Додаток А

Слайди презентації

Харківській національний університет радіоелектроніки
Атестаційна робота магістра

Дослідження методів аналізу емоційності документів про суб'єкта

Виконав: студент групи ІПЗм-18-1 І.В. Гладуш
Керівник: доцент кафедри ПІ О.П. Турута

1

Мета роботи

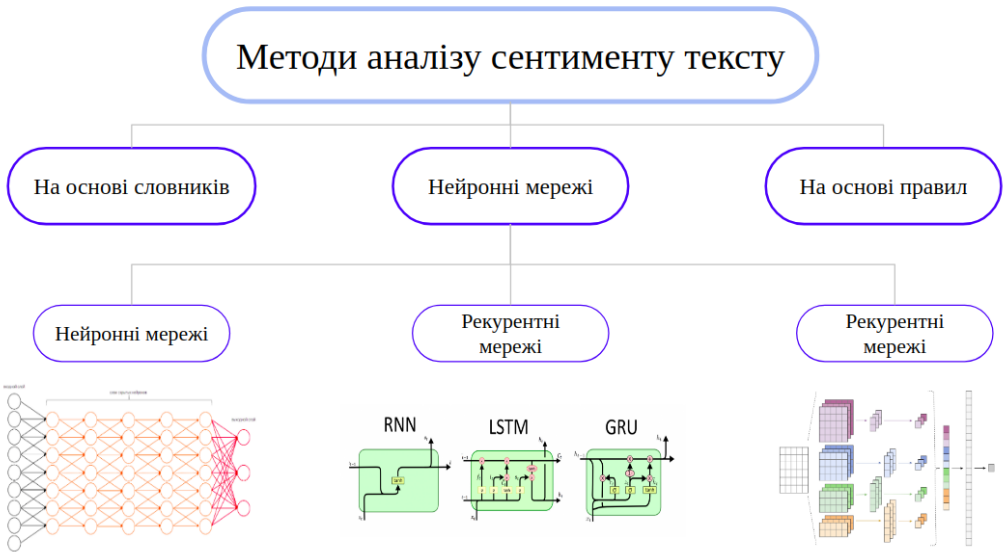
дослідити методи аналізу емоційності документів про суб'єкт написаних українською мовою.

2

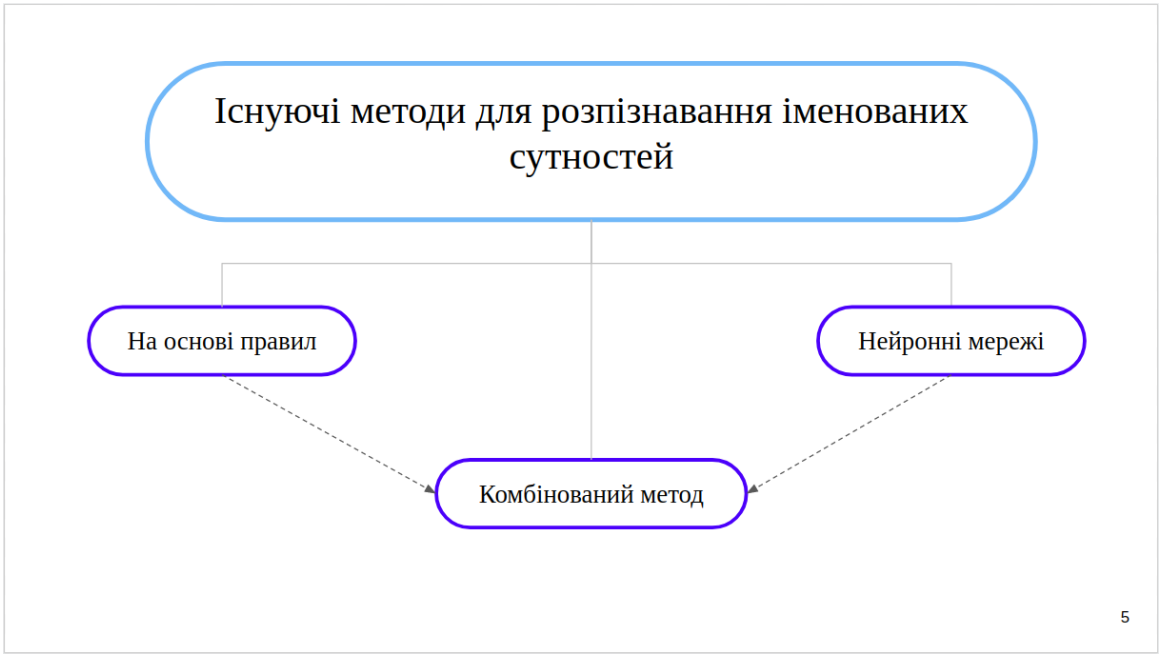
Постановка задач

- проаналізувати методи оцінки емоційності текстів та розпізнавання іменованих сутностей для української і іноземних мов;
- проаналізувати критерії оцінки точності існуючих методів;
- практично перевірити знайдені методи.

