

– розробити REST API для прогнозування класу за текстом у реальному часі;

– провести тестування точності, продуктивності та інтерпретованості системи.

Об'єктом дослідження є текстові повідомлення користувачів соціальних мереж і форумів. Предметом дослідження виступають лінгвістичні, статистичні та семантичні ознаки тексту, що корелюють з ознаками психологічних розладів (депресії, суїцидальних думок). На відміну від більшості сучасних досліджень, що фокусуються виключно на трансформерних архітектурах (BERT, RoBERTa), ця робота пропонує практично орієнтований гібридний підхід, у якому трансформери розглядаються не як єдине рішення, а як компонент багаторівневої системи аналізу. Основний акцент зроблено на досягненні балансу між точністю, швидкодією та масштабованістю, що є критичним для задач реального часу. Наукова новизна роботи полягає не у використанні окремих нових алгоритмів, а у створенні комбінаторно нової гібридної архітектури аналізу текстів, яка поєднує статистичні методи (TF-IDF), глибинне навчання (CNN) та класичні методи (LR, SVM, NB) в єдину багаторівневу систему. Вперше розроблено комплексне рішення, що дозволяє виявляти психологічно ризиковані повідомлення у багатомовному середовищі з високою швидкодією. На відміну від наявних рішень, запропонований підхід забезпечує узгоджену інтеграцію методів різної природи з урахуванням їхніх сильних сторін: швидкодії та інтерпретованості класичних моделей, здатності CNN до виявлення локальних семантичних патернів і глибокого контекстного розуміння трансформерів. Додатковою складовою новизни є адаптація цієї архітектури до багатомовного, зокрема українського, контексту, що реалізується через поєднання локальних лексичних ознак та механізмів cross-lingual переносу знань. Таким чином, новизна роботи має комбінаторний та прикладний характер і полягає у створенні ефективної, масштабованої та адаптованої до реальних умов системи аналізу психоемоційного контенту. Практична цінність проєкту полягає у можливості інтеграції платформи MindGuard у роботу психологічних служб, модераторських команд соціальних мереж, освітніх закладів та сервісів ментального добробуту. Система дозволяє автоматизувати первинний скринінг великого обсягу текстів, своєчасно виявляти загрозливі сигнали та приймати відповідні превентивні дії. Система не замінює лікаря, а слугує «сигналізацією», яка допомагає виявити потенційно небезпечні випадки серед величезного масиву цифрових даних. Зазначимо перелік сфер практичного застосування цієї системи:

1. Модерація та безпека в соціальних мережах (автоматична фільтрація, пріоритезація запитів).
2. Освітні заклади та студентське середовище (раннє попередження, цифрова психогігієна).
3. Психологічні служби та сервіси підтримки (Pre-care, попередній аналіз звернень, масштабування допомоги).
4. Корпоративний сектор (Mental Wellbeing, анонімний скринінг).

Аналіз літературних джерел

Упродовж останнього десятиліття значно зросла кількість досліджень, спрямованих на застосу-

вання методів обробки природної мови (NLP) для аналізу психічного стану користувачів у цифровому середовищі. Актуальність цієї тематики зумовлена як поширенням депресивних і тривожних розладів, так і стрімким розвитком технологій машинного навчання, здатних обробляти великі обсяги текстових даних. Одним з перших систематичних підходів до психолінгвістичного аналізу є Linguistic Inquiry and Word Count (LIWC) — інструмент, який класифікує слова за психологічними категоріями (емоції, когнітивні процеси, соціальні теми) [1]. Подальший розвиток досліджень пов'язаний з використанням соціальних медіа як джерела для виявлення психічних порушень. У [2] запропоновано фреймворк SMHD (Self-reported Mental Health Diagnoses) для вивчення депресії та ПТСР на основі постів користувачів Reddit, де застосовано як класичні, так і сучасні методи класифікації. Важливу роль у дослідженнях психологічних ризиків відіграють моделі трансформерного типу. Наприклад, в [3] продемонстрували можливість ефективного використання BERT для виявлення ризику самопошкодження на форумах підтримки. Аналогічно, в [4] розробили моделі, що вивчають дискурс психічного здоров'я без необхідності ручного маркування, відкриваючи нові горизонти у створенні інтерпретованих моделей. Методологічне підґрунтя для емоційної типології тексту можна знайти у [5], де подано circumplex model of affect — теоретичну основу для аналізу емоційних станів за двома вимірами: валентності та активації. Сучасні платформи для аналізу настроїв, такі як IBM Watson Tone Analyzer або Google Cloud NLP, забезпечують високошвидкісну класифікацію загального емоційного тону, проте не фокусуються на виявленні психічних розладів і не надають інтерпретації на рівні клінічно значущих станів. Дослідження базується на еволюції методів аналізу ментального здоров'я: від лінгвістичних словників до моделей глибинного навчання, зокрема: класичні психолінгвістичні підходи [1], спеціалізовані датасети [2], глибинне навчання [3], емоційні моделі [4] та обмеження наявних рішень [5]—[10].

