НТУУ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. ІГОРЯ СІКОРСЬКОГО»
«ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»
КАФЕДРА СИСТЕМНОГО ПРОЕКТУВАННЯ

Особливості роботи окремих методів автоматизованої обробки текстів для різних мов

•••

Виконав: Мошняга Н.В. ДА-61
Науковий керівник: Булах Б. В.

Мета

Дослідити особливості обробки та аналізу обраними методами текстів українською та англійською мовами (переважно за допомогою інструментів мови Python). Визначити чи покращує, чи погіршує переклад текстів на протилежну мову результати роботи обраних алгоритмів.

Актуальність проблеми

У зв’язку з наявністю багатьох природних мов існує проблема розробки словників, бібліотек, інструментів і іншого для обробки цих мови, оскільки чогось універсального, принаймні поки, не існує. Зокрема стоїть завдання порівняння і аналіз особливостей тих чи інших засобів для застосування до різних мов.

Деякі компанії, які займаються аналізом текстів на українському ринку (наприклад, аналіз емоцій у постах у соціальних мережах для рекламних потреб) перекладають тексти з української на англійську і далі проводять аналіз. Але чи не втрачається властивості тексту при перекладі, і чи є інструменти для різних мов, які працюють приблизно на одному рівні ефективності, і чи затрати на ресурси виправдовують той, чи інший вид підходу до аналізу?

Методологія

Дані:

- українські новини (6 категорій по 850 новин)
- англійські новини (5 категорій по 380 новин)

Методи аналізу (класифікації):

- наївний класифікатор Байєса
- SVM

Використаний перекладач: Google Translate

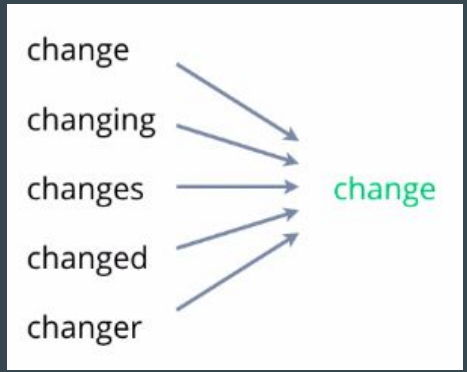
Показники для аналізу ефективності моделей:
точність (accuracy) та F1-показник

Приклад:
Класифікація новин регіональних ЗМІ

Методологія

Використані методи попередня обробки текстів:

- приведення всіх слів у нижній регістр (базова попередня обробка)
- виправлення помилок у словах (спелінг)
- зведення слів до основної форми (лематизація)



Лематизація

Методологія

Послідовність дій:

1. Застосовуємо методи класифікації до базового масиву українських новин.
2. Застосовуємо методи до масиву з виправленими помилками у словах.
3. Застосовуємо методи до масиву з виправленими помилками у словах + лематизацією.
4. Перекладаємо масив на протилежну мову і проводимо пункти 1-3
5. Проводимо пункти 1-5 для масиву англійських новин.
6. Виконуємо аналіз.

Особливості попередньої обробки текстів українською мовою

Виправлення помилок: знайдено лише один словник для виправлення помилок. Він використовується до допомогою бібліотеки hunspell

```
[82]: h = Hunspell('uk_UA', 'uk_UA')

      h.suggest('бугалтер')[0]

[82]: 'бухгалтер'
```

Лематизація слів: знайдено лише 2 інструменти: бібліотека pymorphy та утиліта для мови groovy. Для преобробки обрано pymorphy

```
[103]: import pymorphy2

       morph = pymorphy2.MorphAnalyzer(lang='uk')

       morph.parse('неї')[0].normal_form

[103]: 'вона'
```

Особливості попередньої обробки текстів англійською мовою

Виправлення помилок: використано стандартний словник з бібліотеки hunspell (тої, яка використовувалася для української мови)

```
[12]: h = Hunspell()

      print(h.suggest('Languagy')[0])

      Language
```

Лематизація слів: існує багато відомих бібліотек, зокрема nltk та spacy. Для попередньої обробки обрано nltk.

```
[15]: from nltk.stem import WordNetLemmatizer

      lemmatizer = WordNetLemmatizer()
      lemmatizer.lemmatize('changes')

[15]: 'change'
```