3



4



Попередня обробка тексту

- Видалення стоп слів
- Видалення знаків і символів
- Нормалізація
- Токенизація

Петрик склав задачу :), а Іринка її легко розв'язала і не помітила хо-хо-хо .

↓

Петрик склав задачу :), Іринка її легко розв'язала не помітила.

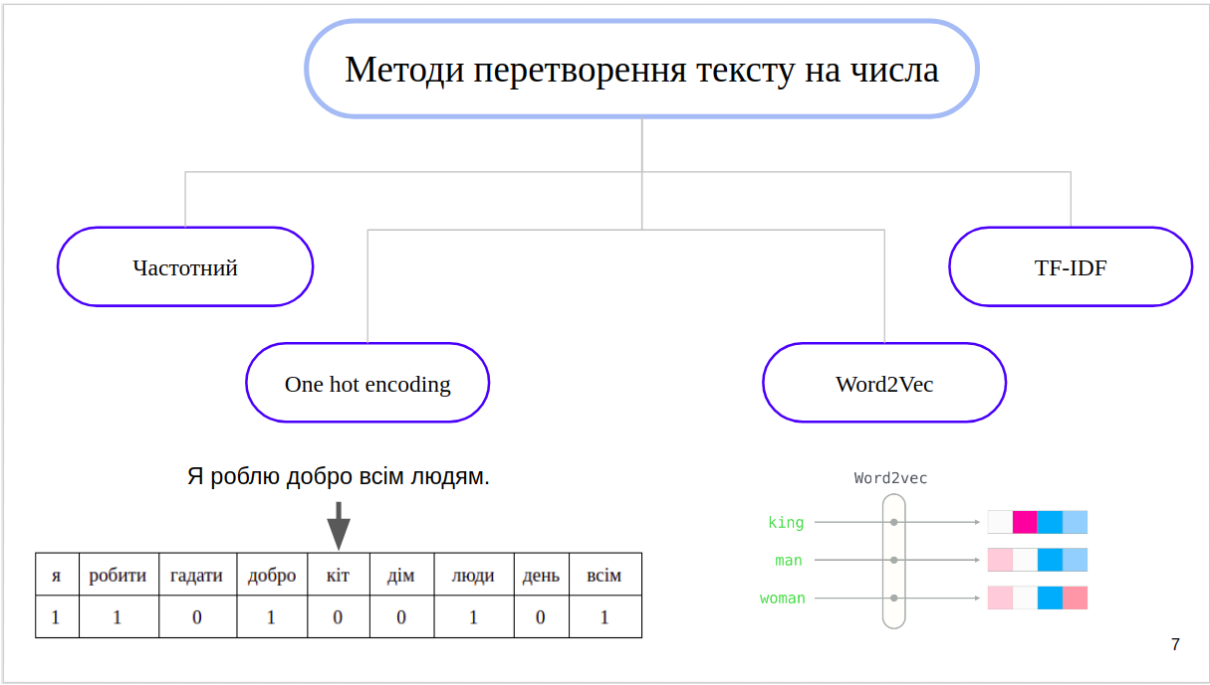
↓

Петрик склав задачу Іринка її легко розв'язала не помітила

↓

"петро", "складати", "задачу", "ірина", "її", "легко", "розв'язувати", "не", "помічати"

6



Створення датасету

	Швидкість	Точність	Ціна
Ручна	Мала	Висока	Безкоштовно
Краутфандін	Середня	Середня	Безкоштовно
Автоматичний збір	Висока	Низька	Безкоштовно
Переклад існуючого датасету	Висока	Середня	Висока

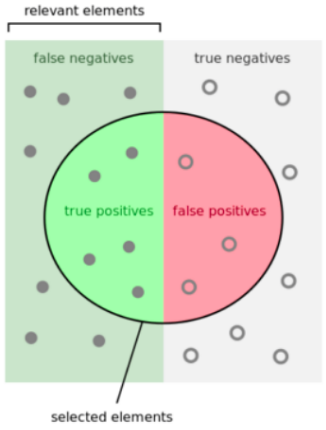
8

Критерії оцінки точності алгоритмів

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$



9

Результати оцінки емоційності тексту

	Точність на великому наборі(%)	Точність на малому наборі(%)	Швидкість навчання однієї епохи(сек)
Simple NN	50	42	3
Simple NN(One hot encoding)	80	35.5	**3**
RNN	60	39	9
LSTM-RNN	**83**	40	105
GRU-RNN	**81**	42	25