Методи

У дослідженні враховано обмеження наявних рішень та адаптовано гібридні методи — комбінацію TF-IDF, CNN та BERT, що демонструють високу точність і практичну застосовність у разі класифікації текстів за ризиковими психологічними ознаками [6]—[10]. У дослідженні реалізовано гібридний підхід до аналізу текстових повідомлень користувачів соціальних мереж для виявлення психологічних ризиків (депресії, тривожності, суїцидальних намірів), що базується на поєднанні класичних статистичних методів та сучасних моделей глибинного навчання. Застосовано як традиційні алгоритми машинного навчання, так і глибинні архітектури з попередньо навченими ембеддингами. Відмова від використання трансформерів як основної моделі є свідомим компромісом між точністю та продуктивністю. У задачах реального часу критичними є латентність і масштабованість, де CNN демонструє суттєві переваги. Водночас трансформери використовуються як додатковий рівень аналізу для підвищення якості в складних випадках. Для реалізації системи використано мову програмування Python та сучасні бібліотеки обробки природної мови і машинного навчання. Класичні алгоритми класифікації реалізовано за допомогою бібліотеки scikit-learn, а глибинні моделі — з використанням TensorFlow/Keras. Для інтеграції трансформерних моделей застосовано бібліотеку HuggingFace Transformers. Серверну частину системи реалізовано на базі FastAPI, що забезпечує можливість високопродуктивного оброблення запитів у режимі реального часу. Контейнеризація застосунку виконана за допомогою Docker, що забезпечує масштабованість і зручність розгортання системи.

Результати дослідження

У дослідженні проведено експериментальне порівняння класичних методів машинного навчання (Logistic Regression, Multinomial Naive Bayes, SVM) та глибинної моделі CNN у задачі класифікації текстів за ознаками психоемоційного стану (таблиця). Вибір не використовувати трансформерні моделі (BERT, RoBERTa) як основний компонент системи є свідомим інженерним рішенням, зумовленим вимогами до продуктивності та масштабованості. Попри високу точність трансформерів у задачах обробки природної мови, їх застосування у режимі реального часу пов'язане зі значними обчислювальними витратами, підвищеною затримкою інференсу та потребою у спеціалізованому апаратному забезпеченні (GPU/TPU). У контексті задачі масового скринінгу текстових повідомлень (де обробляються десятки повідомлень за секунду) ключовими є низька латентність і стабільна продуктивність на обмежених ресурсах (CPU). Саме тому як основну модель вибрано CNN, яка забезпечує оптимальний баланс між точністю (F1-score до 0,92) та швидкодією (≈15 мс на запит), що дозволяє ефективно масштабувати систему. Водночас

трансформерні моделі інтегруються як допоміжний рівень аналізу для складніших випадків, де потрібне глибше семантичне розуміння тексту.

Порівняння якості моделей класифікації

Модель	Клас	Precision	Recall	F1-score	ROC AUC
Logistic Regression	non-psycho	~0,87	~0,92	~0,90	~0,99
	depression	~0,76	~0,75	~0,76	~0,93
	suicide	~0,77	~0,75	~0,76	~0,91
Multinomial NB	non-psycho	~0,86	~0,82	~0,84	~0,96
	depression	~0,69	~0,71	~0,70	~0,87
	suicide	~0,70	~0,72	~0,70	~0,88
SVM	non-psycho	~0,88	~0,93	~0,90	~0,95
	depression	~0,75	~0,74	~0,75	~0,91
	suicide	~0,76	~0,75	~0,75	~0,91
CNN	non-psycho	~0,92	~0,92	~0,92	~0,990
	depression	~0,73	~0,71	~0,74	~0,911
	suicide	~0,78	~0,79	~0,76	~0,910

```
Latency (ms):  min=13.0,  p50=15.0,  max=64.5
ΔRAM (bytes):   avg=6439
ΔCPU (%):       avg=100.9
```

Рис. 1. Результати вимірювання продуктивності моделі (latency, використання CPU та пам'яті)

Такий підхід реалізує багаторівневу архітектуру, у якій швидкий первинний скринінг поєднується з можливістю поглибленого аналізу, що ефективніше для практичного застосування, ніж використання виключно трансформерів.

CNN модель показала найкращий баланс між точністю та ресурсозатратністю. F1-score 0,92 для здорових текстів, 0,74 для депресії та 0,76 для суїцидальних намірів. Високі значення (0,910...0,990) підтверджують якість ранжування текстів за ризиком. Середня затримка обробки запиту становить 15 мс, що дозволяє обробляти до 66 текстів на 1 с на одному CPU-потоці. Виявлено значну лексичну близькість між класами «depression» та «suicide», що є основним викликом для диференціації станів. Проведемо тестування і статичний результат вимірювання роботи збереженої моделі з навантаженням 1000 прикладів (рис. 1):

– Latency (обробка одного запиту): мінімум 13 мс (в ідеальних умовах inference триває близько 13 мс), медіана 15 мс (половина запитів обробляється не довше ніж за 15 мс) та максимум 64,5 мс (поодинокі «затримки» (через збірку сміття, інші системні процеси) не перевищують 0,1 с). Отже, в середньому одна модель здатна обслуговувати ≈ 1000 мс / 15 мс ≈ 66 запитів на секунду на одному CPU-потоці.

– ΔRAM (приріст пам'яті). В середньому близько 6 КБ на inference, тобто модель не створює істотних додаткових алокацій в оперативній пам'яті.

– ΔCPU (% завантаження). Середнє ≈ 100 % на одному логічному ядрі означає, що модель «вантажить» CPU-потік максимально.

