

## ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

## КВАЛІФІКАЦІЙНА РОБОТА

- НА ТЕМУ: Метод аналізу тональності цитування публікації на основі контексту

ВИКОНАВ:  
Студент гр. СПзм-21-1  
Рисенко О. С.

КЕРІВНИК:  
Доц. Ільїна І.В.

ХАРКІВ  
2023р.

2

### МЕТА ТА ЗАВДАННЯ ДОСЛІДЖЕННЯ

**Мета:** є дослідження і розробка методів аналізу тональності цитувань у наукових статтях.

Для досягнення поставленої мети необхідно вирішити такі **завдання**:

1. Провести огляд існуючих методів автоматичного аналізу емоційного забарвлення текстів.
2. Провести дослідження текстових особливостей цитат в наукових публікаціях в контексті розробки методів аналізу їх емоційного забарвлення.
3. Розробити метод автоматичного аналізу емоційно забарвлених цитат.

# Компоненти і атрибути об'єкту

---



3

Другий тип думок – порівняння

---

- порівняння аспектів об'єктів на користь одного (non-equal gradable);
- прирівнювання аспектів різних об'єктів (equative);
- перевага одного об'єкта над іншими (superlative).

4

## Визначення об'єктивного і суб'єктивного речень

---

- об'єктивне речення відображає фактичну інформацію про щось, тоді як суб'єктивне речення виражає чийсь особисті почуття і припущення;
- об'єктивні речення зазвичай не мають емоційного забарвлення, тому аналіз тексту на наявність суб'єктивної інформації часто є підзадачею визначення полярності тексту.

5

## Приклад

---

Розглянемо приклад цитати «The proposed approach yields better accuracy».

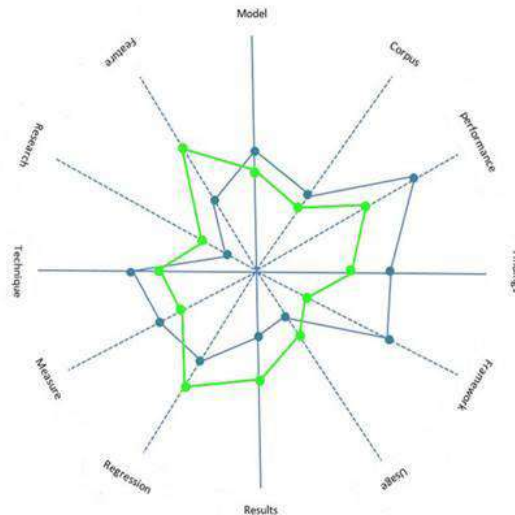
Його уніграми: «The», «proposed», «approach», «yields», «better», «accuracy»

Його триграма: «The proposed approach», «proposed approach yields», «approach yields better», «yields better accuracy»,

Його триграми: «The proposed approach», «proposed approach yields», «approach yields better», «yields better accuracy»,

6

## Профіль настрою цитування на основі аспектів



7

## Список тегів POS та їх значення

Теги частини мови	Значення
JJ, JJR, JJS	Прикметник, порівняльний прикметник, називний відмінок прикметника
NN, NNS, NNP, NNPS	Іменник одна, іменник множина, власна назва
CC, PRP, CD	Узгоджувальний сполучник, особистий займенник, цифрове значення
RB, RBR, RBS	Прислівник, порівняльний прислівник, чудова ступінь прикметників
VB, VBD, VBG, VBN, VBP, VBZ	Дієслово, дієслово минулого часу, дієслово в теперішньому часі, дієприкметник минулого часу

8

# Оцінка настрою ознак біграми

БІГРАМИ	SENT VAL	ТИП
the same	0	Нейтральний
same cut	0	Нейтральний
a reasonable	0.45	Позитивний
reasonable starting	0.45	Позитивний
subsequent research	0.30	Позитивний
neglected notable	-0.49	Негативний

