Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____

Кафедра _____Програмної інженерії_____

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка
рівень вищої освіти – другий (магістерський)стану

Дослідження методів кластеризації для аналізу стану користувача мережі Інтернет

Виконав: студент 2 курсу, групи ІПЗм-18-1___
_____Брижатов Д. С._____
(прізвище, ініціали)

спеціальності 121- Інженерія програмного забезпечення
(код і повна назва спеціальності)

Освітньо-наукова програма_____
(тип програми)

Інженерія програмного забезпечення____
(повна назва освітньої програми)

Керівник _____д. т. н. проф. Четвериков Г. Г___
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2020 р.

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет_Комп'ютерних наук_____

Кафедра_Програмної інженерії_____

Рівень вищої освіти - другий (магістерський)

Спеціальність _121-Інженерія програмного забезпечення_____
(код і повна назва)

Тип програми освітньо-наукова програма_____

Освітня програма_ Інженерія програмного забезпечення_____

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові _Брижатову Дмитру Сергійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи _Дослідження методів кластеризації для аналізу стану користувача мережі інтернеті
затверджена наказом університету від "___" _____20 ___ р № _____
2. Термін подання студентом роботи до екзаменаційної комісії_____
3. Вихідні дані до роботи методи кластеризації, семантичний аналіз, пояснювальна записка.
4. Перелік питань, що потрібно опрацювати в роботі мета роботи, аналіз проблемної галузі і постановка задачі, аналіз можливих методів кластеризації, приведення текстових даних до формату який можливо кластеризувати, семантичний аналіз тексту, аналіз результатів кластеризації.

5. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	д.т.н. проф. Четвериков Г. Г.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка*
1	Аналіз предметної галузі	21 січня 2020 р.	
2	Виявлення пріоритетних властивостей методів кластеризації	10 лютого 2020 р.	
3	Моделювання архітектури	17 лютого 2020 р.	
4	Кластеризація семантичних даних	23 березня 2020 р.	
5	Підготовка пояснювальної записки	13 квітня 2020 р.	
6	Підготовка презентації та доповіді	04 травня 2020 р.	
7	Нормоконтроль, рецензування	07 травня 2020 р.	
8	Занесення диплома в електронний архів	09 травня 2020 р.	
9	Попередній захист	17 травня 2020 р.	
10	Допуск до захисту у зав. кафедри	20 травня 2020 р.	

Дата видачі завдання 20 січня 2020 р.

Студент _____
(підпис)

Керівник роботи _____ д. т. н. проф. Четвериков Г. Г
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи магістра, 65 с., 9 рисунків, 11 таблиць, 10 джерел.

АТЕСТАЦІЙНА РОБОТА, АНАЛІЗ ДАНИХ, МЕТОД K-MEANS, TWEETER, AFINN-111, BIG DATA, КЛАСТЕРИЗАЦІЯ, ІЄРАРХІЧНИЙ АНАЛІЗ, СЕМАНТИЧНИЙ АНАЛІЗ.

Об’єктом дослідження є кластеризація. Предметом дослідження є методи кластеризації для семантичного аналізу користувачів у мережі Twitter.

Метою роботи є кластеризація стану користувачів за допомогою семантичного аналізу.

У даній роботі був запроваджена кластеризація даних, стану користувача, запроваджених семантичним аналізом. Завдяки бібліотеці TweetSharp були отримані дані з мережі Twitter. Семантичний аналіз даних виконувався за допомогою AFINN-111. Кластерний аналіз було запроваджено за допомогою методу k-means та hierarchical clustering. Алгоритми були реалізовані за допомогою мови програмування R.

ATTESTATION WORK, DATA ANALYSIS, K-MEANS, TWEETER, AFINN-111, BIG DATA, CLUSTERIZATION, HIERARCHICAL ANALYSIS, SEMANTIC ANALYSIS.

The object of research is clustering. The subject of the research is clustering methods for semantic analysis of users on Twitter.

The aim of the work is to cluster the state of users using semantic analysis.

In this work, clustering of data, user status, introduced by semantic analysis. Thanks to the TweetSharp library, data was received from the Tweeter network. Semantic analysis was performed using AFINN-111. Cluster analysis was performed using the method hierarchical clustering with help R language.

ЗМІСТ

ВСТУП

Ще з самому початку інтернет, як технологія, створювався як закритий простір для взаємодії різних американських військових та державних організацій. Швидка комунікація між усіма вузлами мережі було основною перевагою такої технології. Через деякий час різні університети і вчені по всьому світу також зацікавилися цією технологією, а через тридцять років інтернетом почнуть цікавитися і звичайні люди.

Саме явище Інтернету породило безліч можливостей для кожної людини. На сьогоднішній день інтернет відіграє важливу роль в житті суспільства. Люди створюють терабайти різної інформації, обмінюючись їй між собою. Інтернет зачіпає важливі соціальні сфери людини, як допомога хворим людям або обговорення екологічних проблем. Різні державні установи або великі приватні організації вже повністю перестроїли своє управління під цю технологію. Однозначно можна сказати що відмова від цієї технології практично неможливий, тому що це стало невід'ємною частиною життя кожної людини. Однак саме цей факт породжує масу проблем яку привносить Інтернет. Ця технологія приховує в собі небезпеку, особливо його темна частина, так званий DarkNet. Але ж крім нього існують різні злочинні організації, спільноти хакерів і терористичні організації завдають шкоди державному ладу, і просто божевільні які можуть нанести звичайним людям.

В основі демократичного суспільства лежить свобода слова, лібертаріанські погляди. Інтернет як технологія, частково, отримав таку популярність через те, що люди, тобто його користувачі, ставали вільними в своїх висловлюваннях та діях. Соціальні мережі такі як Twitter, Facebook, Instagram і різні форуми, ставали майданчиками де люди збираються для дискусій на різні теми, в тому числі і політичні. Після терактів 11 вересня 2001 року, які відбулися в Нью-Йорку, всесвітні державні організації всерйоз

зацікавилися тотальним стеженням за людьми, контролем свободи слова та дій в інтернеті.

Існує дуже багато різних дискусій на моральну складову стеження за громадянами держави. Недавні події з Едвардом Сноуденом викликали величезний колапс в світовому інформаційному полі. Він розкрив факти стеження США не тільки за громадянами, а й стеження за іншими незалежними країнами. Це породило величезний світовий скандал. Уряд США розшукує Сноудена по всьому світу. Деякі американські громадяни вважають його зрадником. Деякі вважають героєм. Думка людей сильно різниться, проте головною причиною за якими держава застосовує таку стеження це запобігання глобальних конфліктів і терактів.

Семантичний аналіз тексту це один з інструментів який може задовольнити цей попит. Але на сьогоднішній день інтернет представляє з себе величезний масив різних даних які потрібно класифікувати за допомогою алгоритмів кластеризації. Тільки після кластеризації даних можна проводити аналіз найбільш ймовірних подальших дій користувача або організації.

1 СИСТЕМНИЙ АНАЛІЗ ПРОБЛЕМИ

1.1 Системний аналіз предметної області

Проведення даної дослідницької роботи ставитися все більш актуальною так як кількість самих користувачів та інформації, якої вони обмінюються між собою, з кожним роком тільки зростає. Відповідно потрібно знайти ефективні алгоритми для підготовки, тобто кластеризації, великого об'єму інформації. Аналізуючи інформацію яку користувач пише, в соціальних мережах або будь-яких інших джерелах, можна зробити різні висновки про його психологічний стан, таким чином сучасні алгоритми машинного навчання можуть брати цю семантично кластерізовану інформацію для подальшого аналізу найбільш ймовірної поведінки користувача. Виконуючи семантичний аналіз найбільш ефективно можна швидко реагувати на різні загрози яку користувач може становити соціуму.

