Харківський національний економічний університет
імені Семена Кузнеця

Кафедра інформатики та комп’ютерної техніки

Дипломний проект
на тему: «Розроблення модулів інформаційної системи класифікації текстів»

Виконала:
студентка ПОТАПОВА Катерина

Керівник:
к.т.н. ПЕРЕДРІЙ Олена

Актуальність

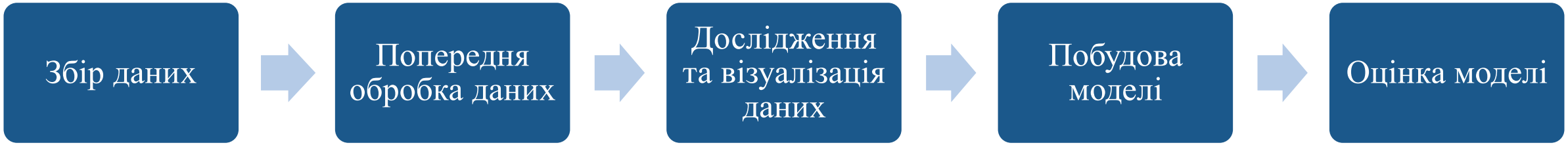
- Сучасний світ потопає у великих обсягах інформації, об’єм якої стрімко зростає.
- По оцінкам, 80% інформації є неструктурованою, при чому текст є одним з найбільш розповсюджених типів неструктурованих даних.
- На сьогоднішній день людині все складніше стає аналізувати та класифікувати дані за категоріями.
- Зростання інформації і одночасне збільшення доступних потужних комп’ютерів дозволяє застосовувати сучасні методи для вирішення задачі класифікації.

Постановка задачі

Для досягнення поставлених цілей треба виконати наступні завдання:

- підготувати набор даних для тестування модулю;
- дослідити та вибрати моделі нейронної мережі для автоматичної класифікації тексту;
- побудувати нейронну мережу та навчити її з використанням отриманих даних;
- провести тестування розробленого модулю;
- оцінити проведену класифікацію.

Етапи класифікації текстових даних



Програмне забезпечення

- Google Colab
- Python
- Keras
- TensorFlow
- NumPy

Бінарна класифікація

Точність навчальної вибірки склала майже 95%, а точність на тестовій вибірці є дещо гіршою, ніж на навчальній, та складає 87,78%.

```
[99] loss,accuracy=model.evaluate(X_train,y_train,verbose=1)
     print('Training accuracy is {}'.format(accuracy*100))

     1250/1250 [==============================] - 12s 9ms/step - loss: 0.2113 - acc: 0.9495
     Training accuracy is 94.94749903678894
```

```
[100] loss,accuracy=model.evaluate(X_test,y_test)
      print('Testing accuracy is {}'.format(accuracy*100))


      313/313 [==============================] - 3s 9ms/step - loss: 0.5075 - acc: 0.8778
      Testing accuracy is 87.77999877929688
```

Бінарна класифікація
Приклад роботи

```
[108] sample_text=('Saw the movie today and thought it was a good effort, good messages for kids.')
      oh=one_hot(sample_text,vocab_size,filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',lower=True,split=' ')
      print(oh)
      pad=pad_sequences([oh],maxlen=max_length,padding='pre')
      predictions=model.predict(pad)
      print(predictions)

      [10212, 2666, 5192, 2336, 1853, 8909, 9409, 6642, 9533, 6944, 6965, 6944, 12349, 9324, 4814]
      [[0.55447656]]
```

Перший відгук є позитивним, результат його класифікації дорівнює 0,554, що ближче до 1 – класу позитивних відгуків, ніж до 0 – класу негативних.

