ДОДАТОК А

Слайди презентації

Рисунок А.1 – Титульний слайд

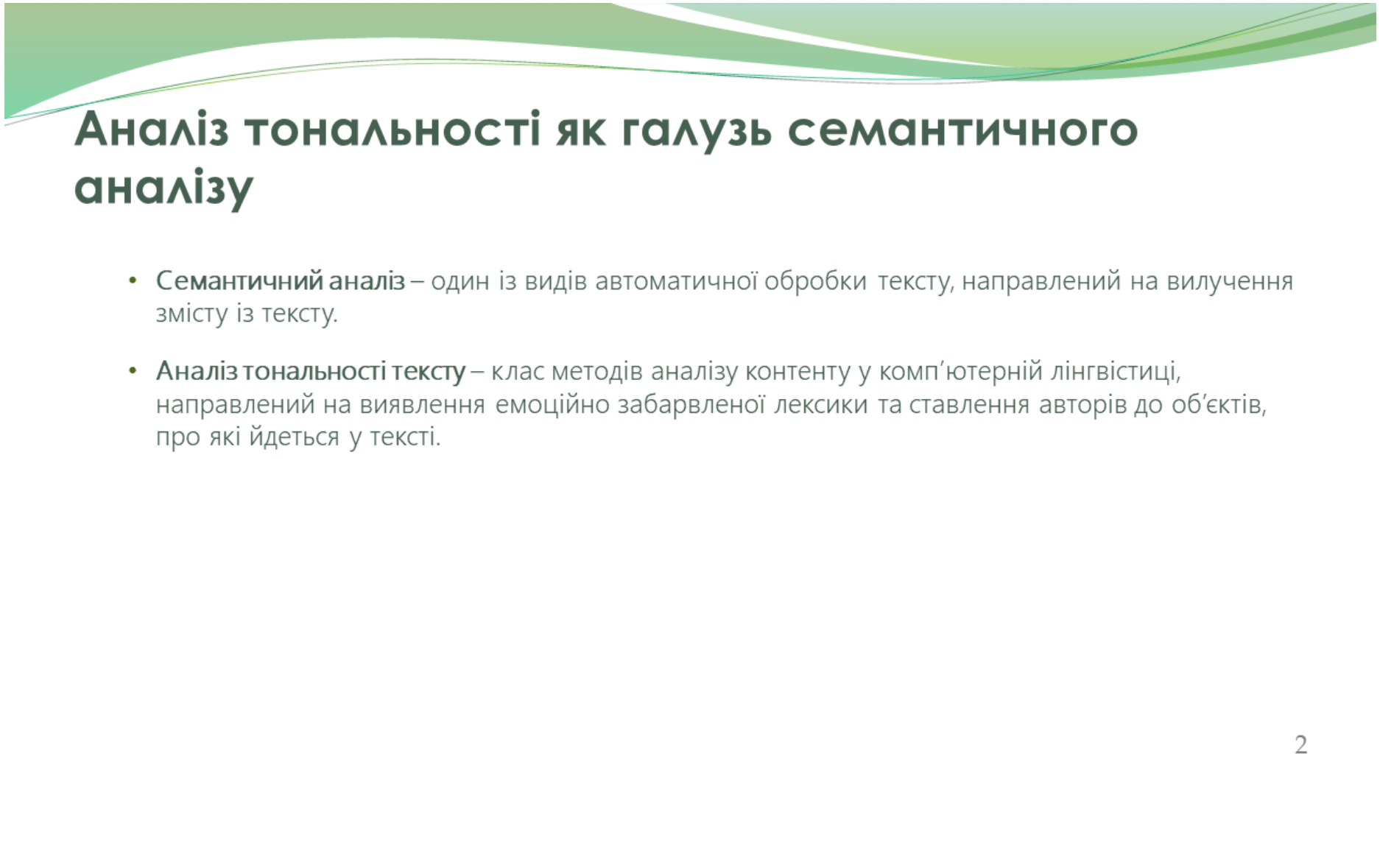


Рисунок А.2 – Слайд «Аналіз тональності як галузь семантичного аналізу»