1.2 Системний аналіз методів кластеризації

Для систематичного аналізу методів кластеризації скористаємося ієрархічним аналізом. Для нього потрібно сформувати чіткі критерії, за якими буде виконуватися аналіз методів кластеризації:

- точність;
- складність;
- універсальність;
- кількість необхідних даних;
- обсяг обчислень.

Існує велика кількість методів кластеризації. У даній дипломній роботі будуть розглянути три найпопулярніших:

– c-means;

– k-means;

– hierarchical cauterization.

Виділимо три структурних рівня ієрархічного аналізу методів кластеризації:

– Рівень 0. Вибір методу розв'язання;

– Рівень 1. Критерії порівняння методів;

– Рівень 2. Множини альтернатив.

На підставі сформованих критеріїв вибору методу кластеризації побудуємо матрицю попарних порівнянь [1]. Попарне порівняння, як правило, - це будь-який процес порівняння сутностей у парах, щоб визначити, яка з кожної сутності є кращою, чи має більший обсяг якоїсь кількісної властивості, чи однакові ці дві сутності. Метод попарного порівняння використовується в науковому дослідженні уподобань, установок, систем голосування, соціального вибору, громадського вибору, розробки вимог та мультіагентних систем штучного інтелекту. У психологічній літературі її часто називають парним порівнянням.

Дана матриця дозволить математично порівняти критерії методів кластеризації. На підставі переваг одного критерію від іншого задається число яке означає важливість (або вагу). Далі на підставі цих чисел шукають відношення відповідності, його індекси, вектори локальних пріоритетів. Наприклад індекс і ставлення відповідності дозволяють обчислити ступінь узгодженості результатів.

Коректними вважаються відносини зі значеннями які <= 0.2. Якщо значення більше то це можна зробити виводь що критерії порівняння методів кластеризації не є правильними, і їх треба переглянути.

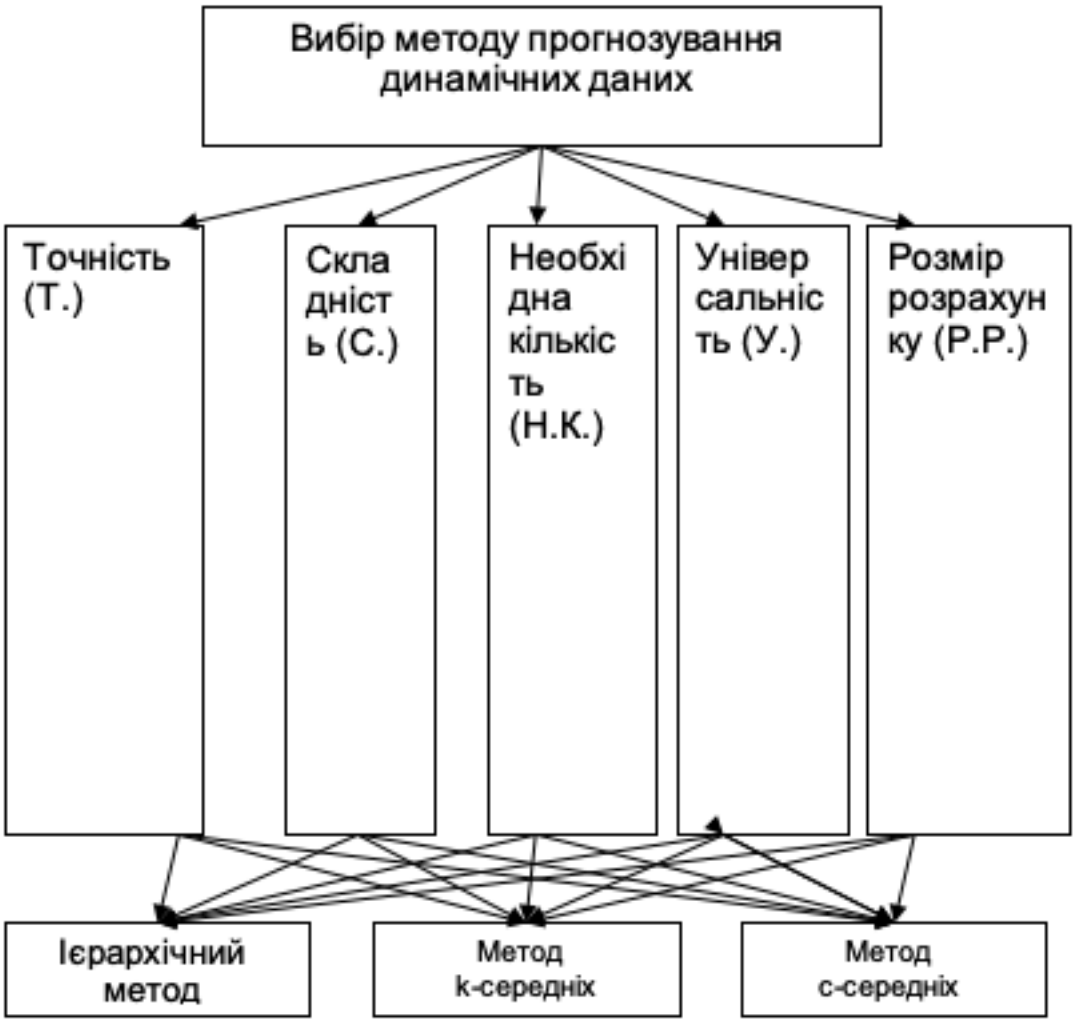

Рисунок 1.1 – Ієрархічна структура пріоритетів

Згідно до цієї ієрархічної структури пріоритетів побудуємо матрицю попарних порівнянь першого рівня.

Таблиця 1.1 – Матриця попарних порівнянь першого рівня

Назва	Точ.	Скл.	Необ. к.	Універ.	Роз. Роз.	Власний вектор	Вектор пріоритетів
Точ.	1	5	2	3	4	2,605	0,428
Скл.	$1/5$	1	$1/2$	$1/3$	2	0,582	0,096
Необ. к.	$1/2$	2	1	$1/2$	2	1	0,164
Універ.	$1/3$	3	2	1	3	1,431	0,235
Роз. Роз.	$1/4$	$1/2$	$1/2$	$1/3$	1	0,461	0,076

Розрахуємо індекс узгодженості (ІУ):

$$\frac{5{,}191 - 5}{5 - 1} = 0.048$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0{,}048}{1{,}12} = 0{,}043$$

З отриманих проміжних результатів можна зробити висновок, що матриця попарних порівнянь заповнена правильно, тобто жодне з тверджень нічому не суперечить. Вектор локальних пріоритетів відносно проблеми вибору набуває вигляду:

$$\vec{p}^{K} = (0{,}428;\ 0{,}096;\ 0{,}164;\ 0{,}235;\ 0{,}076)$$

Відносно кожного з критеріїв слід провести аналіз альтернатив. Матриці попарних порівнянь представлені у вигляді таблиць 1.2 – 1.6. З яких можна побачити якісний аналіз альтернатив та зробити висновки щодо вибору певного методу.