Другий відгук є негативним, відповідно, результат його класифікації дорівнює 0,001.

```
[109] sample_text=('The acting was bad, the dialogs were extremely shallow and insincere.')
      oh=one_hot(sample_text,vocab_size,filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',lower=True,split=' ')
      print(oh)
      pad=pad_sequences([oh],maxlen=max_length,padding='pre')
      predictions=model.predict(pad)
      print(predictions)

      [2666, 12255, 6642, 10707, 2666, 5818, 5128, 3013, 9464, 1853, 7273]
      [[0.00150684]]
```

Парсинг новин з сайту ТСН

Результат роботи модуля парсинг продемонстровані на рисунку. З сайту витягли такі блоки інформації, як заголовок та посилання на першоджерело.

```
data
```

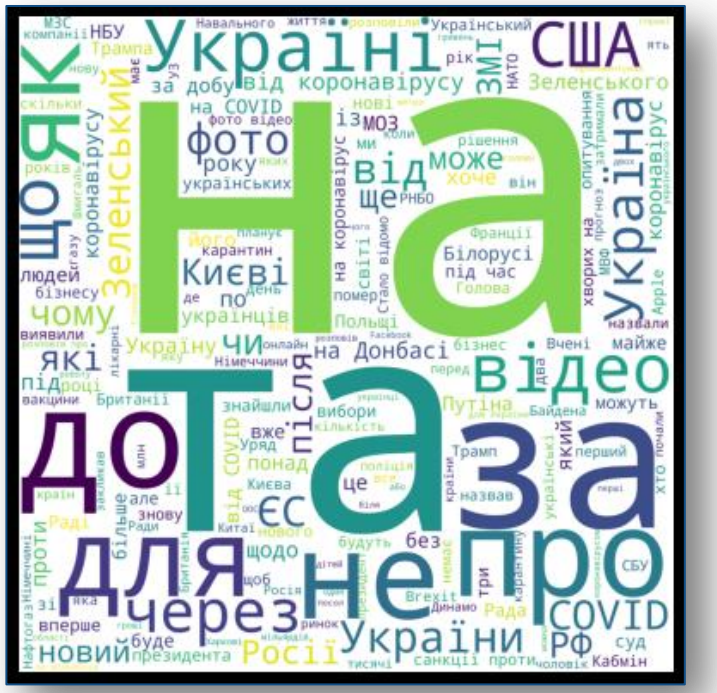
```
[[['Харківщина вкотре під вогнем: окупанти обстріляли житлові будинки та поранили літніх людей'],
  'https://tsn.ua/ato/harkivschina-vkotre-pid-vognem-okupanti-obstrilyali-zhitlovi-budinki-ta-poranili-litnih-lyudey-2086642.html'],
 [['Харків знову опинився в зоні досяжності ствольної артилерії супротивника - Арестович'],
  'https://tsn.ua/ato/harkiv-znovu-opinivsya-v-zoni-dosyazhnosti-stvolnoyi-artileriyi-suprotivnika-arestovich-2085913.html'],
 [['На Харківщині під час обстрілів загинула мати, її 3-місячну дитину - поранено'],
  'https://tsn.ua/ukrayina/na-harkivschini-pid-chas-obstriliv-zaginula-mati-yiyi-3-misyachnu-ditinu-poraneno-2086114.html'],
 [['Російські окупанти гатили по Харкову касетними бомбами - Amnesty International\xa0'],
  'https://tsn.ua/ato/rosiyski-okupanti-gatili-po-harkovu-kasetnimi-bombami-amnesty-international-2085040.html'],
 [['Російські окупанти завдяки штурму закріпилися на околиці Ізбицького на Харківщині - Генштаб'],
  'https://tsn.ua/ukrayina/rosiyski-okupanti-zavdyaki-shturmu-zakripilisya-na-okolici-izbickogo-na-harkivschini-genshtab-2085016.html'],
 [['Окупанти обстріляли населенні пункти Харківщини: є загиблі серед цивільних'],
  'https://tsn.ua/ukrayina/okupanti-obstrilyali-naselenni-punkti-harkivschini-ye-zagibli-sered-civilnih-2084587.html'],
 [['Окупанти б’ють по Харкову бомбами та ракетами більшої потужності: Терехов описав ситуацію в місті'],
  'https://tsn.ua/ato/okupanti-b-yut-po-harkovu-bombami-ta-raketami-bilshoyi-potuzhnosti-terehov-opisav-situaciyu-v-misti-2083501.html'],
 [['"Чесність є зброєю проти усього, що несе Росія": Зеленський – про відновлення українського телевізійного мовлення у Харкові'],
  'https://tsn.ua/ato/chesnist-ye-zbroyeyu-proti-usogo-scho-nese-rosiya-zelenskiy-pro-vidnovlennya-ukrayinskogo-televiziynogo-movlennya-u-harko
 [['"Своєрідна помста з боку Росії: військовий експерт - про збільшення обстрілів Харкова'],
  'https://tsn.ua/ato/svoyeridna-pomsta-z-boku-rosiyi-viyskoviy-ekspert-pro-zbilshennya-obstriliv-harkova-2082772.html'],
```