Для Logistic Regression із 23208 повідомлень без психічних ознак, 21410 правильно класифіковані як non-psycho (рис. 2, 3). Із 23 207 випадків depression модель правильно класифікувала 17345. З 23207 суїцидальних повідомлень 17334 віднесено до класу suicide. Модель демонструє достатньо стабільну здатність відокремлювати healthy від ризикових повідомлень. Залишається лексична плутанина між depression і suicide. Precision для non-psycho 0,87 та для depression і suicide ~0,76...0,77 (кожне 4-те передбачення може бути хибною тривогою).

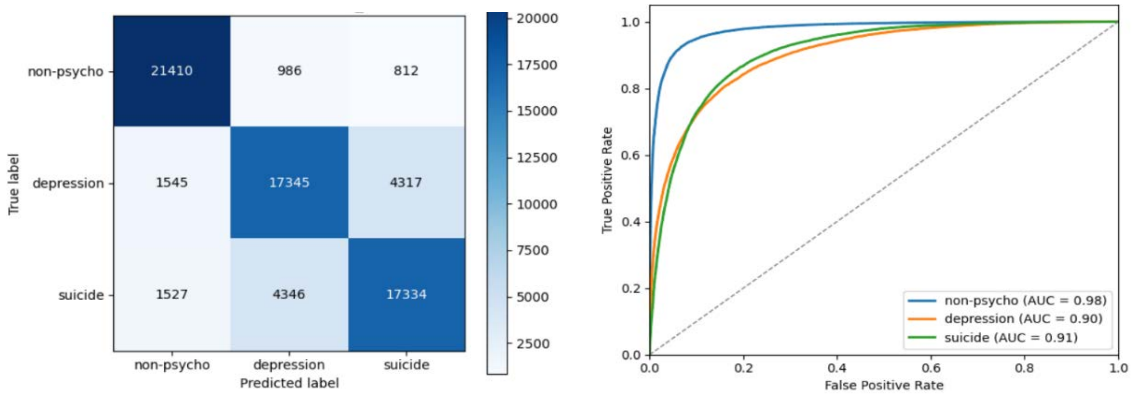
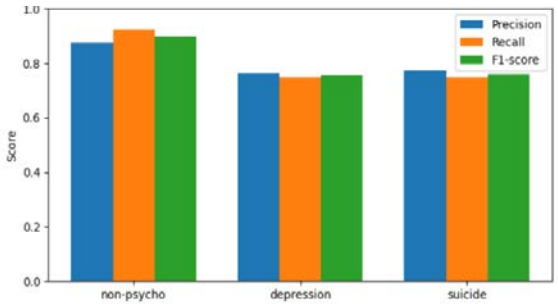


Рис. 2. Матриця помилок для моделі Logistic Regression та ROC-криві для моделі Logistic Regression

Коректні передбачення моделі MultinomialNB для non-psycho 19116 з 23208 класифіковані правильно, для depression —16494 з 23207, а для suicide —16634 з 23207 (рис. 4, 5). Найнижчі показ-

ники серед моделей. Лемматизація дала незначне покращення, але NB менш гнучкий до складної семантики. Модель схильна до загального «перестрахування», внаслідок чого має більше хибнопозитивних спрацювань. Precision для non-psycho 0,86 та для depression і suicide ~0,69...0,70. Модель Multinomial Naive Bayes демонструє значення ROC AUC на рівні 0,96 для класу non-psycho, 0,87 для depression та 0,88 для suicide, що свідчить про задовільну здатність моделі до ранжування текстів за рівнем ризику, проте поступається складнішим моделям у задачі тонкого семантичного розрізнення.



Classification report:

	precision	recall	f1-score	support
non-psycho	0.87	0.92	0.90	23208
depression	0.76	0.75	0.76	23207
suicide	0.77	0.75	0.76	23207
accuracy			0.81	69622
macro avg	0.80	0.81	0.80	69622
weighted avg	0.80	0.81	0.80	69622

Рис. 3. Метрики якості класифікації для моделі Logistic Regression (precision, recall, F1-score)

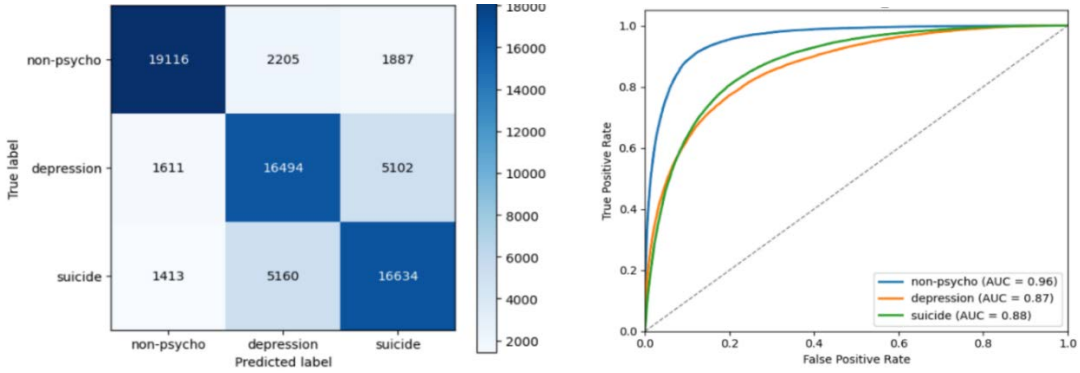
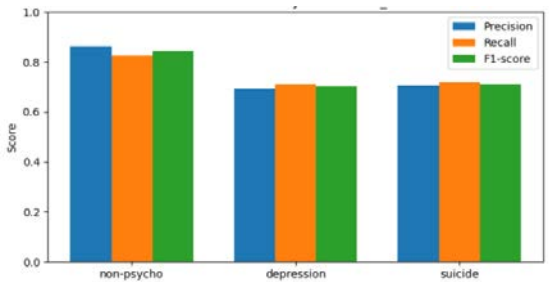