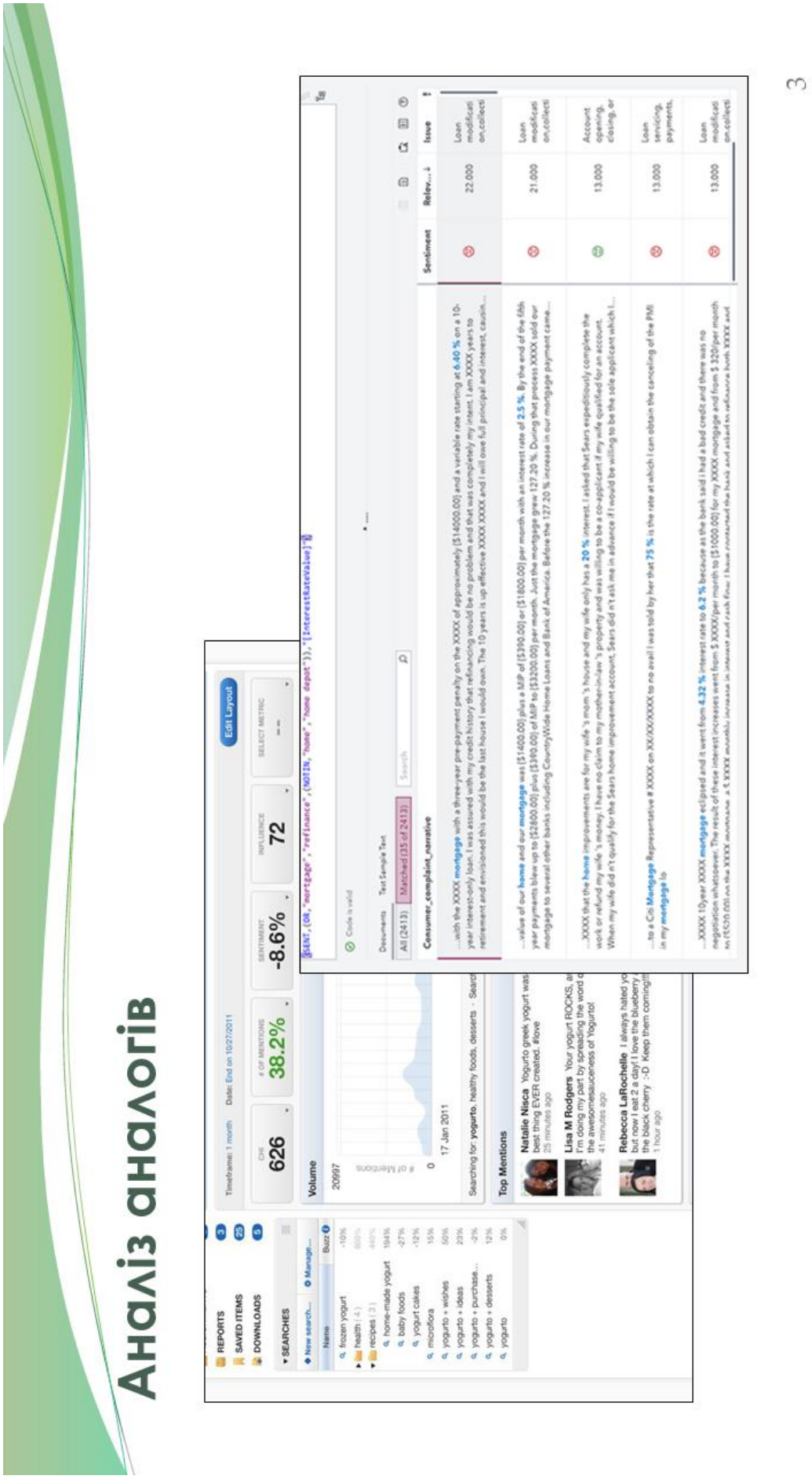


Рисунок А.3 – Слайд «Аналіз аналогів»

Постановка задачі

- Порівняти точність аналізу тональності різних лінійних класифікаторів (логістична регресія та Байєсівські класифікатори);
- Порівняти якість навчання та аналізу лінійних класифікаторів у поєднанні із «мішком слів» та нейронної мережі у поєднанні із Word2Vec;
- Проаналізувати доцільність застосування стемінгу для Word2Vec;
- Проаналізувати доцільність видалення стоп-слів;
- Визначити отпимальні підходи для різних варіантів датасетів (датасети твітів та відгуків різного розміру);
- Розробити прототип системи аналізу тональності.