Бінарна класифікація україномовних новин – попередня обробка тексту (викидання стоп-слів)

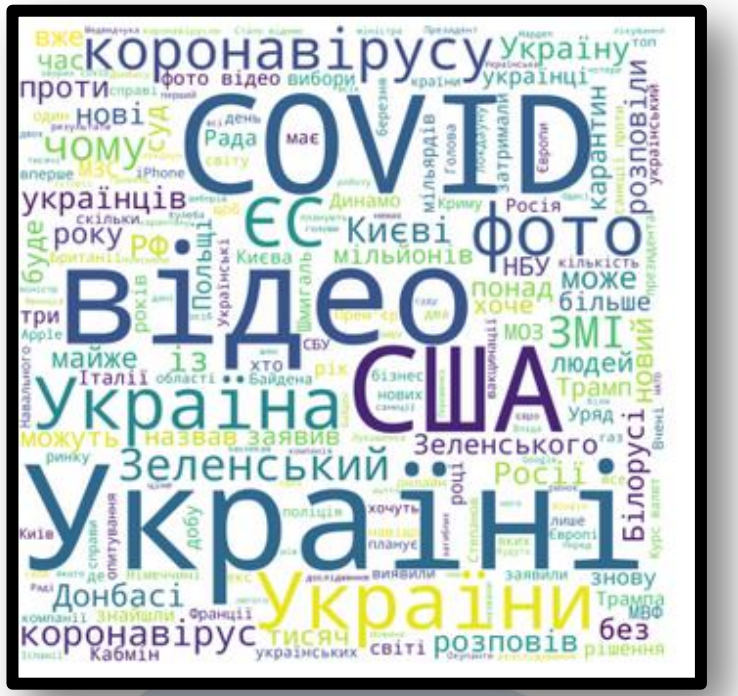
Список стоп-слів для української мови:

```
stop = [
u'я', u'а', u'так', u'але', u'тобі', u'мені', u'ти', u'ми', u'ви', u'і', u'в', u'у', u'на', u'з', u'із',
u'зі', u'за', u'його', u'там', u'як', u'які', u'який', u'яка', u'туди', u'давай', u'зовсім', u'що',
u'ну', u'не', u'ні', u'своя', u'своє', u'свої', u'свій', u'хоча', u'б', u'би', u'например',
u'така', u'такі', u'таке', u'такий', u'нам', u'хм', u'всім', u'ні', u'на', u'про', u'через', u'про',
u'він', u'вона', u'вони', u'воно', u'про', u'нас', u'них' u'тд', u'тощо', u'щодо', u'від', u'під', u'до', u'для',
u'вся', u'хтось', u'щось', u'вам', u'це', u'ця', u'ці', u'цей', u'або', u'чи', u'та', u'після',
u'просто', u'блін', u'дуже', u'самі', u'твої', u'ваша', u'наша', u'доречі', u'по', u'типу', u'пока', u'ок',
u'було', u'її', u'мого', u'моя', u'мої', u'моє', u'мій', u'ще', u'йому', u'їй', u'тепер',
u'ледь', u'потім', u'коли', u'навіть', u'моє', u'тепер', u'навіть', u'якщо', u'хто', u'без'
u'іноді', u'можна', u'нарешті', u'раз', u'той', u'ці', u'може', u'буде', u'вже', u'чому'
]
```

Бінарна класифікація україномовних новин (викидання стоп-слів)



до



після

Бінарна класифікація україномовних новин

```
620/620 [==============================] - 1s 2ms/step - loss: 0.4384 - acc: 0.8339
Training accuracy is 83.38885307312012
155/155 [==============================] - 0s 2ms/step - loss: 0.4571 - acc: 0.8294
Testing accuracy is 82.93963074684143
```

Точність навчальної та тестової вибірки складає – 83%.

Приклад роботи мережі: (новини не відносяться до політичних, тому клас 0 розпізнано вірно)

```
[22] sample_text=('Вакцинацію від COVID-19 планують проводити на КПВВ у Херсонській області – заступник міністра')
     oh=one_hot(sample_text,vocab_size,filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',lower=True,split=' ')
     pad=pad_sequences([oh],maxlen=max_length,padding='pre')
     predictions=model.predict(pad)
     print(predictions)

     [[0.17230341]]

[19] sample_text=('Лікарні Ужгорода заповнені на 100%, тяжкохворих направляють в інші райони області - ОДА')
     oh=one_hot(sample_text,vocab_size,filters='!"#$%&()*+,-./:;<=>?@[\]^_`{|}~',lower=True,split=' ')
     pad=pad_sequences([oh],maxlen=max_length,padding='pre')
     predictions=model.predict(pad)
     print(predictions)

     [[0.1679458]]
```

Мультикласова класифікація новин
Приклад роботи

Після побудови та навчання моделі, оцінимо класифікацію статей. Перша стаття відноситься до категорії бізнес. Друга стаття відноситься до 5 мітки – розваги.

```
[ ]  txt = ["house prices show slight increase prices of homes in the uk rose a seasonally adjusted 0.5% in february  says the nationwide building so

     seq = tokenizer.texts_to_sequences(txt)
     padded = pad_sequences(seq, maxlen=max_length)
     pred = model.predict(padded)
     labels = ['sport', 'bussiness', 'politics', 'tech', 'entertainment'] #orig

     print(pred)
     print(np.argmax(pred))
     print(labels[np.argmax(pred)-1])

     [[1.3026716e-04 7.9453217e-05 9.8438162e-01 7.5572925e-03 7.7977218e-04
       7.0715896e-03]]
     2
     bussiness
```

```
[26]  txt = ["call to save manufacturing jobs the trades union congress (tuc) is calling on the government to stem job losses in manufacturing firms b

      seq = tokenizer.texts_to_sequences(txt)
      padded = pad_sequences(seq, maxlen=max_length)
      pred = model.predict(padded)
      labels = ['sport', 'bussiness', 'politics', 'tech', 'entertainment']

      print(pred)
      print(np.argmax(pred))
      print(labels[np.argmax(pred)-1])

      [[1.4547000e-04 4.3164562e-03 2.4425244e-01 1.1483485e-02 1.7365050e-01
        5.6615162e-01]]
      5
      entertainment
```

Висновки

- Було розглянуто актуальність класифікації текстових даних.
- Проведено огляд та аналіз аналогів.
- Детально описано роботу моделі бінарної та мультикласової класифікації текстів.
- Проведено тестування та аналіз отриманих результатів точності навчання нейронної мережі.
- Результатами роботи є розробка модулів для бінарної класифікації текстів відгуків, збору початкових даних з сайту новин та мультикласової класифікації новин за категоріями.

Дякую за увагу!