**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**

**«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ. І. І. СІКОРСЬКОГО»**

**ННК «ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»**

**Кафедра системного проектування**

**Курсова робота**

**з дисципліни**

**Оброблення надвеликих масивів даних**

Тема: «Передбачення антагоніста та протагоніста літературного твору»

Виконала: студентка групи ДА-11мп

Молчанова В.С.

Перевірив: професор

Рогоза В.С.

Київ – 2021

## ЗМІСТ

[ВСТУП 3](#_Toc90814486)

[ВИЗНАЧЕННЯ ПЕРСОНАЖІВ 4](#_Toc90814487)

[ПОБУДОВА ТА АНАЛІЗ ГРАФУ ВІДНОСИН ПЕРСОНАЖІВ 8](#_Toc90814488)

[АНАЛІЗ РИС ПЕРСОНАЖІВ ЗА МОДЕЛЛЮ BIG5 ЗА ДОПОМОГОЮ DEEP LEARNING 10](#_Toc90814489)

[АНАЛІЗ ОПИСІВ ПЕРСОНАЖІВ 13](#_Toc90814490)

[ВИСНОВКИ 14](#_Toc90814492)

[ВИКОРИСТАНІ ДЖЕРЕЛА 15](#_Toc90814493)

## ВСТУП

Комп’ютерний аналіз та моделювання наративів чи поезії все ще є проблемою для досліджень у цифрових літературознавствах, комп’ютерній лінгвістиці, машинному навчанні чи нейрокогнітивній поетиці. Вивчення спільнот у літературі - відносно незвідана область через безліч рівнів аналізу, необхідних для отримання інформації з книги в мережу.

В даній роботі будуть запропоновані декілька алгоритмів, спрямованих на визначення протагоністів та антагоністів літературних творів. Через новизну цього напрямку досліджень, в якості цільових текстів для визначення ролей персонажів були обрані дитячі казки, адже персонажі в них найчастіше мають простий та чітко визначений характер. Також вони зручні для аналізу завдяки простій мові написання.

## ВИЗНАЧЕННЯ ПЕРСОНАЖІВ

Для того щоб визначити, які герої твору є його протагоністами, а які – антагоністами, потрібно для початку отримати список усіх (або хоча б головних) персонажів, який для кожного з них включатиме якомога більше різних варіантів імені, фамілії, посади або іншого способу ідентифікації, що використовуються в творі.

Одним з найпопулярніших методів вилучення списку персонажів твору є розпізнавання іменованих сутностей (Named Entity Recognition, NER) — підзавдання вилучення інформації, яке прагне знайти та класифікувати названі об’єкти, згадані в неструктурованому тексті, у заздалегідь визначені категорії, такі як імена людей, організації, місця розташування, медичні коди, часові вирази, кількості, грошові оцінки, відсотки тощо.

Більшість досліджень систем NER було структуровано таким чином, щоб взяти блок тексту без коментарів, наприклад цей:

«Джим купив 300 акцій Acme Corp. у 2006 році.»

І створити анотований блок тексту, який виділяє назви сутностей:

[Джим]особа купив 300 акцій [Acme Corp.]організація у [2006]Час.

У цьому прикладі було виявлено й класифіковано ім’я особи, що складається з одного токена, назви компанії з двома токенами та часового виразу.

У виразі named entity слово named обмежує завдання тими сутностями, для яких один або кілька рядків, наприклад слова чи фрази, є (досить) послідовно для певного референта. Наприклад, автомобільну компанію, створену Генрі Фордом у 1903 році, можна називати Ford або Ford Motor Company, хоча «Форд» також може посилатися на багато інших організацій (див. Ford у Вікіпедії). Жорсткі позначення включають власні назви, а також терміни для певних біологічних видів і речовин, але виключають займенники, описи, які виділяють референт за його властивостями, а також назви видів речей на відміну від окремих осіб (наприклад, «Банк»).