Таблиця 1.2 – Матриця попарних порівнянь 1-го критерію

Назва	c-means	k-means	Hierarchical clustering	Власний вектор	Вектор пріоритетів
c-means	1	2	$^1/_2$	1	0,276
k-means	$^1/_2$	1	$^1/_5$	0,464	0,128
Hierarchical clustering	2	5	1	2,154	0,595

Розрахуємо індекс узгодженості (ІУ):

$$\frac{3{,}0015-3}{3-1} = 0.00075.$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0{,}00075}{0{,}58} = 0.0013.$$

Вектор локальних пріоритетів:

$$\overrightarrow{p_1}^A = (0{,}276;\ 0{,}128;\ 0{,}595;)$$

Таблиця 1.3 – Матриця попарних порівнянь 2-го критерію

Назва	c-means	k-means	Hierarchical clustering	Власний вектор	Вектор пріоритетів
c-means	1	$^1/_2$	2	1	0,276
k-means	2	1	5	2,154	0,595
Hierarchical clustering	$^1/_2$	$^1/_5$	1	0,464	0,128

Розрахуємо індекс узгодженості (ІУ):

$$\frac{3{,}0015 - 3}{3 - 1} = 0.00075$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0{,}00075}{0{,}58} = 0.0013$$

Вектор локальних пріоритетів:

$$\overrightarrow{p_2}^{A} = (0{,}276;\ 0{,}595;\ 0{,}128)$$

Таблиця 1.4 – Матриця попарних порівнянь 3-го критерію

Назва	c-means	k-means	Hierarchical clustering	Власний вектор	Вектор пріоритетів
c-means	1	1	2	1,26	0,4
k-means	1	1	2	1,26	0,4
Hierarchical clustering	$1/2$	$1/2$	1	0,63	0,2

Розрахуємо індекс узгодженості (ІУ):

$$\frac{3 - 3}{3 - 1} = 0$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0}{0{,}58} = 0$$

Вектор локальних пріоритетів:

$$\overrightarrow{p_3^A} = (0{,}4;\ 0{,}4;\ 0{,}2)$$

Таблиця 1.5 – Матриця попарних порівнянь 4-го критерію

Назва	c-means	k-means	Hierarchical clustering	Власний вектор	Вектор пріоритетів
c-means	1	3	$^1/_4$	0,91	0,211
k-means	$^1/_3$	1	$^1/_7$	0,362	0,084
Hierarchical clustering	4	7	1	3,036	0,704

Розрахуємо індекс узгодженості (ІУ):

$$\frac{3{,}029 - 3}{3 - 1} = 0{,}015$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0{,}015}{0{,}58} = 0{,}025$$

Вектор локальних пріоритетів:

$$\overrightarrow{p_4^A} = (0{,}211;\ 0{,}084;\ 0{,}704)$$

Таблиця 1.6 – Матриця попарних порівнянь 5-го критерію

Назва	c-means	k-means	Hierarchical clustering	Власний вектор	Вектор пріоритетів
c-means	1	4	$^1/_3$	1,1	0,322
k-means	$^1/_4$	1	$^1/_2$	0,5	0,146
Hierarchical clustering	3	2	1	1,817	0,532

Розрахуємо індекс узгодженості (ІУ):

$$\frac{3{,}042 - 3}{3 - 1} = 0{,}021$$

Розрахуємо відносну узгодженість (ВУ):

$$\frac{0{,}021}{0{,}58} = 0{,}036$$

Вектор локальних пріоритетів:

$$\overrightarrow{p_5^A} = (0{,}322;\ 0{,}146;\ 0{,}523)$$

На основі отриманих результатів розрахуємо вектор глобальних пріоритетів та відповідні показники. Зазначимо, що в даному випадку індекс узгодженості – це сума індексу першого рівня ієрархії та скалярного

добутку вектору пріоритетів критеріїв з вектором індексів узгодженості відповідного критерію.

Таблиця 1.7 – Кінцеві результати задачі вибору методу розв’язання

Критерії / Альтернативи	Точ.	Склад.	Необ. кіл.	Універ.	Роз. Роз.	Вектор пріоритетів
c-means	0,276	0,276	0,4	0,211	0,322	0,283
k-means	0,128	0,595	0,4	0,084	0,146	0,2
Hierarchical clustering	0,595	0,128	0,2	0,704	0,532	0,76

Проаналізувавши дані з таблиці 1.7 можна побачити що все відповідає критеріям узгодженості. Максимальна компонента вектору глобальних пріоритетів знаходиться у третій альтернативі, що є підтвердженням що метод правильний.

Наступним етапом після визначення методу буде аналіз проблем кластеризації. Для того щоб провести аналіз проблем кластеризації потрібно сформулювати бажані, критичні і небажані властивості в процесі її функціонування. Для даної системи головним небажаних властивістю є велика вимога програми до ресурсів.

Щоб вирішити цю проблему потрібно кластеризувати критерії даної системи, який мають певні властивості обраного методу моделювання. Як результат виділимо кілька категорій задоволення і не задоволення. Також охарактеризуємо їх:

- бажані властивості. Це точність та невелика похибка прогнозування психологічного стану;
- небажані властивості. Збільшення похибки результатів та витрат на розрахунків;

– важливі властивості. Це складність програмування методу кластеризації та універсальність.

Згідно до цих категорій задоволення і не задоволення опишемо ієрархічну структуру та її рівні:

– Рівень 0. Цілі, у вигляді проблеми незадоволеності;
– Рівень 1. Класифікація незадоволення;
– Рівень 2. Характеристики незадоволень.

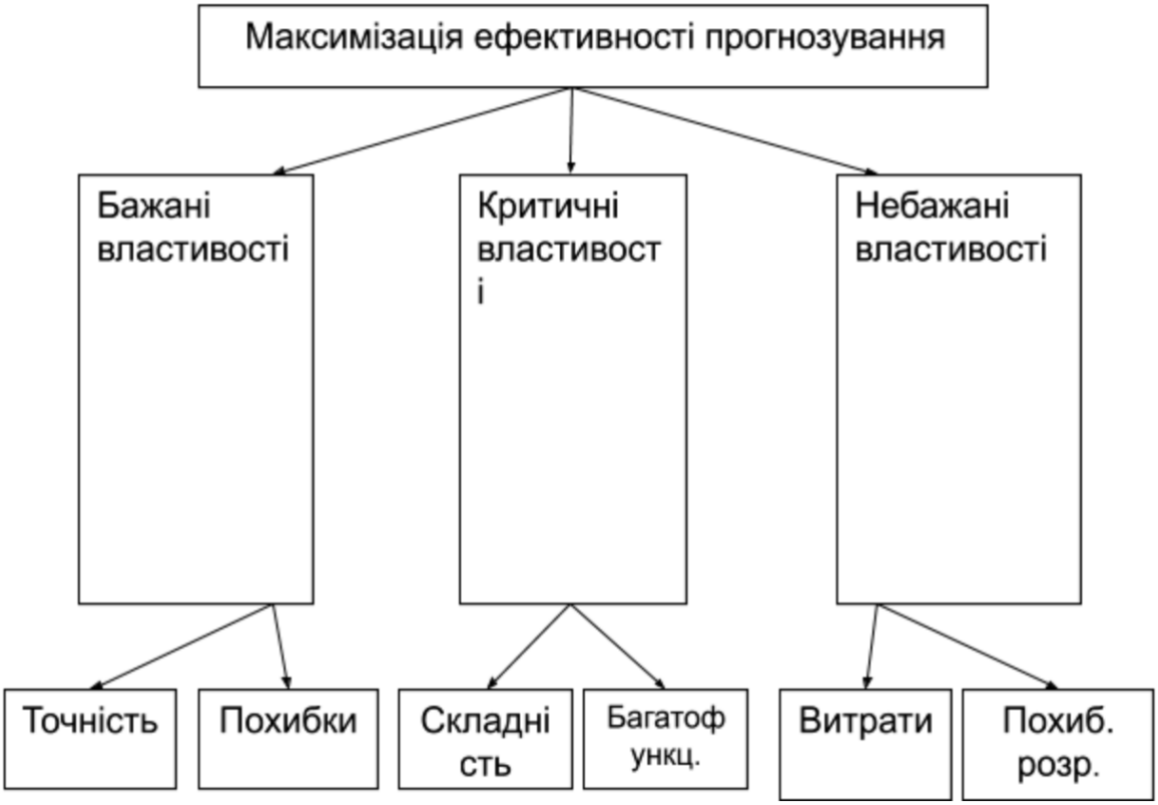

Рисунок 1.2 – Ієрархічна структура категорій задоволення і не задоволення

Згідно до побудованої ієрархічної структури максимізації ефективності прогнозування [2], модель можна вважати закінченою.