Переклад текстів

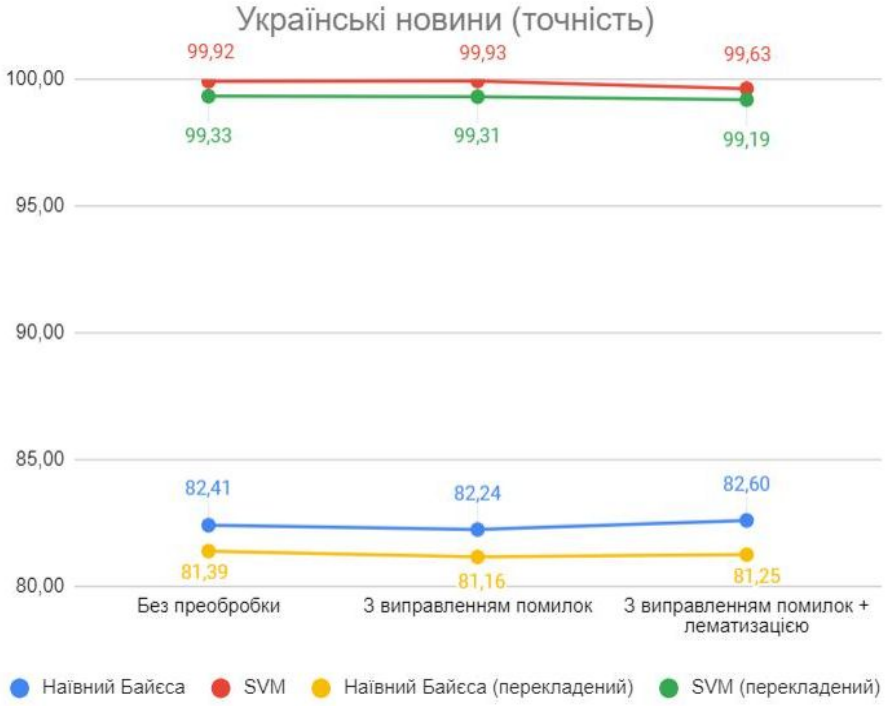
Переклад всього тексту за раз (некоректний):

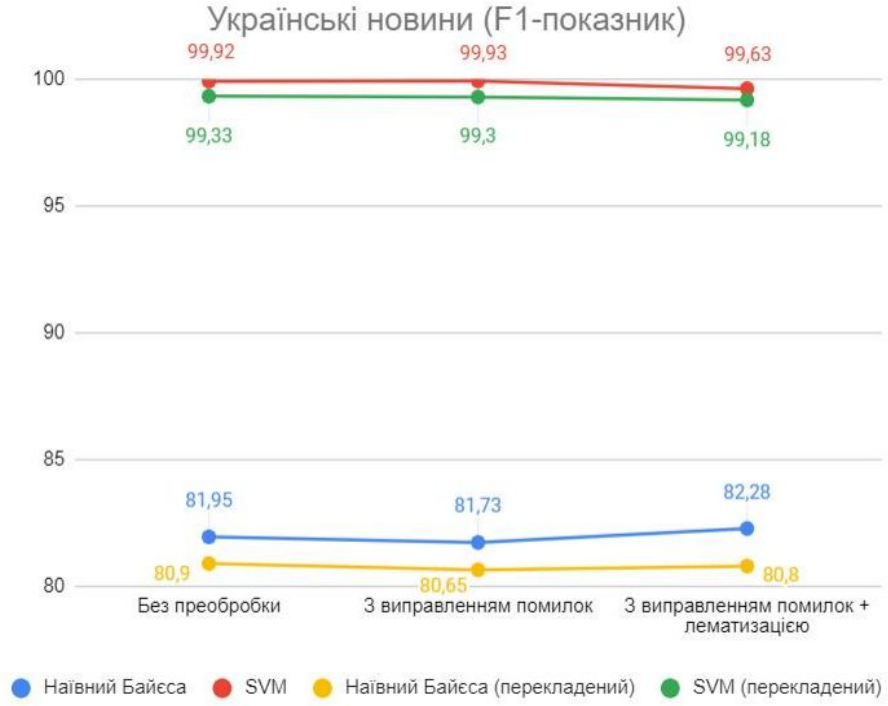
У відповідь представник парламентського комітету з питань бюджету Юрій Кузбит заявив, що це не відповідає дійсності, і комітет підтримує законопроект в остаточній редакції, викладеній у порівняльній таблиці. «Мова йде про те, щоб збалансувати бюджет-2020. Необхідно, щоб він працював і був коректним. От і все. Ніхто не намагається зараз угробити децентралізацію», - підкреслив депутат. Як повідомляв УНІАН, Комітет Верховної Ради з питань бюджету рекомендував парламенту прийняти держбюджет України на 2020 рік у другому читанні і в цілому, приступивши до його розгляду після прийняття змін до Бюджетного кодексу України (законопроект №2144).	In response, Yuri Kuzbyt, a representative of the parliamentary budget committee, said that this was not true and that the committee supported the bill in its final version, as set out in the comparative table. "It's about balancing the 2020 budget. It is necessary that it worked and was correct. That's all. Nobody is trying to ruin decentralization now, "the deputy stressed. As UNIAN reported, the Verkhovna Rada Committee on Budget recommended that the parliament adopt the state budget of Ukraine for 2020 in the second reading and in general, starting its consideration after the adoption of amendments to the Budget Code of Ukraine (bill законо2144).

