Міністерство освіти і науки України

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

СЕМЕСТРОВИЙ ПРОЕКТ

з дисципліни «Проектування програмного забезпечення»

на тему: «Модуль аналізу тональності текстів на українській мові»

Виконали:

студенти 2-го курсу ФІОТ

групи ІВ-82

Кравченко Катерина

Шкардибарда Іван

Київ-2020

ЗМІСТ

ВСТУП

Думки інших людей завжди були важливою частиною інформації для більшості із нас в процесі прийняття рішень. Задовго до того, як використання Всесвітньої мережі стало невід'ємною частиною повсякденного життя суспільства, люди запитували поради в інших щодо якості тієї чи іншої техніки, послуг тієї чи іншої фірми або якому претенденту їх знайомі віддають перевагу на місцевих виборах.

Дослідження в області аналізу тональності знаходяться на початковій стадії, незважаючи на зростаючі потреби суспільства в аналізі соціальних думок. Окрім того, існує досить багато проблем, з якими стикаються дослідники при розробці методів автоматичного аналізу тональності

Метою дослідження є розробка модулю аналізу тональності текстів . Для досягнення такої мети були вирішені наступні задачі:

- проаналізувати існуючі методи та рішення у галузі визначення тональності текстів;
- розробити програмний продукт, що втілює запропонований метод;

Об'єктом дослідження є методи розпізнавання тональності тексту. Предметом дослідження є методи Наївного Байеса, метод опорних векторів та метод Хе Юлан та Жоу Деу у контексті розпізнавання тональності текстів.

Було запропоновано алгоритм аналізу тональності, що базується на методах Наївного байєса та напівавтоматичного навчання.

В першому розділі розглядається актуальність проблеми та існуючі підходи до її розв'язання, формалізується постановка задачі. Другий розділ присвячений обґрунтуванню методу та розробці програми.

1 ОГЛЯД ЗАДАЧІ АНАЛІЗУ ТОНАЛЬНОСТІ

1.1 Аналіз предметної області

Основна задача сентимент аналізу тексту, що містить висловлення думок, може бути сформульована наступним чином: якщо надано текст, що є суб’єктивним висловлюванням, то за припущення, що висловлювання має єдиний об’єкт, виявити емоційний відтінок тексту як одну з двох полярностей: позитивну чи негативну (sentiment polarity).

Визначення, чи має наданий текст в цілому позитивне чи негативне забарвлення називається “класифікацією полярності настроїв” (sentiment polarity classification) або “класифікацією полярності” (polarity classification). Одним з мінусів даного підходу є те, що емоційну складову документа не завжди можна однозначно визначити, тобто документ може містити як ознаки позитивної оцінки, так і ознаки негативної.

В аналізі тональності тексту вважається, що текстова інформація ділиться на два типи: факти і думки. Ключовим поняттям є визначення думки.

Думки поділяються на два типи:

- проста думка;

- порівняння.

Проста думка містить висловлювання автора про один об'єкт. Вона може бути висловлена прямо: «Останні рішення ВТО просто прекрасні», або неявно: «Після важких і малоприємних реформ економіка почала зростати».

Думка першого типу може бути визначена формально: простою думкою називається кортеж з п'яти елементів (entity, feature, sentiment value, holder, time). В цьому визначенні автор (holder) висловив думку про аспект (feature) об’єкту entity в певний момент часу (time). Зазвичай виділяють два види емоцій (sentiment value): позитивна та негативна, тобто класифікація виконується за двома класами. Іноді додається третій - нейтральна думка.

Другий тип думок – порівняння – можна розділити на три види:

- порівняння аспектів об'єктів на користь одного (non–equal gradable);

- прирівнювання аспектів різних об'єктів (equative);

- перевага одного об'єкта над іншими (superlative).

Порівняння першого типу мають вигляд «аспект об'єкта 1 перевершує в чомусь аспект об'єкта 2», наприклад: «Ціни в інтернет-магазині «А» нижче, ніж в інтернет магазині «В». Другий тип виражає схожість аспектів різних об'єктів, наприклад: «Ціни в інтернет-магазинах «А» та «Б» майже однакові». Прикладом третього типу може слугувати речення «В конкурсі на кращий магазин місяця магазин «А» перемагає магазин «Б».

Думка другого типу визначається як кортеж (Obj1, Obj2, A, holder, time). В даному кортежі Obj1 и Obj2 – множини порівнюваних за аспектом A об'єктів, які автор (holder) порівнює у момент часу time. На відміну від кортежу, який визначає думку першого типу, кортеж думки другого типу не містить прямої оцінки емоцій автора.

В аналізі тональності тексту часто зустрічається термін, пов'язаний з поняттям думки – суб'єктивність. Визначення об'єктивного і суб'єктивного речень наступне:

- об'єктивне речення відображає фактичну інформацію про що–небудь, тоді як суб'єктивне речення виражає чиїсь 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.

Отже, аналіз тональності тексту зазвичай включає в себе наступні основні завдання:

- визначення наявності емоційного забарвлення;

- визначення полярності тексту;

- вилучення аспектів з емоційно забарвленого тексту.

Задача визначення полярності тексту формулюється наступним чином: «визначити, яке емоційне забарвлення тексту, позитивне чи негативне?» Визначення полярності тексту зазвичай розглядається на декількох рівнях:

- на рівні документу. Основною задачею на цьому рівні є класифікація, чи повністю весь документ є відображенням позитивної чи негативної думки. Наприклад, для певної рецензії на товар, система автоматичного аналізу тональності тексту визначає, чи висловлює ця рецензія позивний настрій в цілому. Дана задача носить назву «класифікація полярності настроїв на рівні документу». Такий рівень деталізації передбачає, що кожен документ висловлює думку як єдину сутність. Тому він не може застосовуватись для документів, об'єктами яких є декілька сутностей. Однак в рамках данної задачі, зазвичай документи мають одне чітке емоційне забарвлення, бо коментарі до новин зазвичай є досить короткими емоційними текстами;
- на рівні речення. На цьому рівні об'єктом дослідження є окреме речення. Проводиться аналіз чи висловлює певне речення в цілому позитивну чи негативну думку. Аналіз на даному рівні близько пов'язаний з так званою «класифікацією суб'єктивності» (subjectivity classification), яка вірізняє речення (так звані об'єктивні речення) що висловлюють фактичну інформацію від речень, що висловлюють думки та погляди;
- на рівні сутності та аспекту. Обидва попередніх рівня не включають аналіз того, що саме сподобалося чи не сподобалося власнику думки. Аспектний рівень дозволяє виконати більш детальний аналіз. Замість того, щоб аналізувати лінгвістичні конструкції, аспектний рівень аналізує саму думку. Зазвичай, аналіз думки без аналізу її об'єкту має обмежене використання. Окрім того, визнання важливості аналізу об'єкту думки допомагає глибше зрозуміти проблему аналізу

тональності. Наприклад, речення «Не зважаючи на поганий сервіс, мені все одно сподобався цей ресторан» має позитивний сентимент, але ми не можемо стверджувати, що воно є повністю позитивним. Власне, даний сентимент можна вважати позитивним лише в тому випадку, якщо в якості об’єкту обрано «ресторан». Якщо ж в якості об’єкту обрано «сервіс», то думка є повністю негативною. В багатьох дослідженнях об’єкти думки описуються сутностями та їх аспектами. Тобто, метою аналізу такого рівня є виявлення сентименту об’єкту та його властивостей. На такому рівні аналізу, можна отримати структурований підсумок думок щодо не тільки самого об’єкту, але і його властивостей, що перетворить неструктурований текст в структурований масив даних що може бути використаний для будь-яких типів аналізу.

Найбільш важливими індикаторами сентименту є «слова емоційного забарвлення» (sentiment words). Ці слова зазвичай використовуються для висловлення думки, позитивної чи негативної. Наприклад, «добре», «чудово», «неймовірно» – слова для висловлення позитивної думки, в той час як «погано», «жахливо», «сумно» – слова для висловлення негативної думки. Окрім безпосередньо слів, існують також фрази та ідіоми. Слова та вирази є інструментарієм аналізу тональності з очевидних причин. Набір таких слів та виразів називається «лексичним словником».

Незважаючи на те, що слова та вирази для вираження емоційного забарвлення ж дуже важливими при аналізі тональності, просто використання їх не є досить ефективним. Проблема є комплексною та набагато складнішою. Нижче показані основні проблеми, з якими можна стикнутися при виконанні аналізу тональності за словником:

- позитивне чи негативне слово може приймати протилежний відтінок при використанні в іншій предметній області. Так, наприклад, «У цього фільму передбачуваний сюжет» є

негативною характеристикою, а «У цього коду передбачувана поведінка» є позитивною;

- речення, що містить в собі слово емоційного забарвлення, може мати нейтральний сентимент. Цей феномен виникає зазвичай в декількох типах речень. Питальні речення та умовні речення є двома найважливішими типами, наприклад «Чи не могли в Ви порадити, яка з фотокамер Sony є найкращою?», та «Якщо я знайду дійсно хорошу камеру в цьому магазині, я її обов'язково куплю.» В обох цих реченнях є слова, що виражають позитивний настрій («найкраща», «хороша»), але жодне з цих речень не висловлює позитивну чи негативну думку щодо певної камери;
- речення, що містять сарказм з наявністю чи відсутністю слів емоційного забарвлення є дуже складними для аналізу, наприклад «Який чудовий телефон! Перестав працювати вже за два дні». Ця проблема останнім часом має багато уваги, і навіть з'являються деякі практичні результати.

Всі описані вище проблеми є досить серйозними труднощами для виконання аналізу тональності, що базується на лексичному словнику.

1.2 Класифікація думок під час аналізу тональності

В попередньому підрозділі ми визначили, що існують два типи думок – звичайні думки та порівняльні думки. Окрім такої класифікації, звичайні думки також розділяються на явні (explicit) та неявні (implicit) думки (або думки, що маються на увазі).

Звичайна думка. Звичайна думка дуже часто називається просто «думкою» в літературі, та має два основні під типи.

Явна думка. Явна думка – це думка, що була висловлена безпосередньо щодо самої сутності або аспекту цієї сутності. Наприклад: «Якість зйомки просто чудова».

Неявна думка. Неявна думка – це думка, що виражається неявно щодо сутності або аспекту, на основі ефекту, що має ця сутність на інші сутності. Такий підтип досить часто виникає в медичній області. Наприклад, речення «Після введення препарату я почуваюся гірше» висловлює неявну негативну думку щодо препарату, адже ця сутність «препарат» має негативній ефект на іншу сутність «самопочуття».

Більшість сучасних досліджень спираються на явні думки. Їх аналіз відбувається простіше. Аналізувати ж неявні думки досить складно. Наприклад, в області препаратів та медицині, необхідно знати, чи є наданий ефект бажаним чи небажаним. Наприклад, речення «Оскільки я дуже погано себе почуваю, лікар виписав мені цей препарат» не виражає негативної думки щодо препарату, адже «погано себе почуваю» сталося до моменту прийняття препарату.

Порівняльні думки. Порівняльна думка висловлює відношення подібності або відмінності між однією чи двома сутностями, та визначає, яку саме сутність врешті було обрано власником думки. Наприклад, речення «Кола на смак краще, ніж Пепсі» та «Кола найкраща» висловлюють дві порівняльні думки. Порівняльні думки, як правило, висловлюються в

порівняльні формі прикметника або прислівника, хоча і не завжди (наприклад, «я надаю перевагу»).

Явна думка. Явна думка – це суб'єктивне твердження, що надає звичайну чи порівняльну думку, наприклад: «Кола смакує добре» або «Кола смакує краще, ніж Пепсі»

Неявна думка. Неявна думка – це об'єктивне твердження, що має в собі звичайну чи порівняльну думку. Таке твердження, як правило, виражає бажаний чи небажаний факт, наприклад:

«Я купив цей матрац два тижні тому, і він деформувався», та «Заряд батареї телефонів Nokia кращий, ніж в телефонів Samsung».

Явні думки простіше виявити та класифікувати, ніж неявні. Більшість сучасних досліджень фокусується саме на явних думках. Досить мало досліджень було проведено щодо визначення та класифікації неявних думок.

1.3 Суб’єктивність та емоції

Існують дві концепції, що є дуже тісно пов’язаними з класифікацією емоційного забарвлення думок – суб’єктивність та емоції.

Об’єктивне речення визначає певну фактичну інформацію щодо навколишнього світу, в той час як суб’єктивне твердження висловлює почуття та думки окремої людини.

Прикладом об’єктивного твердження є наступне речення: «iPhone є продуктом компанії Apple». Прикладом суб’єктивного твердження є наступне речення: «Мені подобається iPhone».

Задача, що займається визначенням того, чи має документ суб’єктивне чи об’єктивне твердження носить назву «Класифікація суб’єктивності». Ми повинні мати на увазі наступне:

- суб’єктивне речення може не висловлювати ніякої думки. Наприклад, речення «Я думаю, що він пішов додому» є повністю суб’єктивним, але не висловлює нічого. Речення в попередньому прикладі також є повністю суб’єктивним та не висловлює позитивну, негативну чи нейтральну думку щодо камери або одного з її аспектів;
- об’єктивні речення можуть висловлювати думки на основі стверджування бажаних чи небажаних фактів. Наприклад, наступні два речення, що просто висловлюють факти об’єктивно також і висловлюють негативну думку (неявну негативну думку) щодо певних продуктів: «Ці навушники зламалися через два дні» або «Не дивлячись на те, що комп’ютер новий, він перестав вмикатися через місяць».

Емоції досить тісно пов’язані з висловленням думки. Вони були класифіковані та виділені в певні категорії – любов, здивування, радість, сум, страх. Сила емоційного забарвлення, що присутня в думці, як правило

пов’язана з силою певної емоції, тому можна деколи зустріти не бінарну классифікацію на гарні/погані емоції а классифікацію за типами емоцій, що містить висловлювання. Найчастіше така классифікація відбувається за допомогою словникових методів, адже дуже важко отримати достатньо велику вибірку для того щоб гарно розрізняти декілька видів емоцій у висловлюваннях.

1.4 Підзадачі аналізу тональності

Розглянемо модель сутності. Нехай сутність ei представляється скінченним набором аспектів Ai = {ai1, ai2, …, ain}. Ei може бути виражена будь якім іншим набором виразів сутності {eei1, eei2, …, eein}. Кожен аспект сутності може бути представлений як набір виразів {aei1, aei2, …, aein}.

Також, нехай емоційно забарвлений документ d містить в собі думки щодо набору сутностей {e1, e2, …, en} та їх аспектів від певного набору власників думки в певний період часу.

Нарешті, маючи визначення емоційно забарвленого документу можна визначити основні задачі, що потрібно вирішити в процесі аналізу тональності.

Задача 1. Вилучення сутностей та їх категоризація. Вилучити всі вирази сутностей в емоційно забарвленому документі та виконати категоризацію та розбиття їх на класи сутностей. Кожен клас характеризує окрему сутність.

Задача 2. Вилучення аспектів сутностей та їх категоризація. Вилучити всі вирази сутностей в емоційно забарвленому документі та виконати категоризацію та розбиття їх на класи сутностей. Кожен клас характеризує окрему сутність.

Задача 3. Вилучення власників думки та їх класифікація. Вилучити власників думки та класифікувати їх. Ця задача є подібною до задач 1 та 2.

Задача 4. Вилучення часу та стандартизація. Необхідно вилучити всі часові проміжки, коли були висловлені думки їх власниками та стандартизувати різні формати представлення часу.

Задача 5. Класифікація емоційного забарвлення аспектів. Необхідно визначити, чи є думка щодо аспекту aij позитивною, негативною чи нейтральною, або надати числове значення що визначає думку. Кортеж з п'яти елементів, в вигляді якого ми дали визначення думці, надає дуже добре джерело для інформації, а також основу для генерації як якісних, так і

кількісних підсумків. Загальна форма таких підсумків базується на аспектах і має назву підсумки, що базуються на аспектах.

1.5 Підходи до класифікації тональності

Для виконання безпосередньої класифікації існує ряд методів. Всі ці методи відрізняються за точністю та швидкодією. До найбільш популярних методів відносять методи машинного навчання з учителем та без учителя, методи, основані на словниках та правилах, та ряд інших.

Для побудови системи, що виконує автоматичний аналіз тональності, як правило використовуються методи машинного навчання. Методи машинного навчання без учителя, або навіть semi-supervised, як правильно, дають нижчу точність, ніж методи машинного навчання с учителем.

Методи, основані на словниках та правилах дають непогану точність, але вони є дуже залежними від предметної області. Якщо існує необхідність виконати аналіз в декількох областях, необхідно складати декілька словників. Сам процес складання словника є досить важким, тому ці методи важко застосовувати для автоматичного аналізу, що не залежить від предметної області. Лінгвістичний підхід може надати відносно точні результати, будучи реалізованим для наукових або журнальних статей або інших, граматично вірних текстів. Окрім того, підхід, заснований на правилах, сильно прив'язаний до конкретної мови.

2 РОЗРОБЛЕННЯ ПРОЕКТУ

2.1 Попередня обробка даних

Сучасна теорія аналізу та керування великими даними відокремлює два основні напрямки автоматичного аналізу настроїв – це методи на основі використання лексем і методи машинного навчання.

Перед застосуванням будь-якого з методів вилучення настрою, звичайною є практика попередньої обробки даних. Попередньо оброблені дані дозволяють забезпечити високу якість класифікації тексту і зменшити обчислювальну складність. Типова процедура попередньої обробки включає в себе наступні основні кроки:

- Розмітка за частинами мови. Цей процес дозволяє автоматично визначити кожне слово речення як частину мови: іменник, займенник, прислівник, прикметник, дієслово, вигук і т. д. Мета полягає в тому, щоб витягти зразки тексту на основі аналізу частотних розподілів цих частин у мові.
- Зведення до кореня. Процедура відсікання суфіксів та закінчень від кореня. Кількість різних слів для аналізу зменшується, коли корінь схожих слів, наприклад, таких як «читати». «читач» і «читання» відображаються як одне слово «читати».
- Видалення некорисних слів. Це слова, які несуть в собі сполучну функцію в реченнях, наприклад, прийменники, артиклі і т. д. Немає певного списку таких слів, але деякі пошукові машини, не використовують такі слова як, «є», «в», «який» і «на». Ці слова можуть бути видалені з тексту перед класифікацією, так як вони мають високу частоту появи в тексті, але не впливають на його емоційне навантаження.
- Обробка заперечень. Заперечення відноситься до процесу перетворення настроїв з позитивного на негативний, або з негативного на позитивний.

- Токенізація в N-грами. Токенізація – це процес створення словнику зі слів тексту.

2.2 Дослідження методів класифікації тональності

Лексемно-орієнтований підхід обчислює настрій заданого тексту в залежності від полярності слів або фраз у цьому тексті. Методика розрахунку настрою полягає у наступному: після попередньої обробки тексту, відбувається перевірка маркера кожного слова на його полярність в лексиконі. Якщо слово не знайдено у лексиконі, тоді його полярність вважається нульовою. Після призначення балів полярності W всім словам, що містяться у тексті, остаточна оцінка S настрою тексту розраховується діленням суми балів слів, які задають настрій тексту (крім нульових) на кількість m таких слів:

$$S = \frac{1}{m}\sum_{i=1}^{m} W_i,$$

де W_i – бал полярності і-го слова; m – кількість слів, які задають настрій тексту.