1.3 Вектор пріоритетів незадоволення методом аналізу ієрархій

Для того щоб оцінити глобальний вектор пріоритетів потрібно створити матрицю попарних порівнянь 1-го рівня. Теж саме потрібно зробити і для кожної виділеної характеристики.

Таблиця 1.8 – Матриця попарних порівнянь 1-го рівня

Назва	Бажані в.	Критичні в.	Небажані в.	Власний вектор	Вектор пріоритетів
Бажані в.	1	4	2	2	0,558
Критичні в.	$1/4$	1	$1/3$	0,437	0,122
Небажані в.	$1/2$	3	1	1,145	0,319

Згідно до цієї матриці попарних порівнянь першого рівня побудуємо матриці попарних порівнянь для бажаних властивостей, для критичних властивостей та небажаних властивостей.

Таблиця 1.9 – Матриця попарних порівнянь бажаних властивостей

Бажані в.	Точ.	Пох. прог.	Власний вектор	Вектор пріоритетів
Точ.	1	3	1,732	0,75
Пох. прог.	$1/3$	1	0,577	0,245

Таблиця 1.10 – Матриця попарних порівнянь критичних властивостей

Критичні властивості	Скл.	Багатофун.	Власний вектор	Вектор пріоритетів
Скл.	1	$1/5$	0,447	0,166
Багатофун.	5	1	2,236	0,833

Таблиця 1.11 – Матриця попарних порівнянь небажаних властивостей

Небажані властивості	Витрати	Похиб.рез	Власний вектор	Вектор пріоритетів
Витрати	1	$1/3$	1,732	0,75
Похиб.рез	3	1	0,577	0,25

Переконавшись, що дані матриці попарних порівнянь вірні, далі треба знайти значення вектору глобальних пріоритетів (рисунок 1.3).

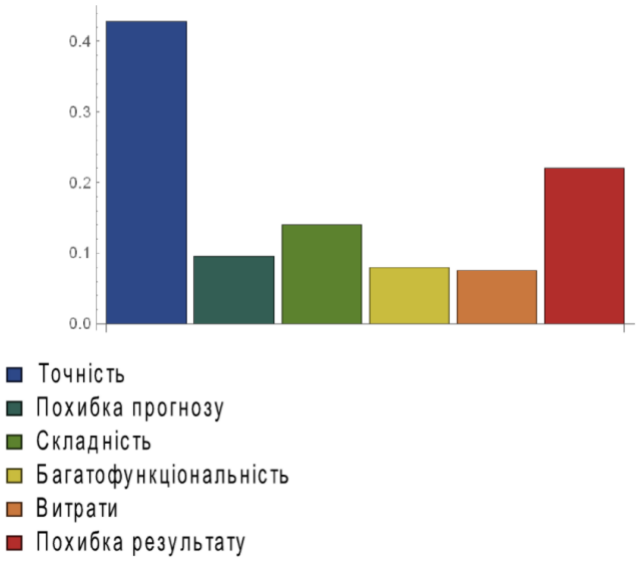

Рисунок 1.3 – Діаграма глобальних пріоритетів

З наведених вище результатів можна зробити висновок що найбільш найважливішими критеріями є точність результатів, зменшення складність програми і зменшення похибки результату. Отже всі зусилля будуть сконцентровані на вирішенні цих завдань.

1.4 Загальна постановка задачі

Дуже важливу роль відіграє в кластеризації даних відіграє вибір алгоритму який вибирається за певними характеристиками вхідних даних. За допомогою аналізу вхідних даних підбираються параметри моделі. Далі за допомогою методу порівняння моделі ці варіанти порівнюються між собою.

Програма працює з даними які показують емоційний стан людини, тому ці дані представлені у вигляді набору числових параметрів. Однак вхідні дані можуть бути не однорідними. В такому випадку виникають різні труднощі в кластеризації даних. Отже можна використовувати методи аналізу і обробки даних для їх перекладу з складного, можливо не структурованого стану в числовий вигляд. Слід виділити наступні кроки для вирішення цього завдання:

- реалізувати програму для отримання та нормалізації даних.
- перетворити текстову інформацію до нормалізованого виду (цифрового);
- реалізувати програму для кластеризації та візуалізації;
- кластеризувати дані.

1.5 Постановка задач дослідження

На підставі проведеного системного аналізу можна сформулювати послідовні завдання для даної дипломної роботи:

- згідно до бажаних характеристик методу кластеризації будуємо модель попарних порівнянь;
- аналізуємо існуючі алгоритми кластеризації;
- обираємо найкращий метод кластеризації згідно до моделі;
- провести кластеризацію використовуючи розроблення програму;
- проаналізувати результати кластеризації;
- зробити висновки.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ'ЯЗАННЯ

2.1 Метод ієрархічної кластеризації

Існує дві основні групи методів кластерного аналізу: зовнішні і внутрішні. Внутрішні виділяються тим, що справедливо, рівнозначно відносяться до ознак кластеризації. Для зовнішніх все навпаки, коли повинен існувати тільки один головний критерій, коли як інші обчислюють його.

Ієрархічні та неієрархічні вдають із себе дві основні групи внутрішніх методів кластеризації [3]. Неієрархічний метод частіше немає ніякої структури, або має але дуже складну. Ієрархічних метод представляє з себе деревоподібний порядок. Так само даний тип кластеризації розділяється на дівозмінні і на агломераційні методи (є найпоширенішими).

Кожен об'єкт слід вважати окремим, самостійним класом, отже для таких кластерів визначимо відстань через функцію (2.1):

$$R(\{x\},\{x'\}) = p(x,x'), \tag{2.1}$$

Всі ітерації алгоритму з пари найближчих один до одного кластерів U та V утворює інший кластер $W = U \cup V$. Відстань від нового кластеру W, до будь-якого іншого кластеру S, розраховується за відстанями $R(U,V),\ R(U,S)\ R(V,S)$, які в цьому моменту повинні бути відомі за формулою (2.2):

$$R(U \cup V,S) = \alpha_{\cup}R(U,S) + \alpha_V R(V,S) + \beta R(U,V) + \gamma R(U,S) - R(V,S), \tag{2.2}$$

Слід зазначити що $\alpha_{\cup}, \alpha_V, \beta, \gamma$ – це цифрові параметри. За допомогою цієї формули можна знайти всі відстані між зазначеними кластерами, де кожна відстань є ребром дерева. Як було визначено, даний метод розраховує відстань між кластерами до їх злиття.

Таким чином за допомогою цього методу кластеризації можна явно подивитися на ієрархію кластерів, та якщо відстань між двома кластерами занадто велика після їх злиття, можна повернутись на попередній крок до злиття цих кластерів [4]. Як результат отримаємо коректну кластеризацію даних. Але треба одразу визначитись з ознаками для цих даних, бо якщо ці ознаки в усіх даних будуть схожі то цей метод кластеризації буде неефективним.

2.2 Метод k-means

Метод k-means – являє собою ітераційний метод кластерного аналізу який поділяє дані наперед відому кількість кластерів. Головна суть полягає у тому щоб з кожною ітерацією знайти найбільш правильний центроїд (центр) у новому кластері та взяти нові дані до кластеру згідного до нового центру. До кластеру потрапляють дані які знаходяться найближче до центру кластеру. Відстань, як правило, розраховується за допомогою формули Евкліда (2.3):

$$p(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2} \quad (2.3),$$

де $x, y \in R^n$ – координати на координатній площині. Слід розглянути ряд спостережень $\left(x^{(1)}, x^{(2)}, \dots, x^m\right), x^{(m)} \in R^n$. Метод k-means розділяє m спостережень (даних) на k кластерів. Та як правило $k \leq m, S = \{S_1, S_2, \dots, S_k\}$. Для мінімізації квадратичного відхилення кластерних точок від їх центрів слід використовувати формулу (2.4):

$$min\left[\sum_{i=1x\in S_i}^{k}\left\|x^{(j)}-\mu_i\right\|^2\right] \quad (2.4),$$

де μ_i – центр кластера S_i. Алгоритм кластеризації k-means починається з вибору кількості кластерів (груп). Це інтуїтивний підхід. Кількість кластерів може залежати від даних і від того як вони відрізняються один від одного. Наприклад якщо ознаки даних були вибрані не правильно (або інші не можна було вибрати) то дані будуть знаходитися дуже близько один до одного, дуже згруповані разом. У таких випадках алгоритм k-means неефективний.