Переклад лише фрази (коректний):

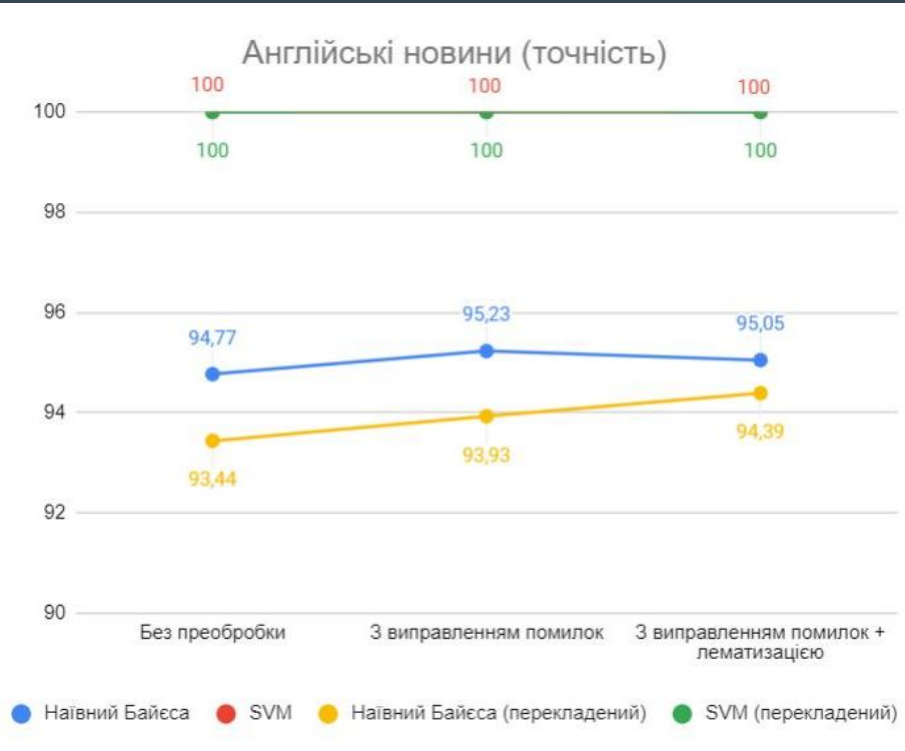
після прийняття змін до Бюджетного кодексу України (законопроект №2144).	after the adoption of amendments to the Budget Code of Ukraine (draft law №2144).

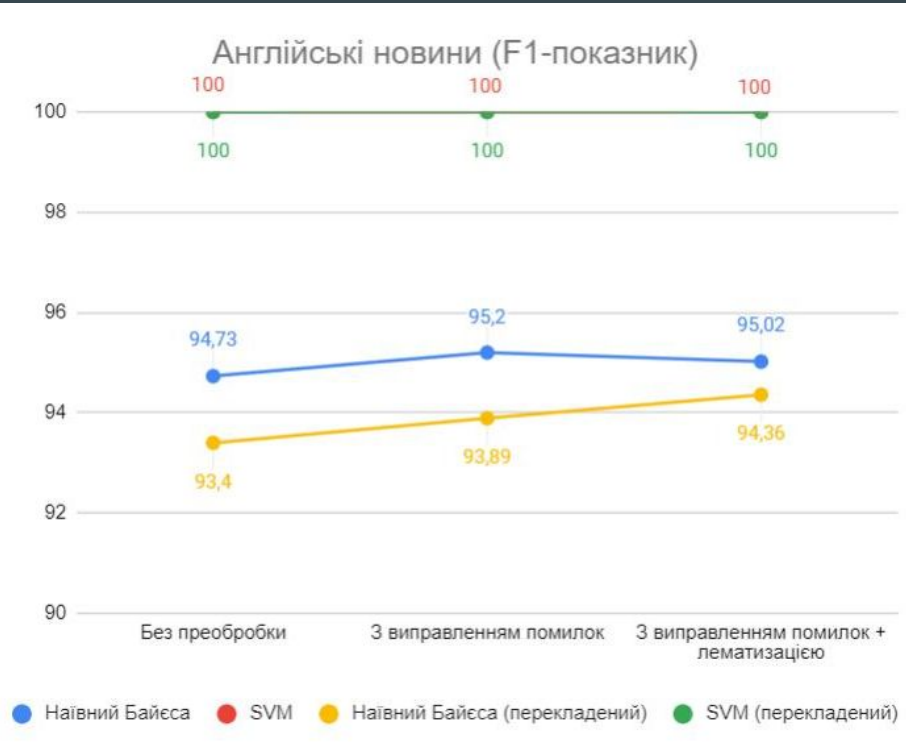
Експерименти з масивом українських новин





Експерименти з масивом англійських новин





Висновки

У результаті порівняння встановлено, що при перекладі як з української на англійську, так і в протилежний бік для розглянутих масивів і методів точність і F1-показник падають. Це може бути пов’язаним з втратами під час перекладу. Тобто переклад з української на англійську для подальшого використання інструментів під неї може виявитися недоцільним, як в плані точності, так і затрат часу на переклад. При тому, що інструменти для обробки української мови існують

Розглянуті методи попередньої обробки для української мови не принесли значного покращення, а подекуди — навіть незначно погіршили результати, ріст ефективності на один відсотковий пункт був лише для масиву, який перекладений з англійської. Для англійської мови також не принесли значного покращення.

Наступні дослідження

Наступними дослідженнями, пов’язаних з поточною темою, можуть бути порівняння іншим методів аналізу тексту (наприклад, аналіз емоцій чи NER) чи іншого виду застосування перекладу — переклад натренованої моделі.

Дякую за увагу!