Рис. 4. Матриця помилок для моделі Multinomial Naive Bayes та ROC-криві для моделі Multinomial Naive Bayes



Classification report:

	precision	recall	f1-score	support
non-psycho	0.86	0.82	0.84	23208
depression	0.69	0.71	0.70	23207
suicide	0.70	0.72	0.71	23207
accuracy			0.75	69622
macro avg	0.75	0.75	0.75	69622
weighted avg	0.75	0.75	0.75	69622

Рис. 5. Метрики якості класифікації для моделі Multinomial Naive Bayes

Коректні передбачення моделі SVM для non-psycho 21547 з 23208, для depression 17 143 з 23207 та для suicide 17331 з 23207 (рис. 6, 7). SVM демонструє трохи кращу точність на non-psycho класі, проте знову спостерігається перехресна плутанина між depression та suicide. Алгоритм краще узагальнює, ніж NB, але гірше від LR на деяких класах.

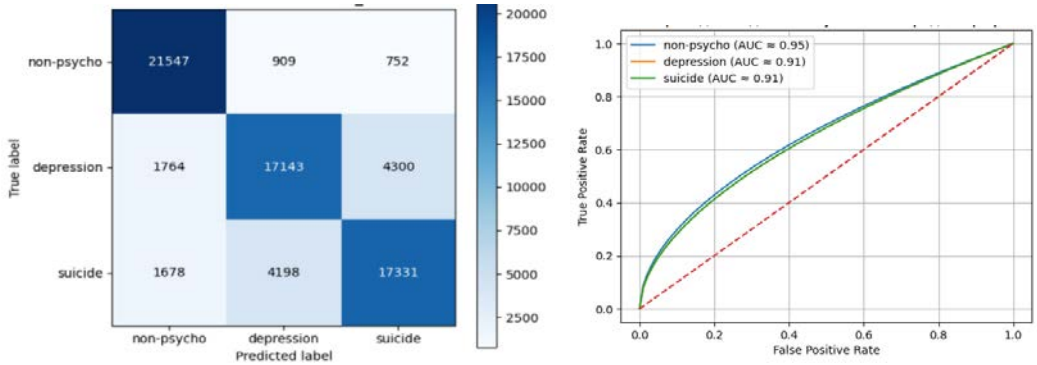
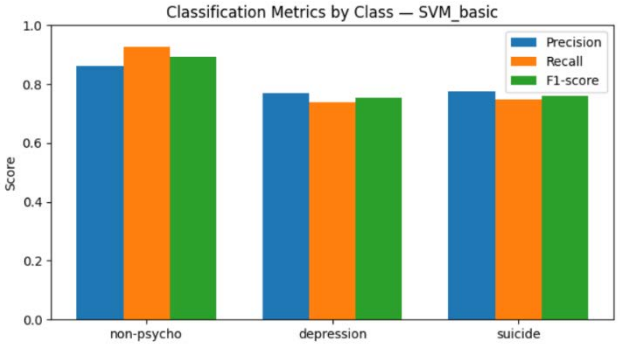


Рис. 6. Матриця помилок для моделі SVM та ROC-криві для моделі SVM



Classification report:

	precision	recall	f1-score	support
non-psycho	0.86	0.93	0.89	23208
depression	0.77	0.74	0.75	23207
suicide	0.77	0.75	0.76	23207
accuracy			0.80	69622
macro avg	0.80	0.80	0.80	69622
weighted avg	0.80	0.80	0.80	69622

Рис. 7. Метрики якості класифікації для моделі SVM

Коректні передбачення основної моделі CNN з понад 23200 повідомлень, що не містять ознак психічного ризику, 21300 віднесені до класу non-psycho (рис. 8, 9). З 23207 реальних дописів із проявами депресії, 16541 модель правильно класифікувала. Із 23207 суїцидальних повідомлень виявлено 18333.