9

D	Оцінка (n=1)	Тип (n=1)	Оцінка (n=2)	Тип (n=2)	Оцінка (n=3)	Тип (n=3)	Оцінка (n=5)	Тип (n=5)
1	0,37	П	0,73	П	1.1	П	1.83	П
2	0	О	0	О	0	О	0	О
4	0,44	П	1.18	П	1.63	П	2.51	П
10	0	О	-0,49	Н	-0,98	Н	-1,96	Н
11	0	О	0	О	0,41	П	0,41	П
17	0	О	0	О	-0,49	Н	-1,96	Н

10

## Шаблони аспектів-кандидатів

Шаблони аспектів-кандидатів можна розділити на три різні категорії:

- іменник-прикметник: відповідає поєднанню герундія іменника та прикметника у якому прикметник описує іменник, наприклад, фрагмент цитати «модель було вдосконалено»;
- іменник-дієслово: відповідає комбінаціям герундія іменника та дієслова, у яких дієслово описує іменник, наприклад, цитатне речення «we refined the corpus»;
- дієслово-прислівник: поєднання герундія дієслова та прислівника означає, що дієслово описується прислівником, наприклад, «техніка використовується широко».

11

## Опис експериментального набору даних

	Вибрані документи	Цитувані речень	Позитивні цитати	Негативні цитати	Нейтральні цитати
Набір даних 1	310	8736	829 (9,5%)	280 (3,2%)	7627 (87,3%)
Набір даних 2	285	4182	702 (17%)	308 (7,36%)	3172 (75,9%)

12

# Приклад контексту цитування

Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass. While it does not adopt a set-based or distributional model of relational similarity. We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.

citationText	Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass.	P
postSentence1	While it does not adopt a set-based or distributional model of relational similarity.	N
postSentence2	We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.	N

```

<!--annotatedPaper id="1"-->
<paper title="Using lexical and relational similarity to classify semantic relations" year="2009" authors="Diarmuid O Seaghdha, Ann Copestake"
  <citationContext>
    <citationText>
      Turney (2008) has recently proposed a simpler SVM-based algorithm for analogical classification called PairClass.
    </citationText>
    <section> Literature Review </section>
    <cite polarity="p"> Turney (2008) </cite>
  </citationContext>
  <postSentence1>
    While it does not adopt a set-based or distributional model of relational similarity.
  </postSentence1>
  <postSentence2>
    We have noted above that PairClass implicitly uses a feature representation which does not produce compatible results.
  </postSentence2>
</paper>
</annotatedPaper>

```

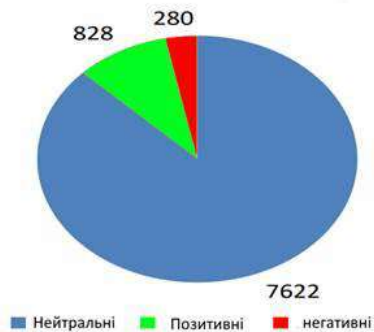
# Топ-20 прикметників

Most Frequent Adjective Уніграми (n=1)	High, good, successful, effective, popular, efficient, important, interesting, robust, difficult, considerable, easy, competitive, suitable, poor, central, appealing, straightforward, satisfactory, dominant
Most Frequent Adjective Біграми (n=2)	such as, statistical machine, natural language, good performance, significant improvements, human judgments, minimum error, maximum entropy, automatic evaluation, the same, human evaluation, more accurate, in recent, current state, Recent work, more sophisticated, very successful, be effective, most notable
Most Frequent Adjective триграми (n=3)	such as the, minimum error rate, statistical machine translation, with human judgments, in recent years, phrase based statistical, from many heterogeneous, In recent years, Introduction In recent, the past few, to be effective, past few years, successful in recent, current state of, greedy search algorithm, various NLP tasks, techniques such as
Most Frequent Adjective пентаграми (n=5)	proven increasingly successful in recent, increasingly successful in recent years, major developments in statistical approaches, current state of the art, notable examples of unsupervised polarity, received much attention in recent, many advances in recent years, such as part of speech, past five years important research, Introduction In recent years statistical