10

Результати розпізнавання іменованих сутностей на великих текстах

	Точність(F1 критерій)	Швидкість	Навантаження на процессор	Кількість оперативної пам'яті
Підхід на основі регулярних виразів	33%	**12000 слів/хв**	15%	165 МБ
Підхід на основі правил	**65%**	8000 слів/хв	**12%**	300 МБ
Підхід на основі BERT	**76%**	4300 слів/хв	30%	1000 МБ
Алгоритм з lang-uk	62%	3900 слів/хв	35%	800МБ
Існуючий алгоритм на основі регулярних виразів з гітхабу	48%	**11600 слів/хв**	**12%**	200 МБ

11

Результати розпізнавання іменованих сутностей на малих текстах

	Точність(F1 критерій)	Навантаження на процессор	Кількість оперативної пам'яті
Підхід на основі регулярних виразів	43%	15%	**155 МБ**
Підхід на основі правил	**89%**	**12%**	280 МБ
Підхід на основі BERT	**88%**	30%	950 МБ
Алгоритм з lang-uk	71%	33%	770МБ
Існуючий алгоритм на основі регулярних виразів з гітхабу	58%	**12%**	190 МБ

12

Висновки

В даній роботі зроблено аналіз існуючих рішень для проведення оцінки емоційності текстів та розпізнавання іменованих сутностей в неструктурованих текстах на українській мові.

Виконано практичну реалізацію, яка показує ефективність методів запропонованих у роботі, а також перевищує точність існуючих методів розпізнавання іменованих сутностей у текстах українською мовою на 10-15%.

13

Мої результати

1) Створений перший відкритий датасет з 28 тисяч текстів(2 мільйони слів) для аналізу сентименту текста.
2) Створене програмне забезпечення для легкої побудови правил з розпізнавання іменованих сутностей.
3) Покращені результати алгоритмів розпізнавання іменованих сутностей на українській мові.
4) По результатам роботи написана статья

14

Задачі для подальших досліджень

1. Дослідження можливостей згорткових нейронних мереж та алгоритмів на основі BERT.
2. Оптимізація алгоритмів розпізнавання іменованих сутностей.

15

ДЯКУЮ ЗА УВАГУ

16

Додаток Б

Відгук керівника

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ
Факультет комп'ютерних наук

ВІДГУК

на атестаційну роботу магістра
Гладуша Івана Валерійовича, ІПЗм-18-1
спеціальність 121 – Інженерія програмного забезпечення
освітньо-наукова програма «Інженерія програмного забезпечення»
Тема атестаційної роботи «Дослідження методів аналізу
емоційності документів про суб'єкта»

Студент Гладуш І.В. під моїм керівництвом виконував атестаційну роботу магістра, проводив дослідження існуючих методів обробки природної мови, досліджував методи машинного навчання для обробки природної мови та різні алгоритмічні підходи.

В роботі Гладуш І.В. самостійно провів аналіз проблеми дослідження, визначив основні етапи вирішення задачі, дослідив актуальні методи обробки природної мови. Дослідив методи визначення іменованих сутностей, в т.ч. для української мови. Продемонстрував високий рівень підготовленості до самостійної роботи, використовував методи наукових досліджень, показав уміння користуватися науково-технічною літературою, ресурсами мережі Інтернет, виявив глибокі знання в області алгоритмізації та мов програмування. Під час дослідження та розробки студент показав знання та вміння використовувати сучасні методи машинного навчання.

В ході роботи студент Гладуш І.В. виявився мотивованим дослідником, який приділяв увагу деталям рішення, надавав самостійну оцінку наданим пропозиціям.

Магістрант гр. ІПЗм-18-1 Гладуш І.В. готовий до самостійної інженерної діяльності. Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121-«Інженерія програмного забезпечення», освітньо-науковою програмою «Інженерія програмного забезпечення».

«_____» _____ 2020 р.
Керівник атестаційної роботи магістра
доцент каф. ПІ, к.т.н., доцент
Турута О.П.