Повне розпізнавання іменованих сутностей часто розбивається, концептуально як дві різні проблеми: виявлення імен і класифікація імен за типом сутності, на яку вони посилаються (наприклад, особа, організація або місце розташування ). Перший етап зазвичай спрощується до проблеми сегментації: імена визначаються як суміжні проміжки маркерів без вкладення, тому «Bank of America» є єдиною назвою, не враховуючи той факт, що всередині цього імені міститься підрядок «America», який сам по собі є ім'ям. Другий етап вимагає вибору онтології, за допомогою якої будуть організовані категорії речей.

Вала, Юргенс, Пайпер та Рус у своїй роботі «Містер Беннет, його кучер і архієпископ заходять до бару, але тільки одного з них розпізнають: про складність виявлення персонажів у літературних текстах» запропонували покращений, восьмиступеневий алгоритм для виявлення персонажів, який будує граф, де вузли є іменами, а ребра з’єднують імена, що належать одному персонажу. Вершини в графі спочатку заповнюються за допомогою NER, включаючи імена та фамілії після почесних звань (Mr., Mrs., Ms., Captain, Coach, Professor тощо). По-друге, виконується аналіз кореферентності (явище, яке позначає відношення між назвами (знаковими одиницями), компонентами висловлювання, при якому кілька компонентів відсилають до одного об'єкта позамовної дійсності (референта)), щоб ідентифікувати імена, які зустрічаються разом у ланцюжку кореференцій, та додаються ребра графу там, де імена двох вузлів зустрічаються разом у ланцюжку. Stanford CoreNLP використовується як для NER, так і для кореференцій. По-третє, застосовується низка правил варіації імен, щоб зв’язати ті з них, які потенційно посилаються на ту саму людину (наприклад, видаляючи почесне звання). По-четверте, для зв’язування варіацій (наприклад, Тім і Тіммі) використовується датасет із 1859 варіантами для 560 імен.

Етапи 2–4 потенційно вводять ребра, що з’єднують імена різних персонажів. Тому на п’ятому етапі застосовуються три евристики для додавання заборон на злиття двох імен в одну особу. Дві вершини не можна об’єднати, якщо (1) статі імен відрізняються, (2) обидва імена мають спільне прізвище, але різні імена, або (3) почесне звання обох імен відрізняється, наприклад, «міс» і «місіс». ». Так само на шостому етапі вводяться заборони шляхом виділення пар імен з твору, де (1) обидва назви з’єднані сполучником, (2) одне ім’я з’являється як мовець, який згадує інше ім’я в прямій мові, і (3) обидві назви з’являються разом у цитаті.

Разом, етапи 1–6 застосовуються спершу шляхом зв’язування всіх вузлів ребрами і далі, визначення пар, заборонених до з’єднання, і видалення ребер уздовж найкоротшого шляху між цими двома вузлами, фактично створюючи два нових роз’єднаних компонента в графі імен. Рисунок 1 ілюструє це перетворення на простому графі.

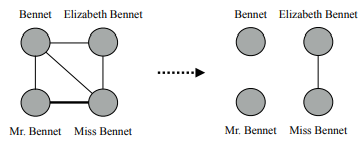


Рисунок 1. Розв’язування імен у графі символів. Кола

представляють окремі назви, а тонкі та товсті лінії позначають

ребра та антиребра відповідно.

Далі, сьомий крок намагається ідентифікувати персонажів, чиї імена можуть не розпізнаватися NER. Наприклад, багато другорядних персонажів не відображаються як названі сутності, а натомість мають загальні рольові референти, такі як «губернатор» або «архієпископ». Однак, незважаючи на відсутність власних імен, такі персонажі поводяться та взаємодіють так само, як і головні персонажі, у тому числі ведуть діалоги. Тому, щоб виявити таких персонажів, використовується техніка завантаження, спрямована на виявлення прототипів поведінки персонажів із самих романів, натхненних семантичними предикатами Флаті та Навіґлі (2013). Корпус художньої літератури Project Gutenberg був аналізований залежностями, щоб ідентифікувати всі дієслова у відношенні залежності з іменниками, де кожен іменник був класифікований як (a) названий об’єкт (b) той, що має своє перше значення в WordNet, відноситься до одушевленої сутності. , або (c) жодного з вищезазначеного. Усі дієслова, пов’язані з ними, були ранжовані відповідно до їх співвідношення типів (a) і (c), щоб визначити дієслова, сильно пов’язані з характерною поведінкою, що дозволяє уникнути включення поведінки іменників у (b), які можуть посилатися на другорядні символи. Використовуючи цей набір предикатів, іменники, що з’являються з дієсловом у відповідному відношенні залежності, додаються як персонажі. Забороняється додавати імена, що містяться в невеликому списку з 22 родових іменників (наприклад, «людина»).