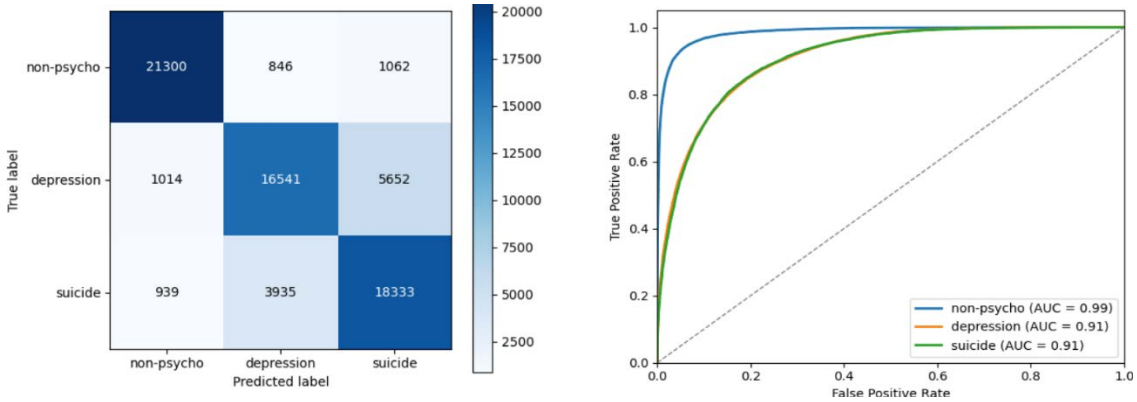
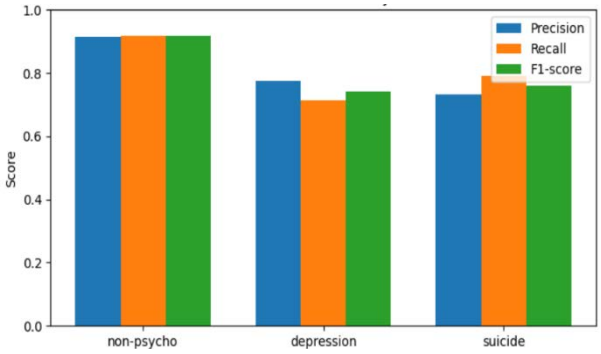


Рис. 8. Матриця помилок для моделі CNN та ROC-криві для моделі CNN



Classification report:

	precision	recall	f1-score	support
non-psycho	0.92	0.92	0.92	23208
depression	0.78	0.71	0.74	23207
suicide	0.73	0.79	0.76	23207
accuracy			0.81	69622
macro avg	0.81	0.81	0.81	69622
weighted avg	0.81	0.81	0.81	69622

Рис. 9. Метрики якості класифікації для моделі CNN

Precision («точність») для non-psycho надзвичайно високий (0,92) — майже всі передбачені як «healthy» тексти насправді не містять ознак ризику. Для обох психічно-ризикових класів precision ≈ 0,78...0,73, тобто кожна четверта передбачена моделью «депресія» чи «суїцид» може виявитися хибною тривогою. Recall («повнота») ≈ 0,92 для “non-psycho” показує, що модель легко відкидає здорові тексти. Recall ≈ 0,71 і 0,79 для «депресії» й «суїциду» означає, що трохи більше чверті таких повідомлень залишається непоміченою (false negatives). F1-score поєднує precision і recall. F1 = 0,92 для “non-psycho” демонструє сильні сторони моделі у простій задачі виявлення відсутності ризику. F1 ≈ 0.74 і 0,76 для «депресії» й «суїциду» свідчить про середній рівень.

ROC AUC для кожного класу 0,990 (non-psycho), 0,911 (depression) та 0,910 (suicide). Значення AUC близькі до 1,0 для «непсихологічних» текстів вказують на майже ідеальну здатність моделі розрізняти цей клас. AUC ≈ 0,91 для «депресії» й «суїциду» показує, що модель добре ранжує тексти за ризиком, але вимагає тонкого налаштування порогів, якщо потрібна вища повнота виявлення. Швидкість inference (~ 15 мс) дозволяє інтегрувати модель у веб-інтерфейс або API з практично миттєвими відповідями. Низький приріст пам’яті гарантує стабільну роботу під високим навантаженням без «роздування» RAM. Повне завантаження одного CPU-потоку дає передбачува-

ний, лінійний шлях до масштабування: кожен додатковий воркер або ядро дають +66 req/s. Успішне виявлення «здорових» текстів. Завдяки високому precision/recall для non-psycho, платформа може автоматично відкидати велику частину безпечних дописів без участі спеціаліста, що оптимізує ресурси операторів.

Адаптація системи MindGuard до українського та багатомовного контексту за використання англомовного корпусу Reddit є складним інженерним завданням, яке вирішується через архітектурні рішення та специфічні методи обробки даних. Хоча основна модель CNN тренувалася на Reddit, підтримка української мови базується на інтеграції моделей типу BERT та RoBERTa. Моделі Multilingual Embeddings використовують спільний векторний простір для багатьох мов. Це означає, що семантичні поняття (наприклад, біль/відчай) кодуються схожими векторами як для англійського “pain”, так і для українського «біль». Завдяки трансформерній архітектурі Cross-lingual Transfer, знання про психологічні патерни, отримані на великому англомовному корпусі, можуть бути перенесені на українські тексти без необхідності мати такий же гігантський обсяг розмічених даних українською мовою. Наукова новизна роботи полягає саме у створенні гібридної архітектури, яка адаптована до локального контексту. Система включає етапи нормалізації та токенізації, що враховують специфіку мови, зокрема видалення стоп-слів та лемматизацію, яка для української мови є складнішою через розвинену флективність (відмінювання). Використання TF-IDF дозволяє вловлювати специфічні локальні ключові слова (лексичні маркери), характерні для українського соціокультурного середовища в умовах соціальної нестабільності. Система розроблена як комплексне рішення, що дозволяє виявляти ризиковані повідомлення у багатомовному середовищі з високою швидкодією, що є критичним для інтеграції в українські освітні та психологічні служби.