4

Рисунок А.4 – Слайд «Постановка задачі»

Засоби реалізації

Мова програмування - Python 3;

Інструмент для аналітичних звітів – Jupyter Notebook;

Бібліотеки:

- **nltk** для символічної та статистичної обробки тексту;
- **Pandas** для аналізу та обробки даних;
- **Tensorflow** для роботи із нейронними мережами.

5

Рисунок А.5 – Слайд «Засоби реалізації»

Лінійні класифікатори

- Байєсівский класифікатор: $P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)}$
- Логістична регресія: $f(x) = \frac{1}{1+e^{-x}}$

Рисунок А.6 – Слайд «Лінійні класифікатори»

Нейронні мережі

$$S = \sum_{i=1}^{n} x_i w_i,$$

$$\varphi = \frac{1}{1+e^{-ax}},$$

де n – число входів нейрона;
xi – значення i-го входу нейрона;
wi – вага i-го синапсу.

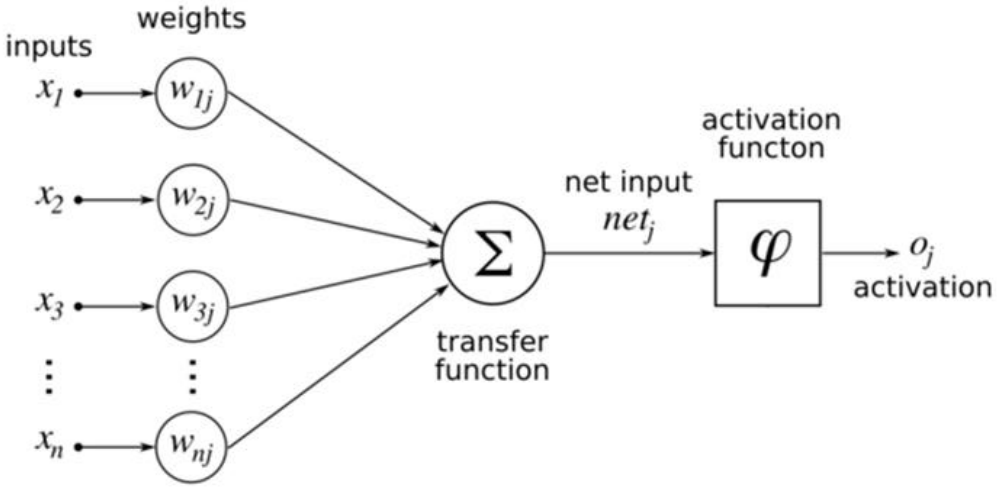


Рисунок А.7 – Слайд «Нейронні мережі»

Метрики оцінки роботи моделі

- Точність (accuracy): $Accuracy = \frac{P}{N}$, де P – кількість документів за якими класифікатор прийняв правильне рішення, а N – розмір навчальної вибірки
- Повнота: $Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}$, де TP – істино-позитивне рішення; FP – хибно-позитивного рішення; FN – хибно-негативне рішення.
- F-міра: $F = 2\frac{Precision*Recall}{Precision+Recall}$

8

Рисунок А.8 – Слайд «Метрики оцінки роботи моделі»

Проведення дослідження

Датасети:

- твіти (1.6 мільйони твітів різної тематики);
- відгуки (3 мільйони відгуків щодо товарів інтернет-магазину Amazon).

Проведені дослідження:

- різні комбінації «мішка слів» та чотирьох лінійних класифікаторів (логістична регресія, наївний Байєсівський класифікатор, Байєсівскі класифікатори із мультиноміальним розподілом та розподілом Бернулі), навчених на 10 000 та 100 000 твітах і відгуках;
- комбінації Word2Vec (навченого самостійно та Word2Vec Google) та рекурентної нейронної мережі для 10 000 та 100 000 твітів та відгуків.