Нарешті, восьмий і останній етап видаляє вузли, які від’єднані від решти графіка і представляють ім’я, яке є частиною одного або кількох імен для інших вузлів. Ці вузли, як правило, є неоднозначними іменами чи прізвищами. Таким чином, набір вузлів об’єднується, щоб створити набори імен, кожне з яких пов’язане з певним персонажем [2].

Після застосування цього метода визначення списку персонажів, можна перейти до трьох запропонованих методів визначення їх ролі у тексті, описаних нижче.

## ПОБУДОВА ТА АНАЛІЗ ГРАФУ ВІДНОСИН ПЕРСОНАЖІВ

Метод аналізу соціальних мереж, заснований на тріадах графів був розроблений студентами Фернандезом та Улмером [1]. Після отримання списку персонажів, вони витягують зв’язки між ними (найчастіше це дієслівний зв’язок) та використовують бібліотеку під назвою SentiWordNet, щоб виконати аналіз настрою цих відносин.

Далі їх алгоритми аналізу мережі беруть дані про взаємозв’язок між персонажами і будують граф, який містить інформацію про позитивність, негативність і важливість кожного персонажа і його відносин. Оброблені дані ідентифікації персонажів та вилучення настроїв зберігаються у вигляді списку з’єднань і позитивних та негативних ваг, відповідних  цьому  з’єднанню. Ці дані дозволяють будувати мережу за допомогою SNAP і створювати мультиграфи для позитивних і негативних категорій.

Аналіз починається з більш простих шарів інформації та тенденцій до більш складних. Перша група функцій (яка враховує позитивні та негативні ваги) включає:

1. Особи з найбільшим ступенем
2. Особи з найбільшим значенням суми всіх ваг ребер.
3. Найбільша різниця між позитивними та негативними вагами для визначення найбільш позитивних персонажів.
4. Найменша різниця (може бути від’ємною) між позитивними та негативними ваговими коефіцієнтами для визначення найбільш негативних персонажів.

Цей набір функцій дає уявлення про структуру графів, але не приводить до значущих результатів для прогнозування інформації про історію. Потім виконується дослідження графів за допомогою аналізу тріад соціальних мереж. Досліджуючи тріади, які існують у соціальному графі, можна знайти дві основні частини інформації про історію. По-перше, можна визначити, чи є граф сильно, слабко чи не структурно збалансований. Ця інформація дозволяє визначити історії, які є складнішими і, швидше за все, містять обман чи складних персонажів, оскільки тріади стосунків є незбалансованими. По-друге, важлива інформація, отримана в результаті аналізу тріади, — це розуміння взаємовідносин. Наприклад, персонажі на негативному кінці тріад Т1, швидше за все, будуть антагоністами, а головний герой часто буде в тріадах Т3 з іншими союзниками або позитивними персонажами (див. Рисунок 2).

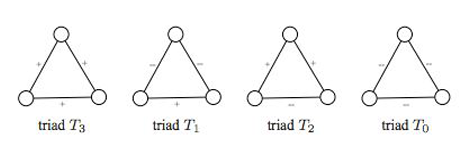


Рисунок 2. Типи тріад.

Цей аналіз тріад, на додаток до інших функцій, які не пов’язані з тріадами, але додають інтуїтивну інформацію про емоційну валентність персонажа в історії, було зосереджено на простому, але ефективному класифікаторі, заснованому на правилах, для ідентифікації головного героя та антагоніста. Додаткові функції:

1. Кількість разів, коли з’являлися в історії в цілому.
2. Ступінь особистості.
3. Значення для суми всіх ваг ребер у позитивному/негативному напрямку.