Коли кількість кластерів визначена, з самого початку роботи алгоритму k-means, випадково вибираються центри для кластерів. Але існує інша форма цього алгоритму. Наприклад у так званому “k-means++” на першому етапі центроїди вибираються згідно максимальної початкової відстані між усіма кластерами. Таким чином ми отримаємо більш точну та швидку роботу алгоритму так як можемо зменшити кількість ітерацій алгоритму.

Центр ваги кожного кластера з кожною ітерацією обчислюють згідно формулі (2.5):

$$\mu_i=\frac{1}{S_{ix^{(j)}\in S_i}}\sum x^{(j)} \quad (2.5)$$

Отже, головною задачею алгоритму k-means є перерахунок центрів кожного кластера для кожної ітерації. Кількість ітерацій і кількість кластерів треба задати самостійно на початку роботи алгоритму. Але існує варіант коли можна не задавати кількість ітерацій. Згідно цього алгоритм треба зупинити або закінчити автоматично коли центроїди кластерів не змінюється.

Слід зазначити, що головним недоліком алгоритму k-means є те, що на самому початку його роботи треба самостійно визначити кількість кластерів. Кількість кластерів визначається згідно розташування даних на координатній плоскості. Але існують специфічні випадки коли датасети не формують явні

кластерні групи на координатній плоскості. Через це, людина може неправильно визначити кількість кластерів. Для цих специфічних випадках розташувань даних слід використовувати більш правильні методи кластеризації.

2.3 Метод c-means

Існує специфічний випадок, котрий k-means не здатний вирішити. Як було зазначено вище, k-means закінчує свою роботу якщо центри кластерів не змінюються або кількість заданих ітерацій була перевищена. При першому випадку алгоритм k-means може ніколи не закінчити свою роботу, тому що дані можуть знаходитися на перетині двох кластерів. Таким чином з кожною ітерацією такі дані будуть переходити будуть переходити з одного кластеру до іншого та алгоритм ніколи не зупиниться.

Для вирішення цієї проблеми треба використовувати метод нечіткої кластеризації даних. За допомогою c-means методу стає можливим визначити з якою ймовірністю дані відносяться до кластеру.

У рамках виконання нечіткої (ймовірнісної) c-means кластеризації вирішується проблема мінімізації через функцію (2.6):

$$E = \sum \sum u_{ij}^{m} \cdot \left\| x_i - c_j \right\|^2 \quad (2.6),$$

з обмеженнями $\sum_j u_{ij} = 1, j = 1 \dots p.$

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Семантичний словник

За допомогою семантичного словника AFINN запроваджується семантичний аналіз тексту користувача. Словник представляє з себе один файл де навпроти кожного слова є семантична оцінка. Кожне слово оцінюється від -5 до 5 балів. Негативні слова мають оцінку зі знаком мінус. Нейтральні слова оцінюються як 0. Усі інші слова являють з себе позитивні.

Для семантичного словника AFINN існують бібліотеки для всіх популярних мов програмування які представляють з себе API для роботи з ним. Це спрощує та зменшує складність роботи. Наприклад, всі слова в AFINN знаходяться в нижньому регістрі, а API бібліотека буде автоматично приводити усі слова до цього формату. Також не потрібно писати парсер для словника AFINN.

З кожним роком з'являються нові слова та вирази, тому існує декілька версій цього словника. Все починалось з AFINN-96 який був створений у 2009 році і мав 1468 слів. У 2011 був створений AFINN-111 який налічує 2477 слів. А найостанніша версія словника була створена у 2015 році під назвою AFINN-165, та налічує 3382 слів.

Треба звернути увагу на те, що люди висловлюють емоції не тільки словами, але і смайликами emoji. Існує open-source проект котрий забезпечує семантичний аналіз цих допоміжних речей тексту. Оцінюються смайли такі як і в семантичному словнику AFINN від -5 до 5.

3.2 Дані для семантичного аналізу

Для роботи з великим обсягом даних, як джерело цих даних, була обрана соціальна мережа Twitter. Ця соціальна мережа має спеціальне API для роботи інших програмних продуктів [5].

Щоб почати роботу з соціальною мережею Twitter треба створити спеціальний акаунт розробника. Цей акаунт повинен пройти верифікацію від компанії Twitter. Після того як акаунт був верифікований треба зареєструвати програмний застосунок та створити спеціальні ключ (так званий bearer-token що використовується у технології OAuth 2) для доступу програмного застосунку до Twitter API.

Після того як акаунт розробника був створений та програмний додаток получив ідентифікаційний ключ, слід вибрати підписку.

Підписка “Enterprise API” є платною. Але має масу переваг в порівнянні зі стандартною безкоштовно:

- виконувати складні пошукові запроси;
- не дуже обмежена за кількістю повідомлень за один пошуковий запит;
- дозволяє зчитувати повідомлення в batch режимі;
- фільтрація повідомлень;
- доступ до всього архіву повідомлень;
- доступ до метаданих повідомлень таких як розширені URL-адреси та геодані профілю;
- більш швидша робота з API (в порівнянні до стандартної підписки);
- дозволяє розраховувати тимчасові інтервали повідомлень;
- доступ до метрик у dashboards для розробників.

Варто зазначити що соціальна мережа Twitter виконує архівацію своїх повідомлень якщо вони застаріли більше ніж на 30 днів. Це такий процес

оптимізації витрат пам'яті через великий обсяги інформації. Саме з цією причино запит на такі старі дані обійдеться дорожче, тому що витрачає більше обчислювальних ресурсів. Для цього існує дві цінові політики. Якщо потрібно отримувати тільки найсвіжіші повідомлення то підписка обійдеться в 150 доларів за 500 запитів, якщо потрібно буде отримувати повідомлення які старіше 30 днів то ціна буде 100 доларів за 150 запитів.

Підписка “Standard API” є повністю безкоштовною і буде використана так як її функціоналу достатньо для вимог даної дипломної роботи. Варто виділити наступні корисні функції даної підписки:

– фільтрація повідомлень;

– пошук повідомлень що не старіше 7 днів.

Для роботи з Standard API буде використовуватися C# бібліотека TweetSharp. Вона є повністю безкоштовною, open-source. За допомогою цієї бібліотеки можливо шукати користувачів та їх інформацію (повідомлення) яку вони оставили на своїй сторінці.

3.3 Розробка програми для семантичного аналізу

За допомогою мови програмування C# буде виконуватись семантичний аналіз повідомлень користувачів Twitter. Ця мова програмування має широкий спектр різних бібліотек, які допоможуть при написанні магістерської роботи. Головною перевагою вибору цієї мови є можливість швидко створювати desktop програмні застосунки за допомогою Windows Forms.

C# – це об'єктно-орієнтована мова програмування, яка також включає функціональну та імперативну парадигму програмування. Працює в віртуальній машині. Є прямим конкурентом мови програмування Java.

Windows Forms – це бібліотека для створення графічних форм. За допомогою цієї бібліотеки стає можливим створювати віконні форми та

елементи що можуть управляти програмою. Також є елементи для візуалізації графів що будуть використовуватись при семантичному аналізі.