Додаток В

Апробація результатів роботи

Автор Гладуш І.В.
Харківський національний університет радіоелектроніки ІПЗм-18-1
факультет КН кафедра ПІ
Науковий керівник – к.т.н., доцен. Турута О.П.
Адреса для листування ivan.hladush@nure.ua
Телефон 0958603217
КОНФЕРЕНЦІЯ „ІНФОРМАЦІЙНІ ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ"
Інтелектуальний аналіз даних;

Розпізнавання іменованих сутностей (NER) (також відоме як ідентифікація об'єктної сутності, фрагментація об'єктної сутності та видобуток об'єктної сутності) є підзадачею видобування інформації, вирішення якої дозволить знайти і класифікувати в неструктурованому тексті іменовані сутності, заздалегідь визначені категорії, такі як імена людей, організації, місця, медичні коди, час, кількості, грошові значення, відсотки тощо. На сьогоднішній день людство навчилося зберігати безліч інформації в електронному вигляді, але ця інформація не є зручною для пошуку в ній, тому вирішення даної проблеми є дуже важливим завданням в сучасному цифровому світі. NER дозволяє збагатити «сирі» дані корисною інформацією. Наприклад, проіндексувати всі судові рішення і для кожного рішення знайти хто був суддя, яке рішення було винесено, було виправдано людину чи ні, які статті використовувалися у ході розгляду справи, імена свідків. Аналогічну дію можна зробити і для нотаріальних рішень. Окремою частиною використання варіантів розв'язання проблеми є чат боти, які автоматизують процес вибору замовлення в онлайн магазинах або ж отримання інформації. Рішення NER легко дозволить зрозуміти чат боту де знаходиться споживач, що він шукає, який розмір одягу потрібен.

В ході даної роботи було вирішено задачу пошуку ПІБ людей в неструктурованих текстах, розглянуто два методи, які відрізняються один від одного швидкістю взаємодії і точністю.

В ході вирішення проблеми знаходження іменованих сутностей були розглянуті два основних підходи, які застосовуються в англійській мові. Останнім

часом набрав популярність алгоритм на підставі попередньо навченої нейронної мережі від Google Bert і розмітка тексту на попередній токенезації тексту та фільтрації його. У нейронній мережі Bert є два великих недоліки. Перший недолік це те, що для її навчання потрібен величезний датасета (Google використовував 100 тис текстів з вікіпедії), другий недолік - ресурси, на яких потрібно навчати цю нейронну мережу. Ці недоліки роблять її застосування для специфічних завдань складним. Рішення на підставі правил, такі як регулярні вирази були відкинуті відразу, так як вони ефективно працюють дуже у вузькому спектрі завдань, наприклад, пошук імейлів, імен законів в структурованих текстах.

Для тестування обох алгоритмів був обраний відкритий, розмічений датасета українських текстів. До цього датасету входить 230 текстів, довжина кожного тексту від 500 до 1500 слів. У кожному датасета від двох до 10 іменованих сутностей, які записані в різних формах, наприклад П. Григорович, Катрусенька, Борис Онуфрійович і т.д. Датасета містив як одвічно українські імена так і іноземні. При підрахунку точності роботи алгоритмів було використано три класи відповідей алгоритму, а саме:

1 Знайдене слово було коректним ім'ям в тексті

2 Знайдене слово не було ім'ям в тексті

3 Не було знайдено ім'я в тексті.

За допомогою вищезазначених класів помилок, були знайдені дві величини precision і recall. Precision можна інтерпретувати як частку об'єктів, названих класифікатором позитивними і при цьому вони дійсно є позитивними, а recall показує, яку частку об'єктів позитивного класу з усіх об'єктів позитивного класу знайшов алгоритм. Враховуючи те, що зазначені метрики недостатньо визначають точність алгоритму незалежно один від одного, для отримання агрегованого критерію був обраний F1 критерій.

В ході роботи алгоритму, заснованому на нейронній мережі з Google, був створений власний алгоритм, який розбиває рядок на слова, кожне слово нормалізується і до нього додається мета інформація (чи є дане слово іменником або прикметником), додається рід і число, а також визначається є це слово ім'ям,

по батькові або прізвищем. Після цього кожне слово отримує свою мітку в нотації IOB. В кінці отримані слова об'єднуються в групи і перевіряються на наявність в словнику імен. Цей алгоритм має такі переваги як швидкість роботи, можливість легко адаптується під конкретну задачу, а також високу точність в структурованих текстах. Недоліки цього алгоритму в тому, що якщо імені немає в словнику, воно не потрапить до результату, а також не високу точність на довільних текстах.

В результаті тестування двох алгоритмів на тестових даних, алгоритм на підставі Bert показав точність 78%, алгоритм розроблений в ході роботи 56%.

В ході роботи описано методи, які можуть використовуватися для вирішення проблеми пошуку іменованих сутностей, визначено критерії, необхідні для оцінки точності роботи алгоритмів, а також створені і протестовані два рішення для вирішення цієї проблеми.

Надалі заплановано проведення дослідження з метою поліпшити алгоритм за допомогою попередньої обробки тексту, а також розширити словник прізвищ, імен та по батькові. Так само планується покращити точність алгоритму заснованого на BERT шляхом навчання нейронної мережі на більшому розмірі даних, які є менш структурованими та більш різноманітними.