## АНАЛІЗ РИС ПЕРСОНАЖІВ ЗА МОДЕЛЛЮ BIG5 ЗА ДОПОМОГОЮ DEEP LEARNING

Для аналізу характеру людини психологи зосереджуються на рисах особистості. Кожна риса зустрічається за певним спектром, і риси незалежні одна від одної, створюючи нескінченну варіативність людської особистості.

Риси, які підтверджують найсильніші дослідження, відомі як Велика п’ятірка (Big 5):

1. Відкритість (openess). Відкритість — це скорочення від «відкритість досвіду». Люди з високою відкритістю люблять пригоди. Вони цікаві та цінують мистецтво, уяву та нові речі. Девіз відкритої особистості може бути таким: «різноманітність — це пряність життя». Люди з низьким рівнем відкритості – навпаки: вони вважають за краще дотримуватися своїх звичок, уникати нових вражень і, ймовірно, не є найбільш авантюрними особами.
2. Сумлінність (conscientiousness). Люди сумлінні організовані і мають сильне почуття обов’язку. Вони надійні, дисципліновані та орієнтовані на досягнення. Люди з низьким рівнем свідомості більш спонтанні та вільні. У крайньому випадку, вони можуть бути схильні до необережності.
3. Ектраверсія (extraversion). Екстраверсія проти інтроверсії є, можливо, найбільш впізнаваною рисою особистості Великої п’ятірки. Чим більше хтось екстраверт, тим більше він соціальний метелик. Екстраверти балакучі, товариські і черпають енергію з натовпу. Вони мають тенденцію бути наполегливими та життєрадісними у своїх соціальних взаємодіях. Інтровертам, з іншого боку, потрібно багато часу на самоті.
4. Приємність (agreeableness). Приємність вимірює ступінь теплоти та доброти людини. Чим приємніший хтось, тим більша ймовірність, що він буде довірливим, корисним і співчутливим. Неприємні люди холодні й підозрілі до інших, і вони рідше співпрацюють.
5. Невротизм (neuroticism). Люди з високим рівнем невротизму часто хвилюються і легко впадають у тривогу та депресію. Якщо все йде добре, невротичні люди схильні знаходити те, про що хвилюватися. Оскільки люди з високим рівнем нейротизму, як правило, відчувають багато негативних емоцій, нейротизм відіграє певну роль у розвитку емоційних розладів, відповідно до статті, опублікованої в Clinical Psychological Science. Навпаки, люди з низьким рівнем нейротизму, як правило, емоційно стабільні та рівномірні.

На мою думку, деякі з властивостей Big5 можна застосувати в визначенні ролі героїв у творі. В залежності від жанру та складності твору, набор цих властивостей та їх вага буде змінюватися. Наприклад, якщо продовжити фокусуватися на казках, то наявність у персонажа приємності та/або сумлінності – дуже сильне свідоцтво на користь того, що він є протагоністом історії, а наявність невротизму може свідчити про те, що персонаж є антагоністом. І хоча в більш складних творах не можна використовувати цей метод як основний, адже персонажі в них найчастіше мають складний та неоднозначний характер, при правильному визначенні ваг властивостей великої п’ятірки, він може добре використовуватися як допоміжний.

Для визначення, якому персонажу які властивості притаманні, можна використовувати метод, запропонований італійцями Даніеле Піццоллі та Карло Страппарава в статті «Personality Traits Recognition in Literary Texts» [3]. Вони підійшли до цієї задачі як до проблеми навчання під керівництвом, використовуючи позначений набір даних, а потім передавши результат у їх набір даних.

У літературознавстві важко знайти класифікацію персонажів за якоюсь моделлю особистості. Літературні критики часто воліють глибше аналізувати персонажа, а не відносити їх до простих категорій. В основі використаної моделі і взагалі в згаданих рамках лежить лексична гіпотеза: принаймні в якійсь мірі, дозволено вивести риси особистості з мови та зі слів. З психологічної точки зору, підтримка лексичної гіпотези є в Ashton and Lee [4].

Як розмічений набір даних вони використали «есе», спочатку від Pennebaker and King (1999). «Есе» — це великий набір даних із текстів потоку свідомості (близько 2400, по одному для кожного автора/користувача), зібраних між 1997 та 2004 роками та позначених класами особистості. Тексти підготували студенти, які заповнили анкети Big5. Основна причина використання цього набору даних полягає в тому, що він єдиний містить «золоті» мітки, які підходять для виконуваної задачі.