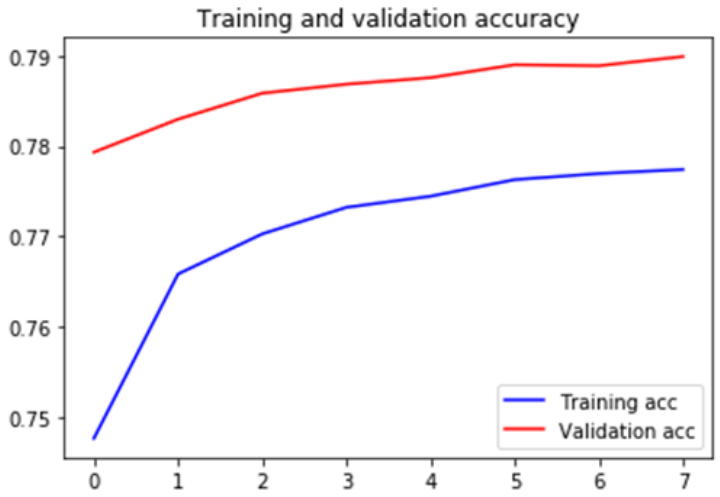
Рисунок А.9 – Слайд «Проведення дослідження»

Рисунок А.10 – Слайд «Результати досліджень – приклад класифікації текстів за тональністю (логістична регресія)»

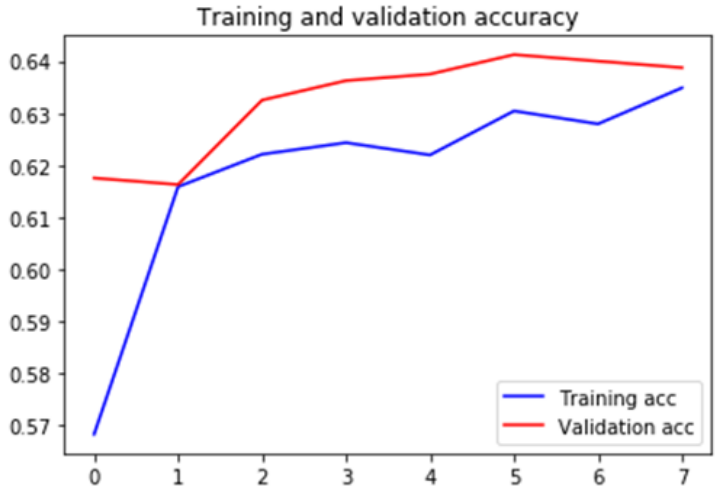
Рисунок А.11 – Слайд «Результати досліджень – приклад класифікації текстів за тональністю (Байєсівський метод)»

Рисунок А.12 – Слайд «Результати роботи нейронної мережі»

Результати роботи нейронної мережі



Навчання нейронної мережі на датасеті із 100 000 твітів



Навчання нейронної мережі на датасеті із 10 000 твітів

Результати дослідження – робота Word2Vec

Word2Vec, навчений на 100 000 твітах:

- 'luv' – 0.5732780694961548;
- 'loves' – 0.5623787045478821;
- 'loved' – 0.5373271703720093;
- 'amazing' – 0.5026600360870361;
- 'adore' – 0.4942743480205536;
- 'looove' – 0.47235167026519775;
- 'awesome' – 0.4598265290260315;
- 'lovee' – 0.45823752880096436.

Word2Vec, навчений на 10 000 твітах:

- 'hi' – 0.9325231909751892;
- 'thank' – 0.9258179664611816;
- 'thx' – 0.9088349342346191;
- 'hun' – 0.9048795104026794;
- 'kind' – 0.8882864713668823;
- 'tweets' – 0.8815054297447205;
- 'p' – 0.8811770081520081;
- 'follow' – 0.8776755332946777.

Рисунок А.13 – Слайд «Результати дослідження – робота Word2Vec»