3.4 Розробка програми для кластеризації та її візуалізації

Для допомогою мови програмування R буде здійснена кластеризація та візуалізація результатів [6]. Це спеціальна мова програмування орієнтована на статистичні обчислення, математичний аналіз та візуалізацію результатів обчислень. Більшість професійних математиків, статистів, та навіть у медичних університетах використовують цю мову програмування для побудови математичних моделей [7]. Як і мова програмування Fortran, у мові R, є дуже велика кількість безкоштовних бібліотек для роботи з даними.

Конкурентними аналогами для цих задач є Fortran, SAS Analytics, StatSoft STATISTICA, Minitab, IBM SPSS. Більшість з цих інструментів не потребують навиків програмування, всі налаштування виконуються мануально. Мову програмування Fortran слід вважати застарілою. Її використання не є перспективним. Математики її використовують за деякою інерцією, тому що навіть зараз, більшість компьютеризированных математичних робот існує на мові Fortran.

3.5 Опис готової програми

При запуску програми відкривається головна форма (рис. 3.1). На цій формі можна побачити кілька кнопок таких як “AddNew” – додати нового користувача, “Show Full Info” – показати всю інформацію цього користувача і кнопка “Save” – зберегти всіх користувачів в excel форматі.

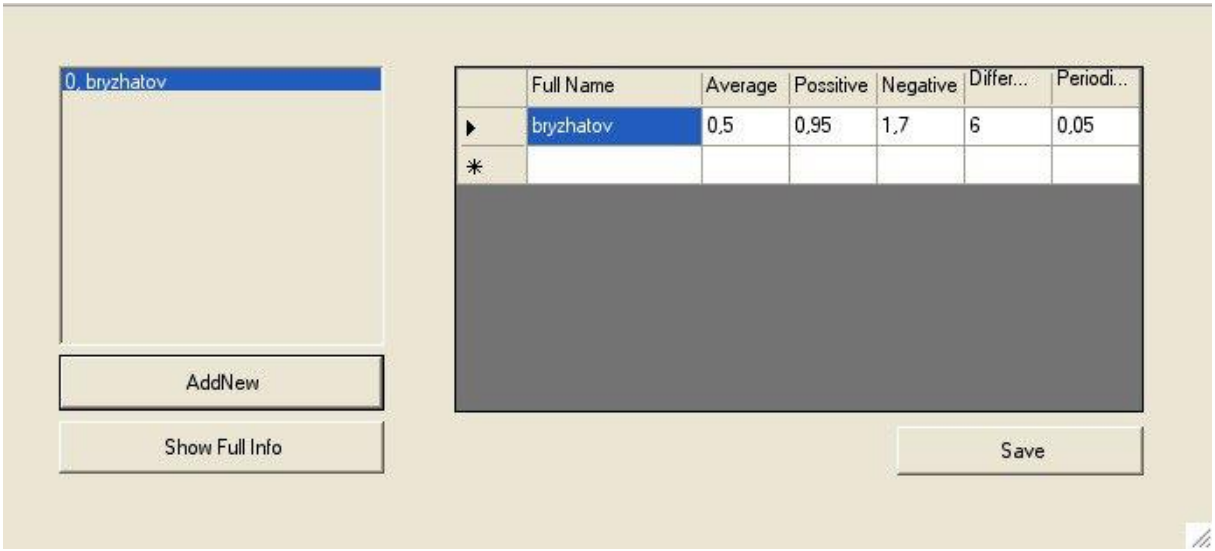

Рисунок 3.1 – Головне вікно програми

Для того щоб приступити до семантичного аналізу користувачів, слід додати користувачів до програмного додатку. Це можна зробити якщо натиснути на кнопку “AddNew”. Далі з'явиться нова форма додавання користувача (рис. 3.2). Щоб додати нового користувача треба ввести у поле “Full Name” його ім'я, у поле “Birthday” його день народження, у поле “Twitter Nick” його логін при реєстрації.

У поле “Twitter Id” буде автоматично знайдено ідентифікаційний ключ користувача Twitter завдяки API бібліотеці TweetSharp.

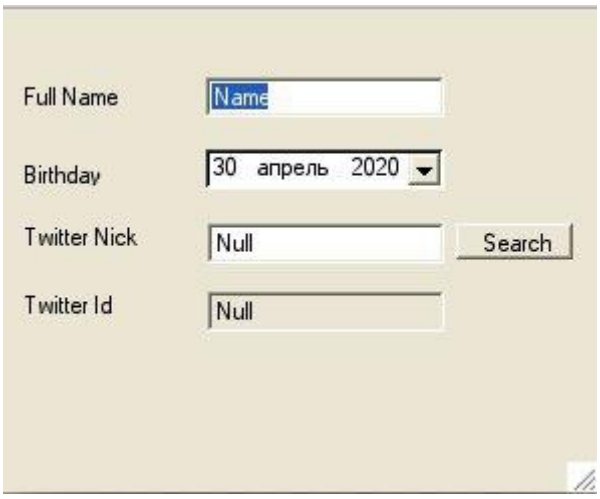

Рисунок 3.2 – Форма додавання користувача

Якщо користувач не був знайдений, то з'явиться вікно помилки (рис. 3.3).

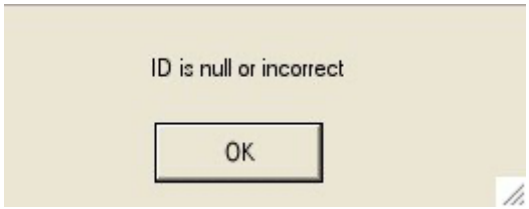

Рисунок 3.3 – Форма додавання користувача

Як тільки користувач додається до програми, починається семантичний аналіз його Twitter сторінки. Аналізуються всі повідомлення з початку реєстрації користувача у соціальну мережу. Завершеним аналіз можна враховувати якщо дані як на рисунку 3.1 не змінюються деякий час.

Якщо на головній формі вибрати будь-якого користувача та натиснути “Show Full Info”, відкривається форма з більш детальною інформацією про нього (рис. 3.4).