Усереднення балу дозволяє отримати числове значення балу настрою у діапазоні від -1 до 1, де 1 означає сильний позитивний настрій, -1 означає сильний негативний настрій і 0 означає, що текст є нейтральним. Якість класифікації багато в чому залежить від якості словника.

Словники можуть бути створені з використанням різних методів:

- Вручну побудовані словники (простий, але не дуже швидкий метод). Наприклад General Inquire, який складається зі слів суспільствознавчих категорій для контент-аналізу. Ці категорії аналізу контенту намагаються охопити тон, ставлення, зовнішній вигляд.
- Словники з підготовлених даних бувають напівавтоматичними (наприклад, використовують такі ресурси, як WordNet або UNL, або автоматичними, коли словник може бути отриманий автоматично через асоціацію, де оцінка для кожного нового прикметника розраховується з використанням частоти близькості від прикметника до одного або більшої кількості затравочних слів.

Методи машинного навчання для аналізу текстів – це сукупність методів, заснованих на алгоритмах штучного інтелекту, які використовують для навчання дані раніше помічені як позитивні, негативні або нейтральні.

У спрощеному вигляді, задача класифікації текстів може бути описана наступним чином – задано набір маркованих даних:

$$T_{data} = \{(t_i, L_i), \dots (T, n)\},$$

де кожен текст належить до набору даних T і мітка L_i є попередньо встановленим класом всередині групи класів L, мета полягає в тому, щоб побудувати алгоритм навчання, який буде приймати в якості вхідних даних навчальний набір T_{data} і створити модель, яка буде точно класифікувати немарковані тексти t_i у кількості n.

Найпопулярніші алгоритми навчання для класифікації тексту – це метод опорних векторів, наївний класифікатор Баєса, дерева прийняття рішень, метод максимальної ентропії та нейронні мережі.

Метод опорних векторів (SVM) – це метод навчання з учителем, що використовується для бінарної класифікації. Даний алгоритм машинного навчання будує розділяючу поверхню у гіперпросторі з точок (об'єктів вибірки), що лежать між полярними підмножинами, тобто розмежовує класи. Точки побудованої поверхні називаються опорними векторами. Цей класифікатор може змінювати нейронні мережі, але має дуже повільний процес навчання.

Знаходження SVM відповідає опуклій оптимізації. Завдання класифікації, як правило, включає в себе поділ даних на навчальні та текстові набори. Кожен екземпляр в навчальному наборі містить одне «цільове значення» (тобто клас-мітку) і кілька «атрибутів» (функції спостереження за змінними). Метою SVM є вироблення моделі (на основі навчальних даних), яка визначає цільове значення тексту, та побудова оптимальної

розмежувальної гіперплощини. SVM для класифікації використовується, щоб знайти лінійну модель такого вигляду:

$$y(x) = \omega^{\gamma} x + b,$$

де x вхідний вектор, ω і b є параметрами, які можуть бути скориговані для певної моделі, що оцінюється емпіричним шляхом, γ – вектор двоїстих змінних.

Для простої лінійної класифікації завдання полягає в тому, щоб звести до мінімуму функцію помилок, що визначається рівнянням:

$$C \sum_{n=1}^{N} \varepsilon_n + \frac{1}{2} \|\omega\|^2 \rightarrow min,$$

де C – обрана константа;

ω – вектор коефіцієнтів;

ε – параметр для обробки неподільних даних (входів);

n – номер процедури навчання.

Дерева рішень можуть бути адаптовані до практично будь-якого типу даних, тому цей спосіб широко використовується в алгоритмах машинного навчання. При контрольованому машинному навчанні використовується алгоритм, який ділить підготовлені дані на більш дрібні частини, з метою визначення моделі, яка може бути використана для класифікації. Дані потім представляються у вигляді логічних структур, подібних до древовидної, які можуть бути легко зрозумілі без будь-яких статистичних знань. Алгоритм особливо добре підходить для випадків, коли може бути знайдено багато ієрархічних категоріальних відмінностей. Вони побудовані з використанням евристичних алгоритмів, які називають рекурсивним розбиттям. Це, як правило, відомо, як підхід «розділяй і володарюй», оскільки він використовує значення функцій для поділу даних на менші підмножини подібних класів. Структура дерева рішень складається з кореневого вузла, який представляє

собою весь набір даних, рішень вузлів, які виконують обчислення і листових вузлів, які здійснюють класифікацію.