# Precision, Recall and Accuracy

Шаблони	Manual			SEMANTRIA			THEYSAY		
	P	R	F-score	П	P	F-бал	П	P	F-бал
Adj. (1-g)	70,54	40	55,27	75,5	29,17	52,34	44,45	80	62,23
Adj. + Adv. (1-g)	75,54	55,56	65,55	94,11	55,56	74,84	53,35	78,17	65,76
Adj. (2-g)	75	50	62,5	70	54,17	62,09	55,56	76,92	66,24
Adj. + Adv. (2-g)	88,89	80	84,45	85,5	92,30	88,9	58,06	85,71	71,89
Adj. (3-g)	65,55	50	57,78	96	58,33	77,17	70,54	80	75,27
Adj. + Adv. (3-g)	88,89	72,72	80,81	96	68,57	82,29	59,38	82,60	70,99
Adj. (5-g)	75,54	60	67,77	94,11	66,67	80,39	54,17	81,26	67,72
Adj. + Adv. (5-g)	90	81,82	85,91	96,15	96,15	96,15	55,88	82,60	69,24

15



Запропонований метод



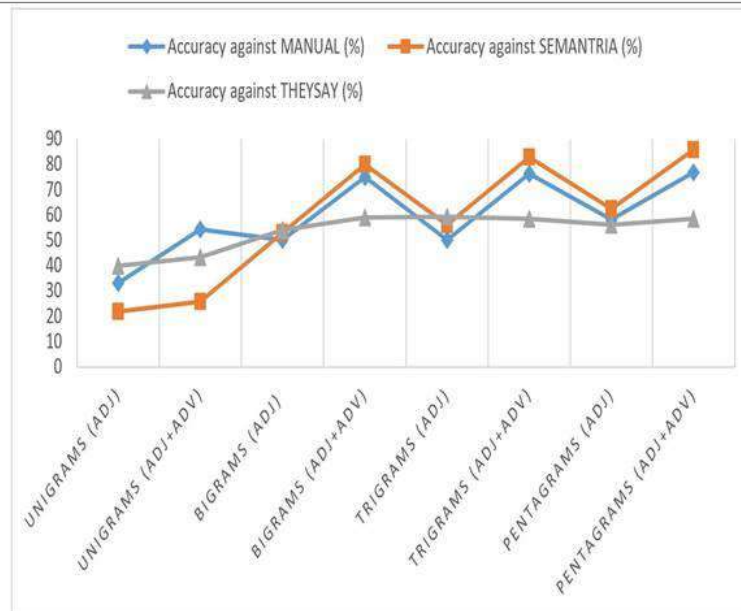
Відсоток міток класів, створених SEMANTRIA



Відсоток міток класів, створених THEYSAY

16

## Порівняння точності



17

## ВИСНОВКИ

В кваліфікаційній роботі було досліджено ефективність використання прикметників, прислівників та їхніх комбінацій для класифікації настрою цитованого тексту на рівні документа за допомогою золотого стандарту набору настрою цитування.

Після цього було досліджено ефективність застосування комерційних інструментів до набору цитат для виявлення настроїв. Аналіз результатів виявив, що n-грами високого порядку ( $n = 5$ ) для комбінацій прикметників і прислівників відіграють основну роль у покращенні точності класифікації настрою.

Результати дослідження показали точність понад 80% порівняно з анотованим вручну набором даних із використанням комбінацій прикметників і прислівників із n-грамою вищого порядку. Порівняльний аналіз результатів із сучасним дослідженням, заснованим на використанні n-грамів різної довжини, показав, що уніграма, біграма та триграма показали кращі результати у комбінації порівняно з уніграмою та уніграмою плюс біграма.

Експериментальні результати також визначили, що поточні інструменти аналізу настроїв, точніше THEYSAY, недостатньо ефективні для точного визначення настроїв, оскільки більшість цитат і триграма показали кращий результат у комбінації порівняно з уніграмою та уніграмою плюс біграма.

18