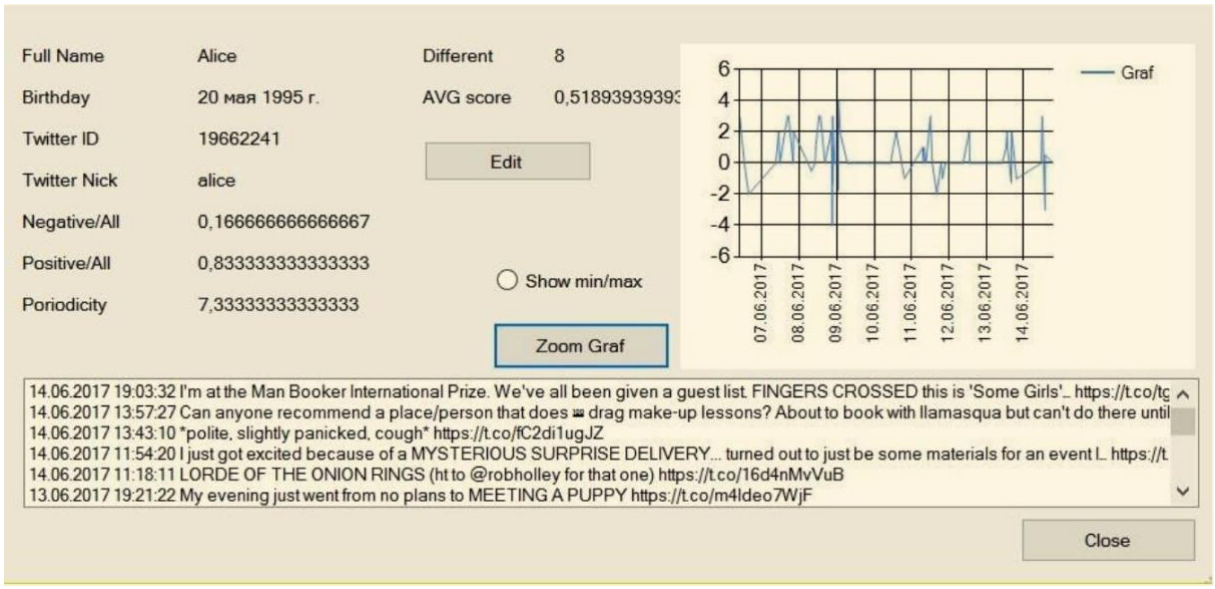

Рисунок 3.4 – Повна інформація про конкретного користувача

На цій формі можна побачити повне ім’я, день народження, ідентифікаційний ключ, логін. Нижче є форма де відображається дата створення, час та текст повідомлення. У правій частині форми є графік який показує перепади настрою. На осі X можна побачити дати створення повідомлень, на осі Y бачимо шкалу від -6 до 6. Так як бібліотека AFINN розрахована на числа від -5 до 5, треба було створити деяке порожнє простір щоб графік настрою не торкався границь форми. Слід зазначити що такі поля як “Negative/All”, “Positive/All”, “Periodicity”, “Different” та “Average” будуть розглянуті далі. Повертаючись до головної форми (рис. 3.1), після проведеного семантичного аналізу слід натиснути на кнопку “Save”.

Після чого статистичні дані що є в формі будуть завантажені у excel файл у виді таблиці (рис. 3.5).

№	Full Name	Common Average	Positive %	Negative %	Difference	Periodicity
1	Paul	0.57	0.79	0.21	4	2.3
2	Alice	0.4964	0.845	0.1549	8	7.1
3	Anton	-0.2941	0.7647	0.2352	5.5	4.25
4	Marry	1	1	0	0	1
5	Kirill	2.5	1	0	1	1.5
6	Elizabet	NaN	NaN	NaN	NaN	NaN
7	Dmitry	0.5438	0.7926	0.2073	9	4.68
8	Dmitry	0.8151	0.983	0.02	7	2.5
9	Boris	0.7096	0.8115	0.1846	7	4.64
10	Illya	0.9731	0.8309	0.169	9	10.1
11	Alex	0.2043	0.798	0.2019	7	37
12	Vadim	NaN	NaN	NaN	NaN	NaN
13	Igor	0.3	1	0	0	2
14	Daria	NaN	NaN	NaN	NaN	NaN

Рисунок 3.5 – Excel файл з семантично проаналізованими користувачами

Бібліотека AFINN представляє з себе лише семантичний словник, але повідомлення користувача це фрази, набір слів [8]. Саме тому, коли обробляється будь яке повідомлення користувача, воно парситься, далі семантичного аналізується кожне слово. Середнє арифметичне оцінок цих слів у повідомленні і є “Message Average”. Тобто це середня емоційна навантаження цього повідомлення. Але не можна судити о користувачеві по одному повідомленню, тому “Message Average” кожного повідомлення підсумовується та ділиться на їх кількість, тобто знаходиться “Common Average” серед інших “Message Average” окремих повідомлень користувача.

Значення для поля “Difference” беруться з різниці самого максимального та мінімального “Message Average”. За цим полем визначається емоційна стабільність користувача за деякий період часу.

Поля “Positive” та “Negative” відображають процентне відношення негативних та позитивних слів користувача Twitter. Якщо скласти ці числа можна получити 100 відсотків, тобто всі слова користувача.

Поле “Periodicity” відображає середню кількість повідомлень яку робить користувач Twitter у день. Це поле розраховується як відношення всієї кількості повідомлень користувача та кількості всіх днів коли він їх писав.

Далі дані у цьому файлі треба кластеризувати та візуалізувати за допомогою мови програмування R (рис. 3.6).

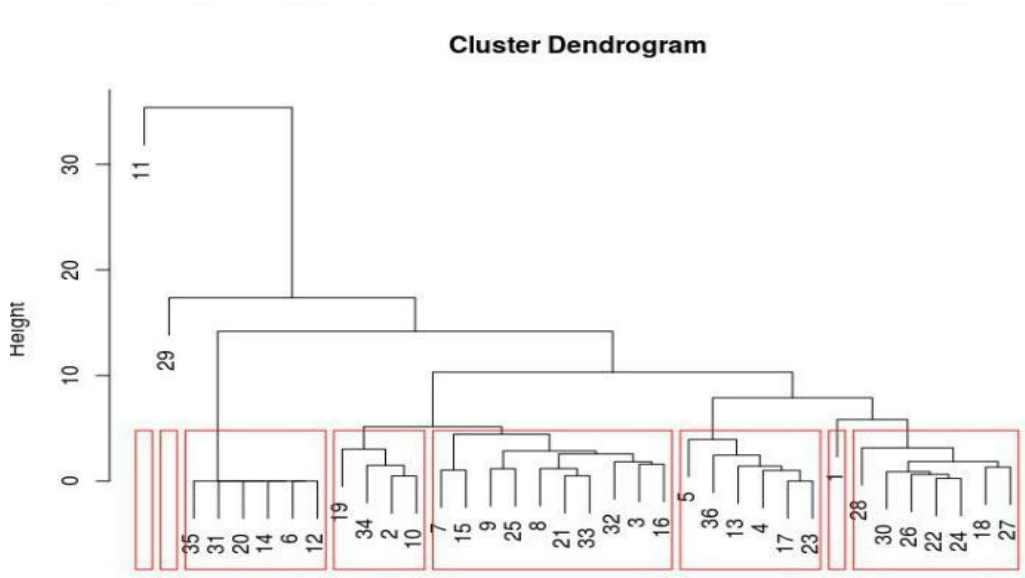

Рисунок 3.6 – Кластеризований excel файл методом ієрархічної кластеризації

Ієрархічна кластеризація виконувалась за такими полями як “Common Average” (середня емоційна навантаження користувача серед усіх його повідомлень), поле “Difference” (різниця або дельта між максимальним та мінімальним значенням “Common Average”) та поле “Periodicity” яке визначає середню кількість повідомлень на день.

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

За результатами ієрархічної кластеризації можна побачити дуже різноманітну структуру. По перше слід звернути увагу що користувачі під номером 35, 31, 20, 6, 12 та 14 на рисунку 3.6. Вони мають в усіх колонках значення “NaN”, тобто “Not a Number”. Це означає що ці Twitter сторінки не використовуються по призначенню. Вони не мають даних для аналізу. Саме тому на дендограмі можна побачити як всі ці користувачі опинилися в одному кластері.

Подивимось на рисунок 3.6, там знайдемо користувачів де в полі “Common Average” задані числа від 0 до 1, а в “Periodicity” можна побачити дуже невелику кількість повідомлень за день. Це означає що ці користувачі не є активними користувачами соціальної мережі Twitter. На дендограмі це користувачі під номером 5, 36, 13, 4, 17, 23.

Користувачі 19, 34, 2 та 10 мають велику кількість повідомлень на день (“Periodicity”). Великою будемо вважати значення, які більш за всього виділяються у інших користувачів, тобто це є числа які більше 7. Також ці користувачі мають велике значення у полі “Difference”. Для нього великим значенням також будемо вважати числа від 7. Отже, за результатами цих двох показників (частоти повідомлень на день та різниці між найбільшим та найменшим значенням поля “Common Average”) можна створити деяке припущення що ці користувачі емоційно нестабільні та потенційно можуть нашкодити соціуму.