Для того, щоб класифікувати невідомий екземпляр, дані передаються через дерево. На кожному вузлі рішення певної функції, отриманої з вхідних даних, порівнюються з константою, яку було визначено на етапі підготовки. Обчислення, яке відбувається в кожному вузлі рішення, зазвичай, порівнює обрану функцію з цією, заздалегідь заданою, константою, тоді рішення буде ґрунтуватися на функції, створюючи два способи поділу на дереві. Дані будуть, в кінцевому підсумку, проходити через ці вузли рішення до тих пір, доки не досягнуть листового вузла, який представляє собою визначений клас.

В якості найпростішого методу для класифікації тональності тексту використовується *наївний класифікатор Баєса*. У даному класифікаторі використовується теорема Баєса для визначення ймовірності приналежності елемента вибірки до одного з класів при припущенні незалежності ознак. Для підвищення якості класифікації застосовується метод максимальної ентропії. Класифікатор максимальної ентропії є класифікатором ймовірності, який належить до класу експоненціальної моделі. На відміну від наївного класифікатора Баєса, він не припускає, що ознаки умовно незалежні одна від одної. Цей класифікатор засновано на принципі максимальної ентропії усіх моделей, які відповідають даними навчання з найбільш рівномірним розподілом. Класифікатор максимальної ентропії може бути використаний для вирішення великої кількості різноманітних завдань класифікації тексту, таких як виявлення спаму, сленгу, тематичної класифікації тощо.

Для автоматичного визначення емоційного забарвлення контенту соціальних мереж можна використовувати *нейронні мережі*. За допомогою нейронних мереж можна проводити аналіз емоційного навантаження смайлів та картинок, що постяться у соціальних мережах, а також аналіз емоційного забарвлення текстових даних.

Перевагами застосування нейронних мереж є: можливість рішення задач при невідомих закономірностях, здатність до навчання, стійкість до шумів у вхідних даних.

Алгоритми навчання штучних нейронних мереж поділяються на алгоритми навчання з учителем та без учителя. Навчання нейронної мережі в першу чергу полягає в зміні вагових коефіцієнтів синоптичних зв’язків між нейронами.

Для аналізу текстових даних, доцільно застосовувати глибоке навчання рекурентних нейронних мереж, яке не викликає складнощів із перенавчанням, на відміну від згорточних та повнозв’язних нейромереж.

Переваги та недоліки досліджених методів автоматичного визначення тональності тексту представлені у таблиці1.

Таблиця 1 – Порівняльний аналіз методів автоматичного визначення тональності у текстах та мультимедійних даних

Назва методу	Необхідність застосування словників	Необхідність попередньої лінгвістичної обробки тексту	Можливість застосування до різних типів даних	Можливість застосування для виявлення інформаційно-психологічних впливів
Лексемний метод	+	+	Тільки до текстів	+
Метод опорних векторів	-	-	До різних типів даних	+
Дерева прийняття рішень	-	+	До різних типів даних	-
Наївний класифікатор Баєса	+	-	Тільки до текстів	+
Метод максимальної ентропії	+	-	Тільки до текстів	+
Нейронні мережі	-	-	До різних типів даних	+

ВИСНОВКИ

В даній роботі розв’язувалась задачі побудови системи для оцінювання емоційного відклику за коментарями. В роботі отримані наступні результати:

- Проведено аналіз існуючих методів (лексемний метод, метод опорних векторів, дерева прийняття рішень, наївний класифікатор Баєса, метод максимальної ентропії, нейронні мережі). В результаті аналізу для даної роботи було обрано лексемний метод.
- Проаналізовано переваги та недоліки розглянутих методів.
- Було розроблено модуль аналізу тональності текстів на українській мові