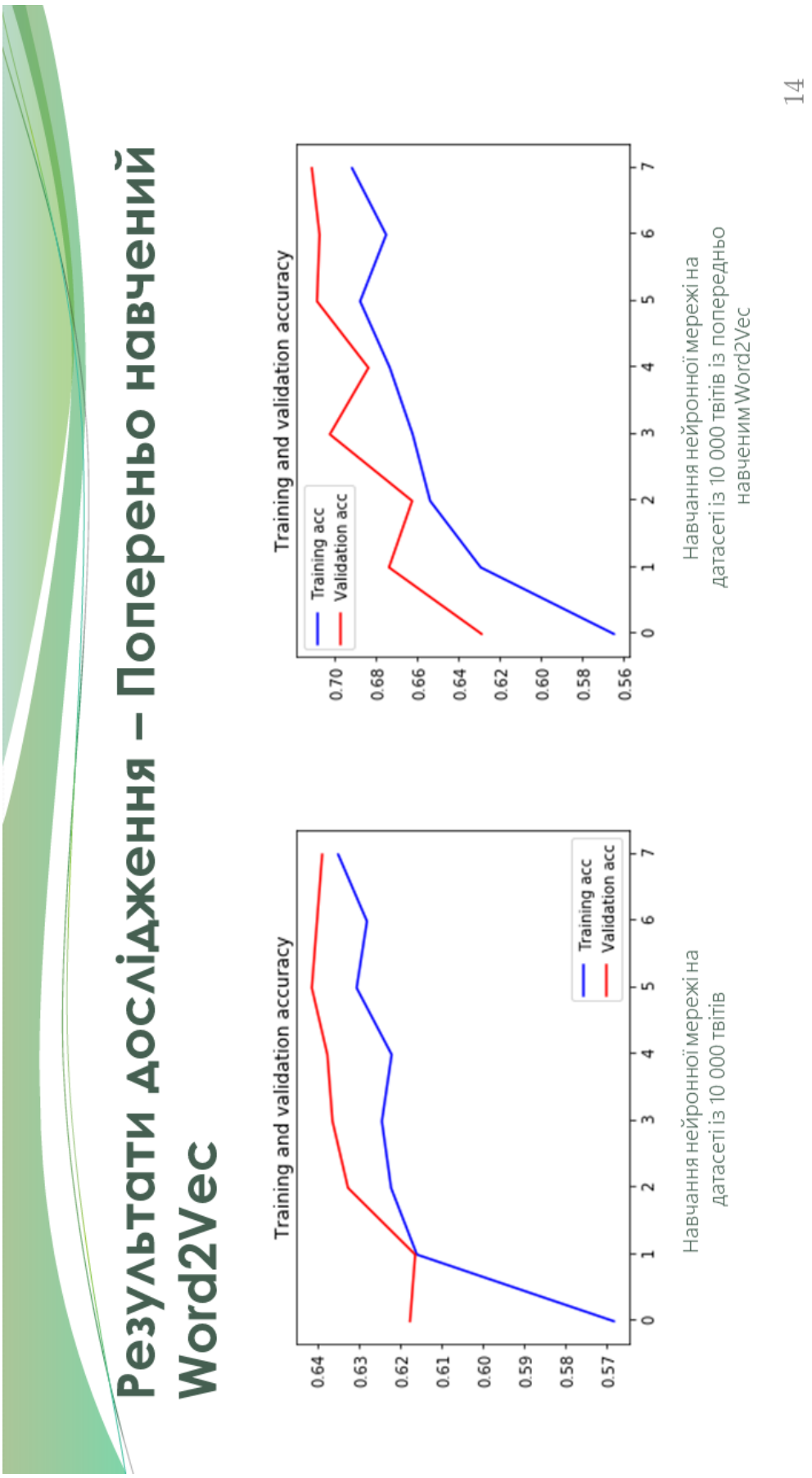


Рисунок А.14 – Слайд «Результати дослідження – попередньо навчений Word2Vec»

Результати дослідження - висновки

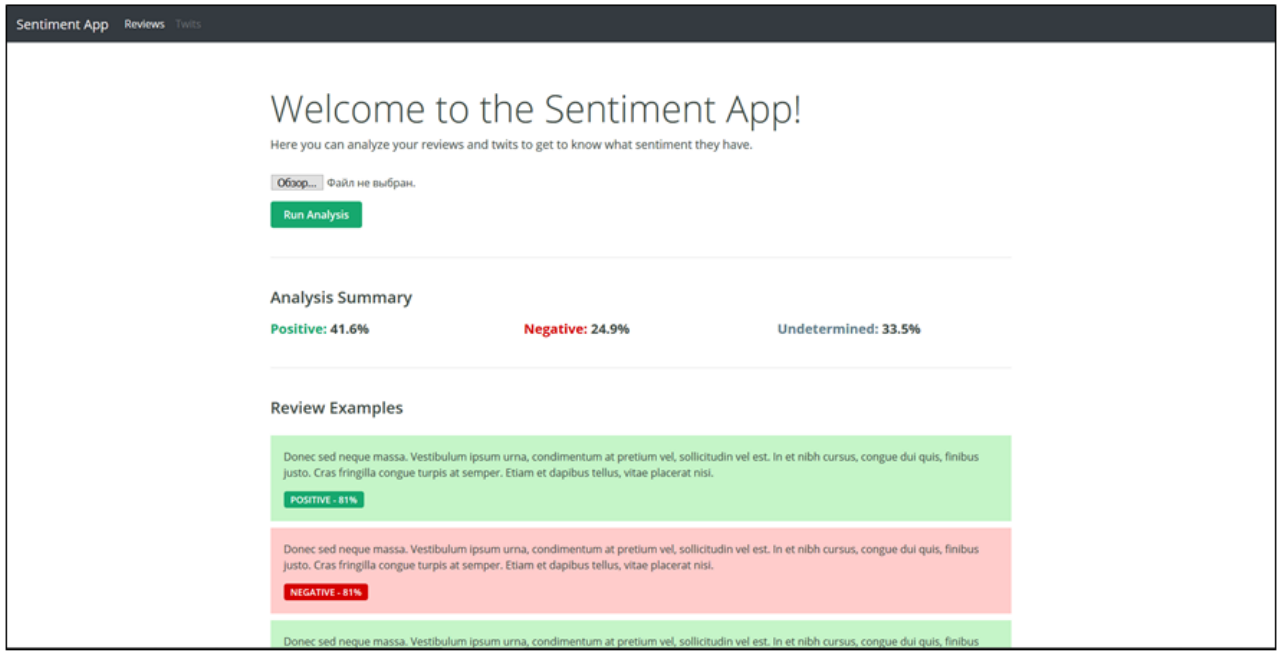
- лінійні класифікатори менше страждають від невеликої навчальної вибірки, аніж нейронні мережі. Нейронні мережі гарно працюють лише на великих датасетах;
- залишення стоп-слів сприяє незначному підвищенню точності аналізу тональності;
- при використанні нейронної мережі на невеликих датасетах має сенс використовувати попередньо навчений Word2Vec, оскільки це збільшує точність класифікації;
- нейронна мережа класифікувала відгуки із більшою точністю, ніж твіти, при однаковій кількості документів;
- застосування стемінгу у поєднанні із Word2Vec не призводить до покращення результатів класифікації, тому його застосування не виправдане.

15

Рисунок А.15 – Слайд «Результати дослідження – висновки»

Рисунок А.16 – Слайд «Прототип програмного забезпечення»

Прототип програмного забезпечення



Сторінка аналізу відгуків

Висновки

- виконано аналіз предметної галузі та методів семантичного аналізу;
- розглянуто підходи та методи реалізації автоматичного аналізу тональності;
- розглянуто методи попередньої обробки, векторизації та класифікації порівняно роботу поєднання «мішка слів» та лінійних класифікаторів із Word2Vec та нейронної мережі довгої короткочасної пам'яті на датасетах твітів та відгуків із різною кількістю документів;
- порівняно роботу чотирьох лінійних моделей: логістичної регресії, наївного Байєсівського класифікатору, класифікаторів Байєса із поліноміальним розподілом та розподілом Бернулі;
- оцінено доцільність застосування стемінгу із Word2Vec;
- порівняно роботу попередньо навченого Word2Vec та навченого на власному датасеті;
- розроблено прототип системи аналізу тональності твітів та відгуків.

17

Рисунок А.17 – Слайд «Висновки»

ДОДАТОК Б

Відгук та рецензії

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет комп'ютерних наук

ВІДГУК

на атестаційну роботу магістра
Тура Дмитра Володимировича, ПЗСм-18-1,
спеціальність *121 – Інженерія програмного забезпечення*
освітньо-професійна програма - *«ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ СИСТЕМ»*

Тема атестаційної роботи «Дослідження методів семантичного аналізу для автоматизації обробки тексту»

Проведена робота відповідає заявленій актуальній тематиці (аналізу тональності). Результати, що були отримані у результаті дослідження, можуть бути застосовані для підбору оптимального підходу для аналізу тональності в залежності від виду та кількості даних, що може покращити роботу систем автоматичного визначення емоційного забарвлення.

Під час проведення дослідження студент продемонстрував гарні вміння роботи із науково-технічною літературою та пошуку у Internet.

Тур Д. В. продемонстрував самостійність та ініціативність, а також показав вміння застосовувати сучасні методи та засоби дослідження і проектування.

Студент сумлінно та вчасно виконував усі етапи атестаційної роботи, готовий до самостійної діяльності.

Відхилень від календарного плану не було.

Магістрант гр. *ПЗСм-18-1* Тур. Д. В. готовий до самостійної інженерної діяльності. Атестаційну роботу можна подати до захисту в ЕК за спеціальністю *121- «Інженерія програмного забезпечення»,* освітньо-професійною програмою *Програмне забезпечення систем.*

«12» грудня 2019 р.

підпис

Керівник атестаційної роботи магістра
д. т. н. проф. Четвериков Г. Г.

Рецензія
на атестаційну роботу магістра
магістранта групи **ПЗСм-18-1** ***Тура Дмитра Володимировича***
спеціальність – **121- Інженерія програмного забезпечення**
освітньо-професійна програма **Програмне забезпечення систем**
«Дослідження методів семантичного аналізу для пошукових механізмів».
(Тема атестаційної роботи)

Структура атестаційної роботи: пояснювальна записка 83 сторінки; графічна частина 17 аркушів; програмне застосування 87 файлів, загальним обсягом 236 Мбайт.

Атестаційна робота, яка запропонована до рецензування, присвячена застосуванню методів семантичного аналізу для автоматичної обробки текстів. У межах роботи проведено дослідження методів, що використовуються для аналізу емоційного забарвлення текстів для датасетів різного розміру та із різними даними (твіти та відгуки), що є актуальним.

Об'єм пояснювальної записки та її розділів відповідає вимогам, матеріал викладено послідовно та структуровано. Записка відповідає стандартам. Робота містить цитування наукових публікацій та патентів, а також технічної літератури. Цитування адекватне та доцільне.

Програмне забезпечення являє собою веб-застосунок, у якому реалізований аналіз емоційного забарвлення відгуків та твітів. Програмна система виконана на високому рівні із дотриманням існуючих правил проектування та розробки програмного забезпечення, відповідає поставленому завданню.

Приведені результати дослідження можуть знайти застосування у системах та модулях аналізу тональності.

Однак атестаційна робота має недоліки: недостатній аналіз статистичної інформації щодо використаного датасету та відстуність аналізу тональності за окремими сутностями у тексті.

Атестаційна робота магістранта групи ПЗСм-18-1 Тура Д. В. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно - 90». Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121- Інженерія програмного забезпечення, освітньо-професійною програмою Програмне забезпечення систем.

Рецензент
Кандидат технічних наук, доцент, доцент
кафедри програмної інженерії В.В. Голян

Рецензія
на атестаційну роботу магістра
магістранта групи **ПЗСм-18-1** ***Тура Дмитра Володимировича***
спеціальність – **121- Інженерія програмного забезпечення**
освітньо-професійна програма **Програмне забезпечення систем**
«Дослідження методів семантичного аналізу для автоматизації обробки тексту»
(Тема атестаційної роботи)

Структура атестаційної роботи: пояснювальна записка 83 сторінки; графічна частина 17 аркушів; програмне застосування 87 файлів, загальним обсягом 236 Мбайт.

Надана для рецензування атестаційна робота відповідає заявленому поставленому завданню. Тема семантичного аналізу, зокрема аналізу тональності, є актуальною та широко застосовується для аналізу різного виду даних.

Пояснювальна записка відповідає вимогам щодо обсягу та оформлення. Матеріал викладено структуровано та логічно. У записці висвітлено аналіз предметної галузі та методів семантичного аналізу (векторизація та кластеризація текстів), етапи та результати проведеного дослідження, а також описано розроблений програмний застосунок. Цитування використаних джерел відповідає контексту. Джерела включають актуальні наукові публікації та пантенти.

Результати дослідження можуть бути застосовані для систем аналізу тональності.

Програмне забезпечення представлене у вигляді прототипу системи для аналізу тональності твітів та відгуків із застосуванням найбільш оптимального підходу, відповідно до результатів досліджень. Застосування відповідає стандартам проектування систем та розробки програмного забезпечення.

До недоліків атестаційної роботи слід віднести направленість дослідження виключно на тексти англійскою мовою.

Атестаційна робота магістранта групи ПЗСм-18-1 Тура Д. В. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно - 90». Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121- Інженерія програмного забезпечення, освітньо-професійною програмою Програмне забезпечення систем.

Рецензент
Доктор технічних наук, професор,
завідувач кафедри ІУС К.Е. Петров