Користувачів 7, 15, 9, 25, 8, 21, 33, 32, 3, 16 можна вважати нормальними, але в них знайдені деякі аномалії. Наприклад користувач 7 та 9 має завищений “Difference”, але “Common Average” та “Periodicity” досить нормальні. Мабуть цей користувач мав нещодавно короткостроковий стрес. Згідно інших показників, він соціально безпечний. Користувач 8 дуже позитивний,

семантичний аналіз визначив 98 відсотків його слів як позитивні. Також ця людина має завищений “Difference” та відноситься до такого ж случаю як користувачі 7 та 9. Користувач 3 має невеликий, негативний “Common Average”, тобто майже всі його повідомлення мають досить нейтральний або декілька негативний характер. Мабуть це його нормальна поведінка, бо його “Difference” стабільний, не завищений. Згідно “Periodicity” він не несе загрози соціуму.

Користувачі під номером 5, 36, 13, 4, 17, 23 мають у полі “Common Average” задані числа від 0 до 1, а в “Periodicity” можна побачити дуже невелику кількість повідомлень за день. Це означає що ці користувачі не є активними користувачами соціальної мережі Twitter.

Користувачі 1, 28, 30, 26, 22, 24, 18, 27 є звичайними користувачами соціальної мережі Twitter.

Отже, за результатами ієрархічної кластеризації сформувалися такі групи акаунтів як:

- соціальні (групи, блогери, інформаційні акаунти). Акаунти 11 та 29;
- не мають жодних повідомлень. Акаунт 6, 12, 14, 20, 31 та 35;
- емоційну нестабільні. Акаунти 2, 10, 19 та 34;
- звичайні, але з аномаліями (мають один або декілька аномально);
- завищених або зважених показників. Акаунт 3, 7, 8, 9, 15, 16, 21, 25, 32 та 33.
- звичайні, але не дуже активні, мають невелику кількість повідомлень. Аккаунты 4, 5, 13, 17, 23 та 36;
- звичайні. Акаунты 1, 18, 22, 24, 26, 27, 28 та 30.

Кластеризацію за такими ознаками як “Common Average” (середня емоційне навантаження користувача серед усіх його повідомлень), поле “Difference” (різниця або дельта між максимальним та мінімальним значенням “Common Average”) та поле “Periodicity” (визначає середню кількість повідомлень на день) можна вважати успішною, так як сформувалися досить чіткі кластери даних [9]. Виникає питання до соціальних акаунтів. Можна

побачити що 11 та 29 акаунт не знаходиться в одному кластері тому що дані в них різняться за періодичністю. Детерменізацію соціальних акаунтов можна вирішити додаванням ще однієї ознаки для кластеризації як “Followers Number”, тобто кількість послідовників цього акаунту.

5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Аналіз стану користувача може бути корисний для виявлення психологічно хворих людей, які мають різні відхилення або тимчасово агресивно налаштовані. По-перше це дозволило б в захистити таких користувачів від решти спільноти. По-друге вжити заходів щоб допомогти цій людині. Однак потрібно розуміти яка інформація аналізується, тому що вторгнення в особисте життя може викликати величезні громадські протести. Однак якщо аналізувати лише загальну інформацію, яку користувач сам показує, то такий семантичний аналіз може не відображати реальної, повної картини психологічного стану користувача.

Почати його застосування можна з громадських ресурсів, таких як форуми або сайти на яких люди висловлюють свої думки і ці думки загальнодоступні. Чи застосовувати семантичний аналіз для приватних повідомлень? Це дуже неоднозначне питання. Однак, теоретично, такий інструмент був би дуже ефективний для спецслужб. Особливо якщо зібрати всі соціальні мережі користувача в єдине джерело. Сам по собі семантичний аналіз може показати наміри користувача. Однак потрібні спеціальні словники і навчена нейронна мережа яка показувала б ймовірність виконання якої-небудь дії. Можливо слід використовувати розглянутий алгоритм кластеризації “c-means”, який показує на скільки відсотків той чи інший елемент відноситься до кластеру. Також потрібно буде ретельно продумати ознаки кластеризації. На жаль все це буде не можливо якщо не відстежувати особисту переписку користувача.

Також семантичний аналіз може застосовуватися локально. Наприклад батьки могли би його застосувати для своїх дітей. Це дозволило б набагато знизити відсоток суїциду серед молодих людей, допомогти їм зберегти життя. Особливо це актуально коли існують такі ігри як “Синій кит”. Також семантичний аналіз допоможе психологу краще зрозуміти природу

психологічного відхилення дитини і надати правильну допомогу. На жаль і в цьому випадку виникає етична проблема стеження, навіть за своїми дітьми, це може викликати деякі проблеми у стосунках.

Така система також була б корисною при прийомі на роботу на спеціальні органи державної безпеки. Проаналізувавши потенційного працівника можна було б зрозуміти наскільки він психологічно стійкий та може виконувати свої обов'язки.

ВИСНОВКИ

Через складність людської природи (наприклад жартів, сарказму або провокативних дій щоб звернути на себе увагу) звичайний семантичний аналіз, напевно, ніколи не зможе з високою ймовірністю стверджувати чи є людина соціально небезпечною, чи ні. Однак семантичний аналіз та кластеризація допоможуть звернути увагу на деякі відхилення в поведінці користувача. Наступним етапом розвитку даної магістерської роботи слід було б вважати створення нейронної мережі для аналізу отриманих кластерів [10]. Наданий аналіз обчислювального експерименту вдає із себе приклад людського аналізу і може бути некоректним, суб'єктивним, емоційним. Через великий обсяг даних і швидкість роботи – людський ресурс не повинен розглядатися як варіант аналізу цих кластерів.

Для кластеризації семантичного стану користувача використовується метод ієрархічної кластеризації. На відміну від “k-means” кластеризації, або “c-means”, цей метод не вимагає на початку своєї роботи задавати кількість очікуваних кластерів [11]. Це дуже важлива властивість цього методу, так як неможливо передбачити кількість кластерів для задачі семантичного аналізу користувачів соціальної мережі. Також метод ієрархічної кластеризації при об'єднанні кластерів відображає їх відстань, тобто показує наскільки ці кластери схожі один з одним. Ця відстань відображається через висоту ребра. Це дозволяє правильно виділити отримані кластери, що збільшує точність семантичного аналізу.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Wenhong Tian, Yong Zhao. Optimized Cloud Resource Management and Scheduling: Theories and Practices. – Morgan Kaufman, 2014. – 284 p.

2. Ирина Чубакова. Data Mining. Методы классификации и прогнозирования. URL: http://www.intuit.ru/studies/courses/6/6/lecture/174 (дата звернення: 05.04.2020).

3. Ulrich Meyer, Peter Sanders. Algorithms for Memory Hierarchies: Advanced Lectures. – Springer Science & Business Media, 2003. – 428 p.

4. Seetha Hari, Murty Narasimha, B.K. Tripathy. Modern Technologies for Big Data Classification and Clustering. – IGI Global, 2015 – 360 p.

5. Рассел Метью, Классен Михайло. Data mining. Извлечение информации из Facebook, Twitter, LinkedIn, Instagram, GitHub. – Питер, 2020. – 464 с.

6. Hadley Wickham, Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. – O'Reilly Media, 2017. – 520 p.

7. Бокс Дж. Дженкинс Г. Анализ временных рядов: прогноз и управление. – Мир, 1974. – 405 с.

8. Cliff Goddard. Semantic Analysis: A Practical Introduction. – OUP Oxford, 1998. – 428 p.

9. Джоел Грас. Data Science. Наука о данных с нуля. – БХВ-Петербург, 2016. – 336 с.

10. Шакла Ниш. Машинное обучение и TensorFlow. – Питер, 2019. – 336 с.

11. Гвоздинский А.Н., Губин В.А., Юрдига Л. А. Кластеризация слабоструктурированных текстовых документов // ХНУРЭ: электронная версия научно-технического журнала 2011. № 1 URL: http://openarchive.nure.ua/bitstream/document/1784/1/RI_2011_1-044-047.pdf (дата звернення: 17.05.2020).