Висновки

У результаті проведеного дослідження розроблено, реалізовано та протестовано прототип інтелектуальної інформаційно-аналітичної платформи MindGuard, призначеної для автоматизованого виявлення депресивних і суїцидальних патернів у текстах соціальних мереж. Робота посідає проміжне місце між класичними підходами та сучасними трансформерними моделями. На відміну від підходів, що орієнтовані виключно на BERT/RoBERTa, запропонована система реалізує гібридну архітектуру, де CNN використовується для високошвидкісного первинного скринінгу, а трансформери — як перспективний інструмент для поглибленого семантичного аналізу. Такий підхід дозволяє поєднати переваги обох парадигм і забезпечити ефективність у реальних умовах обмежених ресурсів та багатомовного середовища. Побудовано та розмічено корпус даних, що включає понад 340000 прикладів повідомлень з Reddit з категоріями: non-psycho, depression та suicide. Розроблена гнучка схема анотації із залученням експертів дозволила забезпечити якісну диференціацію текстів за психологічними станами. Реалізовано гібридний підхід до аналізу текстів, що поєднує класичні методи (TF-IDF з Logistic Regression / SVM / Naive Bayes), глибинні нейромережі (CNN з власними ембеддингами) та трансформерні архітектури (BERT, RoBERTa). У результаті експериментів встановлено, що модель CNN із GloVe-ембеддингами забезпечує найкращий баланс точності, швидкодії та інтерпретованості. Досягнуто таких показників F1-score 0,92 (non-psycho), 0,74 (depression), 0,76 (suicide); AUC ROC 0,990, 0,911, 0,910 відповідно; Latency ≈ 15 мс за навантаження до 66 запитів/секунду на одному CPU-потоці. Проведено системний аналіз архітектури платформи та реалізовано вебінтерфейс із REST API, що дозволяє швидко інтегрувати систему у практику психологічного скринінгу, модерації контенту та освітніх ініціатив. Аналіз помилок показав, що найбільші складнощі виникають у разі класифікації проміжних станів між депресією і суїцидом. Визначено напрями для покращення: збільшення кількості прикладів з «помірними» ознаками депресії; використання трансформерів з тонким тюнінгом; аугментація даних та розширення емоційного спектру. Система демонструє високу практичну цінність: вона дозволяє своєчасно виявляти психологічно ризиковані тексти, знижує навантаження на фахівців і може бути ефективно використана у психологічних службах, навчальних закладах, медіаплатформах та громадських ініціативах. Таким чином, платформа MindGuard підтвердила свою ефективність як інструмент ранньої діагностики емоційних порушень у цифровому середовищі, що відкриває перспективи для подальшого розвитку та масштабування в контексті цифрової психогігієни. MindGuard може бути інтегрована в інструментарій модераторських команд для автоматичного виявлення контенту, що порушує політику безпеки щодо самоушкодження. Завдяки високій точності у виявленні «здорових» текстів (F1 = 0,92), система здатна автоматично відсіювати безпе-

чний контент, фокусуючи увагу модераторів лише на ризикових повідомленнях. Оскільки модель добре ранжує тексти за ступенем ризику (ROC AUC ~0,91), вона дозволяє в першу чергу реагувати на критичні випадки суїцидальних намірів. Система може використовуватися для моніторингу внутрішніх форумів або анонімних чатів підтримки в університетах та школах. Це стосується виявлення ознак тривожності чи депресії у студентів на етапі їх зародження, що дозволяє психологам закладу ініціювати профілактичну бесіду до розвитку клінічного стану. Також є допоміжним інструментом в підтримці здорового емоційного клімату в онлайн-спільнотах навчальних закладів. MindGuard виступає як технічний асистент для гарячих ліній та волонтерських організацій. Система може обробляти вхідні текстові запити в реальному часі (з затримкою всього ~15 мс) і надавати фахівцю швидку оцінку стану користувача ще до початку діалогу. Продуктивність у 66 текстів на секунду дозволяє невеликим командам охоплювати значно більшу аудиторію, ніж у разі ручної обробки. Компанії можуть впроваджувати MindGuard у платформи ментального здоров’я працівників. Виявлення загальних тенденцій вигорання або депресивних настроїв у колективі без порушення приватності (на основі деперсоніфікованих даних). Платформа є засобом превентивної дії. Дозволяє «підняти прапорець» там, де людина може потребувати допомоги, але ще не звернулася до лікаря або її стан залишається непоміченим у потоці цифрової інформації. Паралельно застосування платформи MindGuard як інструменту первинного скринінгу вимагає глибокого аналізу етичних викликів, оскільки робота з ментальним здоров'ям безпосередньо впливає на права та безпеку користувачів. На основі проведеного дослідження та результатів тестування моделей, можна детально окреслити такі аспекти:

1. Ризики хибних висновків та їхні наслідки, тобто невідповідність між лінгвістичними маркерами та реальним станом людини створює два типи критичних ризиків:

– Хибнопозитивні результати (False Alarms), оскільки модель CNN демонструє рівень помилок менше 1 % для здорових текстів, проте для категорій депресії та суїциду точність (precision) становить близько 0,73...0,78. Це означає, що кожне четверте передбачення про ризик може бути помилковим. Надмірна кількість «хибних тривог» може призвести до необґрунтованого втручання в особисте життя користувача, стигматизації або перевантаження психологічних служб.

– Хибнонегативні результати (False Negatives), оскільки повнота виявлення (recall) для суїцидальних намірів становить 0,79, що означає пропуск понад 20 % потенційно небезпечних повідомлень. Найбільші складнощі виникають через лексичну близькість депресивних та суїцидальних висловлювань. Пропуск реального сигналу про допомогу є найкритичнішим ризиком, оскільки це може призвести до відсутності вчасної превентивної дії.

2. Конфіденційність та захист даних, оскільки робота з контентом соціальних мереж та месенджерів у контексті ментального здоров’я потребує жорсткого дотримання цифрової психогігієни. Система має обробляти тексти без прив’язки до персональних ідентифікаторів на етапі автоматизованого аналізу. Використання Docker для контейнеризації та FastAPI для побудови REST API забезпечує можливість створення закритих корпоративних або медичних контурів обробки даних. Користувачі мають бути поінформовані про використання алгоритмів моніторингу, особливо в освітніх закладах чи модераторських середовищах.

3. Роль експертної верифікації, оскільки MindGuard є системою підтримки прийняття рішень, а не автономним діагностом. Платформа дозволяє автоматично відкидати безпечні дописи (non-psycho), звільняючи ресурси фахівців для роботи з реальними групами ризику. Всі повідомлення, класифіковані як «suicide» або «depression», повинні проходити обов’язкову перевірку людиною (психологом або модератором) перед ухваленням будь-яких санкційних чи терапевтичних заходів. Попри те, що CNN забезпечує високу швидкість (15 мс), вона має середній рівень диференціації проміжних станів, через що людська експертиза є фінальним фільтром у складних випадках.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, *The Development and Psychometric Properties of LIWC2015*. Austin, TX, USA: University of Texas at Austin, 2015. [Electronic resource]. Available: http://hdl.handle.net/2152/31333 .

[2] A. Cohan, et al., “SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions,” *arXiv preprint arXiv:1806.05258*, 2018. https://doi.org/10.48550/arXiv.1806.05258 .

[3] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” *arXiv preprint arXiv:1709.01848*, 2017. https://doi.org/10.48550/arXiv.1709.01848 .

[4] P. Resnik, et al., “Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015. [Electronic resource]. Available: https://aclanthology.org/W15-1212.pdf .

[5] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980. https://doi.org/10.1037/h0077714 .

[6] Hugging Face, *Transformers Documentation*. [Electronic resource]. Available: https://huggingface.co/docs/transformers .

[7] TensorFlow, “TextVectorization layer,” *TensorFlow API Documentation*. [Electronic resource]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/layers/TextVectorization

[8] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA, USA: O’Reilly Media, 2023. [Electronic resource]. Available: http://14.139.161.31/OddSem-0822-1122/Hands-On_Machine_Learning_with_Scikit-Learn-Keras-and-TensorFlow-2nd-Edition-Aurelien-Geron.pdf .

[9] Y. Chen, “Convolutional neural network for sentence classification.” M.S. thesis, University of Waterloo, Waterloo, ON, Canada, 2015. [Electronic resource]. Available: http://hdl.handle.net/10012/9592 .

[10] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA, USA: O’Reilly Media, Inc., 2009. [Electronic resource]. Available: https://www.researchgate.net/publication/220691633_Natural_Language_Processing_with_Python .

Додатки

Для навчання моделей застосовано масштабну базу даних, яка відображає реальну поведінку користувачів у мережі. Використано платформу Reddit, яка є репрезентативною для збору текстів про ментальні стани (корпус із понад 340000 повідомлень). Дані розподілені за трьома семантичними класами: non-psycho» (тексти без ознак розладів), «depression» (повідомлення, що свідчать про депресивні стани) та «suicide» (висловлювання з намірами суїциду). У дослідженні для оцінки моделей використано стандартний підхід до розділення даних 20 % на 80 %. Загальний обсяг фінального тестування становив близько 69622 повідомлень. Проведено нормалізацію (нижній регістр, видалення URL та пунктуації) та токенізацію. Для забезпечення повторюваності результатів наведемо конкретні гіперпараметри для основної моделі TextCNN та класичних алгоритмів. Для класичних моделей машинного навчання використовувалась модель «мішка слів» із зважуванням TF-IDF Term Frequency–Inverse Document Frequency). Вибрано векторизатор TfidfVectorizer(max_ features=5000, ngram_range=(1, 2)), підтримка уніграм і біграм, видалення стоп-слів та нормалізація тексту перед векторизацією (нижній регістр, видалення URL, пунктуації, згадок, токенізація).У глибинних моделях застосовано ембеддинги (GloVe) та векторизація, зокрема: TextVectorization шар (довжина послідовності: 100 токенів), GloVe-ембеддинги з розмірністю 100, ембеддинги не оновлювалися під час навчання (trainable=False) та додатково застосовано SpatialDropout1D(0.2) для регуляризації. Реалізовано три основні класичні ML-моделі пайплайни з бібліотекою scikit-learn: Logistic Regression (LR_basic), Multinomial Naive Bayes (NB_advanced) та Linear SVM (SVM_basic). LR_basic — зручна для інтерпретації, забезпечила найкращу точність серед класичних методів. NB_advanced — швидкий та ефективний для роботи з TF-IDF, з обмеженнями в оцінках ймовірності. SVM_basic — точний у задачах текстової класифікації, проте чутливий до налаштувань гіперпараметрів. Реалізована архітектура TextCNN на основі глибинної моделі CNN, що включає:

– вхідний шар — Input(shape=(), dtype=tf.string) + TextVectorization;

– GloVe-ембеддинг — Embedding(input_dim=20000, output_dim=100, weights= [embedding_ matrix], trainable=False);

– згорткові шари — п’ять паралельних Conv1D із ядрами розмірів від 2 до 6;

– пулинг – GlobalMaxPooling1D та GlobalAveragePooling1D для кожного шару;

– об’єднання ознак — Concatenate() у вектор довжини 1280;

– щільні шари — Dense(64, activation='relu') з BatchNormalization та Dropout(0.5);

– вихідний шар – Dense(3, activation='softmax').

Параметрами навчання CNN є оптимізатор AdamW (lr=1e-3, weight_decay=1e-5), функція втрат sparse_categorical_crossentropy та Callbacks (EarlyStopping, ReduceLROnPlateau, ModelCheckpoint).

У межах дослідження реалізовано гібридний підхід, де кожна модель вирішує специфічні завдання аналізу. Вибір CNN+Embedding як фінальної моделі обґрунтований її здатністю забезпечувати продуктивність у 66 текстів/сек на одному CPU, що критично для первинного скринінгу. Застосування BERT/RoBERTa потенційно дозволить перемикатися між швидким скринінгом (CNN) та глибоким експертним аналізом (Transformers) у багатомовному середовищі.

Рекомендована кафедрою комп'ютерних систем управління ВНТУ

Дата надходження: 27.02.2026
Дата прийняття до друку після рецензування: 27.04.2026
Дата публікації 7.07.2026

Висоцька Вікторія Анатоліївна — д-р техн. наук, доцент, професор кафедри інформаційних систем та мереж, e-mail: victoria.a.vysotska@lpnu.ua. https://orcid.org/0000-0001-6417-3689 ;

Чирун Любомир Вікторович — канд. техн. наук, доцент кафедри інформаційних систем та мереж, e-mail: lyubomyr.v.chyrun@lpnu.ua. https://orcid.org/0000-0002-9448-1751;

Бичков Ілля Олегович — студент Інституту комп'ютерних наук та інформаційних технологій, e-mail: illia.bychkov.sa.2021@lpnu.ua. https://orcid.org/0009-0004-2170-440X .

Національний університет «Львівська політехніка», Львів.

V. A. Vysotska[1,2]
L. V. Chyrun[1,3]
I. O. Bychkov[1]

Information Technology for Suicidal and Depressive Intentions Detection in social Networks Based on Logistic Regression, Multinomial Naive Bayes, Linear SVM and CNN

[1]Lviv Polytechnic National University;
[2]Kharkiv National University of Internal Affairs;
[3]Ivan Franko National University of Lviv

The article considers the current problem of developing and implementing automated systems for monitoring users' mental health in the digital environment. Given the rapid increase in cases of depressive disorders and suicidal intentions, which is recorded by the World Health Organization, there is an urgent need to create tools for early detection of psychological risks based on the analysis of text content of social networks and messengers. Special attention is paid to the specifics of the Ukrainian context in conditions of social instability, where the resources of medical institutions are limited. The purpose of the study is to develop an intelligent information and analytical platform MindGuard, which allows for deep analysis of texts to identify signs of depression, anxiety and suicidal behavior. To achieve this goal, a large-scale training corpus was formed and labeled, including over 340,000 examples of messages from the Reddit platform, distributed by categories: "non-psycho" (texts without signs of disorders), "depression" (depressive states), and "suicide" (suicidal intentions). Within the framework of the study, several approaches to text classification were implemented and compared: classical machine learning algorithms (Logistic Regression, Multinomial Naive Bayes, Linear SVM) using TF-IDF statistical weighting, as well as deep learning methods, in particular the TextCNN architecture using pre-trained GloVe embeddings. The results of the experimental evaluation showed that the CNN model provides the best balance between accuracy and computational efficiency. The achieved F1-scores are 0.92 for the class "non-psycho", 0.74 for "depression" and 0.76 for "suicide". High values of the ROC AUC metric (0.910–0.990) confirm the system's ability to effectively rank texts by risk level. A detailed analysis of the inaccuracies matrices was conducted, which revealed lexical proximity between the categories of depression and suicidal thoughts, which causes the main difficulties in differentiating intermediate states. The practical implementation of the platform includes the development of a REST API based on FastAPI and a client part based on React. Performance testing confirmed the high speed of the system: the average latency is about 15 ms per request, which allows processing up to 66 texts per second on one CPU thread. This opens up opportunities for integrating MindGuard into the work of social media moderation teams, psychological services, and educational institutions to automate initial screening and timely adoption of preventive measures. The scientific novelty of the work lies in the creation of a hybrid architecture adapted to the local context, which ensures high accuracy in detecting psychologically risky messages in real time.

Keywords: natural language processing (NLP), machine learning, deep learning, mental health, depression, suicidal intent, information and analytical platform, CNN, TF-IDF, social networks, psycholinguistics, digital mental hygiene.

Vysotska Victoria A. — Dr Sc. (Eng.), Associate Professor, Professor of the Chair of Information Systems and Networks, e-mail: victoria.a.vysotska@lpnu.ua. https://orcid.org/0000-0001-6417-3689 ;

Chyrun Lyubomyr V. — Cand. Sc. (Eng.), Associate Professor of the Chair of Information Systems and Networks, e-mail: lyubomyr.v.chyrun@lpnu.ua. https://orcid.org/0000-0002-9448-1751;

Bychkov Ilia O. — Student of the Institute of Computer Science and Information Technology, e-mail: illia.bychkov.sa.2021@lpnu.ua