Модель була побудована, використовуючи інструментарій Scikitlearn для машинного навчання на Python (Pedregosa et al., 2011). Початкове завдання полягало в тому, щоб отримати належну продуктивність набору даних «есе». Проблема відноситься до класу міток з кількома виходами. Для простоти кожну мітку (відповідає особистій рисі) можна розглядати як незалежну, розділивши проблему на 5 задач класифікації. Згідно з пропозицією (Celli et al., 2013, 2014), модель була побудована як конвеєр, що включає як модель пакета слова, так і вихідний текст TextPro. Для покращення продуктивності моделі можливо багато налаштувань (наприклад, створення іншої моделі для кожної ознаки або використання різних функцій чи класифікатора).

Після навчання на «золотом» наборі даних, модель можна запустити на цільовому творі, та, після отримання її результатів, застосувати розроблені для даного типу твору ваги кожної властивості особистості для виявлення антагоніста та протагоніста, як було описано вище.

## АНАЛІЗ ОПИСІВ ПЕРСОНАЖІВ

Описані вище методи аналізу персонажів базуються в основному на їх взаємовідносинах та діалогах. Однак в простих сюжетно, але багатих на епітети та образи творах (якими як раз є казки), на мою думку, позитивність/негативність персонажу можна дослідити за допомогою аналізу його опису, в тому числі зовнішнього.

Зробити це можна, за допомогою парадигми MapReduce та використання датасету прикметників з розбиттям їх на позитивні та негативні, наприклад, Opinion Lexicon [5] за таким алгоритмом:

1. Після отримання списку персонажів, за методом, описаним в другому розділі, зробити інвертований індекс, що буде показувати на кожне місце, де з’являється певне ім’я якогось персонажу.
2. Обрати інтервал пошуку опису персонажу (наприклад, 2 речення після імені).
3. Для кожного індексу в обраному інтервалі знайти слова, що є в обраному датасеті прикметників та порахувати «позитивність» даного опису.
4. Виконати reduce в 2 етапи – спочатку по конкретному варіанту імені, потім зібрати разом різні імена кожного персонажу, сумуючи «позитивність»

Отримаємо дані, які можна використовувати як самостійну оцінку персонажів, або як допоміжний критерій в комплексному аналізу за допомогою декількох методів.

## ВИСНОВКИ

Проблема комп’ютерного аналізу літературних творів, а саме виявлення його персонажів та їх ролей в сюжеті, є складною та багатоаспектною задачею, яку можна намагатися вирішувати з різних сторін.

В даній роботі було розглянуто метод визначення списку персонажів літературного твору та три різних метода виявлення антагоністу та протагоністу твору: на основі відносин персонажів, на основі їх висловів та на основі їх описів. Кожен метод використовує свій унікальний підхід до вирішення поставленої задачі та може використовуватися як сам по собі, так і в комбінаціях з іншими. Через те, що другий та третій методи поки що не перевірялись на тестових датасетах, при поєднанні результатів різних методів їм треба призначити меншу вагу, ніж першому. Також ваги впливу результатів методів різних методів та ваги різних факторів самих методів можуть залежати від складності твору та його жанру.

Застосування розглянутих методів було описане в основному на прикладі дитячих казок, але можна вірити, що при певному доопрацюванні вони зможуть показати високі результати навість на більш складних творах.

## ВИКОРИСТАНІ ДЖЕРЕЛА

1. Matt Fernandez, Ben Ulmer. Extracting Social Network from Literature to Predict Antagonist and Protagonist. – 2015. – 10 с.
2. Hardik Vala, David Jurgens, Andrew Piper, Derek Ruths. Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. – 2015. – 6 с.
3. Daniele Pizzolli, Carlo Strapparava. Personality Traits Recognition in Literary Texts. – 2019. – 5 c.
4. Michael C. Ashton, Kibeom Lee. A defence of the lexical approach to the study of personality structure. – 2005. – 19 c.
5. Hu Xu, Bing Liu. Opinion Lexicon. Режим доступу: <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon>