

Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____ Комп'ютерних наук _____
(повна назва)
Кафедра _____ Програмної інженерії _____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____ другий (магістерський) _____
(рівень вищої освіти)

_____ (позначення документа)

_____ Дослідження методів емоціонального окрасу тексту з емотіконами _____

_____ (тема)

Виконав: студент 2 курсу, групи ПЗСм-16-2
спеціальності 121 – Інженерія програмного
забезпечення

_____ (код і повна назва спеціальності)
спеціалізації програмне забезпечення
систем

_____ (повна назва спеціалізації)

_____ Пугачов Є.А. _____
(прізвище, ініціали)

Керівник _____ доцент Вечур О.В. _____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри

_____ (підпис)

_____ Дудар З. В. _____
(прізвище, ініціали)

2018 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)

Спеціалізація _____ Програмне забезпечення систем _____
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

« _____ » _____ 20 ____ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові _____ Пугачову Євгену Анатолійовичу _____
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів емоціонального окрасу текст емотіконами

затверджена наказом по університету від 16 квітня 2018 р. № 451 Ст _____

2. Термін подання студентом роботи до екзаменаційної комісії _____ 20 ____ р.

3. Вихідні дані до роботи _____ Дослідити методи аналізу емоціонального окрасу тексту з емотіконами. Розглянути методи машинного навчання для визначення полярності тексту. Розробити систему для автоматичного визначення тональності на основі досліджуваних методів. Визначити практичне значення та можливості застосування отриманих результатів на практиці.

4. Перелік питань, що потрібно опрацювати в роботі _____ Мета роботи, аналіз проблемної галузі і постановка задачі, опис досліджуваних методів, архітектура програмної системи та структура даних, опис розробленої програмної системи, можливість використання отриманих результатів у науковій і практичній діяльності.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Мета роботи, системи з аналізу тональності, постановка задачі, сентімент аналіз, емотікони, метод на основі словника, наївний Баєс, SVM метод, метод на основі емотиконів, головна сторінка веб-системи, сторінка результатів, результати, прикладне використання, висновки.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів	Примітка
1	Об'єктний аналіз поставленої задачі	19-04-2018	виконано
2	Аналіз предметної області	23-04-2018	виконано
3	Опрацювання літератури	26-04-2018	виконано
4	Дослідження методів	30-04-2018	виконано
5	Розробка структури взаємодії даних	05-05-2018	виконано
6	Створення коду програми	13-05-2018	виконано
7	Тестування і налагодження	20-05-2018	виконано
8	Підготовка пояснювальної записки	29-05-2018	виконано
9	Підготовка презентації та доповіді	03-06-2018	виконано
10	Попередній захист	06-06-2018	виконано
11	Нормо контроль, рецензування	08-06-2018	виконано
12	Занесення диплома в електронний архів	12-06-2018	виконано
13	Допуск до захисту у зав. кафедри	12-06-2018	виконано

Дата видачі завдання _____ 20__ р.

Студент _____
(підпис)

Керівник роботи _____ доц. Вечур О.В. _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: __ с., __ рис., __ табл., __ додатки, __ джерел.

АНАЛІЗ ТОНАЛЬНОСТІ, ЕМОТІКОНИ, МЕТОД ОПОРНИХ ВЕКТОРІ, НАЇВНИЙ БАССІВ КЛАСИФІКАТОР, СОЦІАЛЬНІ МЕРЕЖІ, ТВІТТЕР

Об'єктом дослідження є методи аналізу тональності тексту, які застосовуються до аналізу тональності англomовного тексту з емотіконами у соціальних мережах.

Метою роботи є дослідження методів аналізу емоційного забарвлення тексту з емотіконами та порівняння ефективності роботи різних алгоритмів аналізу тональності.

У результаті роботи розглянуті методи аналізу тексту, що ґрунтуються на наївному бассівському класифікаторі, методі опорних векторів з урахуванням емотіконів, методі в основі якого лежить словник та методі визначення оцінки ґрунтованому на значенні забарвлення смайла. Здійснена програмна реалізація системи для автоматичного визначення тональності на основі досліджуваних методів.

SENTIMENT ANALYSIS, EMOTONICS, METHOD OF OPPORTUNAL VECTORS, IMAGE BAYS CLASSIFIER, SOCIAL NETWORKS, TWITTER

The object of the study is the methods of analysis of the tonality of the text, which are used to analyze the tone of the English text with emoticons in social networks.

The purpose of the work is to study the methods of analyzing the emotional color of the text with emoticons and compare the performance of various algorithms of tonality analysis.

As a result of the paper, the methods of analysis of the text based on the naive Bai's classifier, the method of reference vectors taking into account emoticons, the method based on which is the dictionary and the method of estimation based on the meaning of smile coloring, are considered. The program implementation of the system for the automatic determination of tonality based on the research methods is carried out.

ЗМІСТ

ВСТУП.....	5
1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ	7
1.1 Аналіз предметної області	7
1.2 Аналіз існуючих аналогів.....	9
1.3 Постановка задачі.....	12
2 АНАЛІЗ ЕМОЦІЙНОГО ОКРАСУ ТЕКСТУ З ЕМОТІКОНАМИ.....	14
2.1 Інтелектуальний аналіз даних.....	14
2.2 Аналіз тональності текстів.....	17
2.3 Емотікони.....	20
2.4 Методи дослідження.....	22
2.4.1 Метод на основі емотіконів	25
2.4.2 Метод на основі словника	27
2.4.3 Наївний Баєсівський метод.....	28
2.4.4 Метод опорних векторів.....	31
3 ПРОГРАМНА РЕАЛІЗАЦІЯ.....	34
3.1 Джерело вхідних даних	34
3.2 Архітектура програмної та структура даних.....	35
3.3 Опис програмної системи	45
3.4 Результати	48
4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ.....	55
ВИСНОВКИ.....	61
ПЕРЕЛІК ПОСИЛАНЬ	63
Додаток А Слайди презентації.....	65
Додаток Б Тезиси.....	73
Додаток В Код програми	75

ВСТУП

Завдання аналізу емоційного забарвлення текстів, розвиток методів фільтрації в мережі Інтернет набувають все більшої актуальності в зв'язку з величезною аудиторією мережі, зростаючим середнім часом перебування в ній, а також великим охопленням серед дітей і підлітків. Аналітика та моніторинг соціальних мереж становить величезний інтерес для соціологів, лінгвістів, психологів, маркетологів і політологів. Для вирішення завдань аналізу емоціональної забарвлення тексту в комп'ютерній лінгвістиці використовуються методи контент-аналізу, загальна назва для яких – Sentiment Analysis (аналіз тональності тексту).

Під словами аналіз тональності слід розуміти область комп'ютерної лінгвістики, що займається вивченням думок і емоцій в текстових документах. Аналіз тональності використовується для знаходжень думок і визначення їх властивостей, відносно вхідного тексту. При аналізі визначаються різні властивості, це може бути, автор, тема. Способи спілкування в соціальних мережах сильно відрізняються від норм літературної мови. І характеризуються використанням сленгових слів, авторської пунктуації, помилок і більш частим використанням смайлів.

Аналіз тональності тексту відносно новим напрямком автоматизації аналізу емоційної складової тексту. Правильне його застосування дозволяє оцінити реакцію користувачів на той чи інший об'єкт і врахувати її в подальшому. Однак проблемою такого аналізу є те, що не завжди можна просто визначити точне емоційне забарвлення тексту опираючись тільки на окреме слово[1]. Поширене використання набули емотікони та аббревіатурні скорочення, які в сукупності можуть нести зовсім інший емоційний зміст ніж по одинці. Або ж текст може містити велику кількість негативних або позитивних слів і все одно виражати зовсім протилежну думку. Тому одним з напрямків аналізу тональності тексту є вибір методів таким чином, щоб проводити класифікацію максимально точно.

У даній роботі розглядаються способи визначення тональності текстів відгуків і коротких повідомлень які містять в собі емотікони. Емотікон – це піктограма або послідовність друкованих знаків, що відображає емоцію.

Об'єктом дослідження є методи аналізу тональності тексту, які застосовуються до аналізу тональності англomовного тексту з емотіконами у соціальних мережах.

Метою роботи є дослідження методів аналізу емоційного забарвлення тексту з емотіконами та порівняння ефективності роботи різних алгоритмів аналізу тональності.

Розглянуті в даній роботі методи аналізу тексту ґрунтуються на наївному баєсівському класифікаторі, методі опорних векторів з урахуванням емотіконів, методі в основі якого лежить словник та методі визначення оцінки ґрунтованому на значенні забарвлення смайла.

Отримані результати дозволяють стверджувати, що емотікони є відносно точним показником для визначення ставлення автора до об'єкту висловлювання. Врахування емотіконів при аналізі тексту за допомогою інших методів підвищує точність отриманих результатів.

Методи аналізу повідомлень в соціальних мережах можуть також стати кроком до створення принципово нових автоматизованих соціологічних і маркетингових досліджень тональності в конкретній предметній області.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Аналіз предметної області

У сучасному світі багато говорять про обробку природного тексту – причому, не тільки в наукових кругах, де ця концепція справедливо вважається основоположною для подальшого розвитку штучного інтелекту, але і серед маркетологів, політологів та представників ІТ-індустрії.

Серед найцікавіших і популярних методів цього широкого наукового напрямку окремо стоїть одна, що носить назву аналізу сентиментальності, що означає – аналіз тональності текстів. Загальне визначення свідчить, що аналіз тональності текстів – це клас методів контент-аналізу, призначений для автоматичного виявлення в тексті емоційно забарвленої лексики, а також думок автора з приводу об'єктів, про які йде мова в тексті.

З визначення можна зробити кілька висновків про те, де теоретично (і, якщо вже на те пішло, практично) концепція аналізу тональності тексту могла б знайти застосування і прояснити деякі її деталі.

По-перше, аналіз тональності текстів здатний допомогти розібратися в законах, за якими живе природна мова і навчити комп'ютер сприймати його на рівні, наближеному до людського. До недавнього часу машина розуміла тексти на абстрактному рівні – в основному, через лексеми, які для неї мають формою і зміст[2]. Дана концепція пропонує ввести ще одну функцію – так звану лексичну тональність тексту в найпростішому випадку вона визначається як сума лексичних тональностей кожної окремої лексеми.

По-друге, аналіз тональності здатний значно покращити якість. Відомо, що еталоном машинного перекладу служить результат перекладу тексту людиною – професійним перекладачем. За 50 з гаком років розробок в цій області дослідники переконалися в тому, що навчити машину думати, як перекладач можна лише взявши до уваги всі ті міркування, якими користується професіонал, переводячи той чи інший текст. Природно, при перекладі не обійтися без первинного аналізу

тексту та окремих слів – в тому числі, аналізу тональності як такого.

По–третє, метою аналізу тональності тексту може бути якась думка автора або сам автор. Це – найбільш цікава сфера застосування, оскільки тут бачиться не тільки спосіб делегування машині деяких повноважень вченого, наприклад, філолога, який досліджує твір того чи іншого автора, але і знову спроба наблизити образ мислення комп'ютера до людського[2]. З цієї точки зору аналіз тональності, можливо, є одним з найбільш важливих і перспективних кроків до розвитку штучного інтелекту.

Аналіз тональності використовується для знаходжень думок і визначення їх властивостей, відносно вхідного тексту. При аналізі визначаються різні властивості, наприклад:

- автор – суб'єкт висловлює думку соціологія;
- тема – об'єкт про яких йде мова;
- тональність – ставлення автора до теми тексту.

У мережі Інтернет міститься величезна кількість різноманітних текстів, авторами яких є звичайні користувачі. Це можуть бути статті в блогах, відгуки на продукти, повідомлення в соціальних мережах і т. п. У цьому контенті міститься велика кількість цінної інформації яка може біти корисна та тих фахівців, діяльність яких залежить від думок людей.

У сучасному світі на наш вибір в будь–яких ситуаціях найчастіше впливає думка інших людей – ми читаємо відгуки про товар, перш ніж замовити його в інтернет–магазині, дізнаємося думку інших людей, перш ніж проголосувати на виборах за того чи іншого кандидата, довго і ретельно вибираємо собі ВНЗ, місце роботи або ресторан, який ми збираємося відвідати. Ця інформація становить значний інтерес для маркетологів, соціологів і багатьох інших фахівців. Крім того, для власників інтернет–ресурсів життєво важливо знати думку користувачів – будь це думка щодо зробленого на порталі нововведення, свіжої новини на сайті або оцінка користувачами товару в інтернет–магазині. Все вищесказане робить актуальним завдання аналізу тональності тексту.

1.2 Аналіз існуючих аналогів

В результаті проведення аналізу веб–систем з визначення тональності текстів було виявлено лише декілька систем які відповідають сучасним реаліям на критеріям які до них ставляться користувачами.

Веб–сервіс Sentiment140 дозволяє аналізувати інформацію про продукти, котрі згадують користувачі (рис 1.1), за допомогою даних з соціальних мережі Twitter.



Рисунок 1.1 – Результати роботи сервісу Sentiment140

Користувачеві Twitter Sentiment досить ввести слово, і програма проаналізує до 100 останніх записів про цьому слові. при цьому буде побудований графік співвідношення позитивних і негативних відгуків. В сукупності, надається легкий спосіб проаналізувати думки користувачів про будь–яких продуктах. Система, також, надає доступ до власного API, що дає можливість використовувати систему в інших ресурсах.

До недоліков системи можна віднести, відсутність можливості проаналізувати конкретний твіт або текст і використання лише одного методу при аналізі повідомлень.

Система Social Mention дозволяє легко відстежувати і вимірювати відгуки про компанії, нові продукти або з якоїсь іншої аналізованої темі в режимі реального часу (рис. 1.2). Система проводить моніторинг понад сотні соціальних ресурсів, включаючи Twitter, Facebook, і YouTube.

За ключовим словом можна отримати цілий ряд параметрів, в тому числі дізнатися кількість позитивних, негативних і нейтральних згадок.

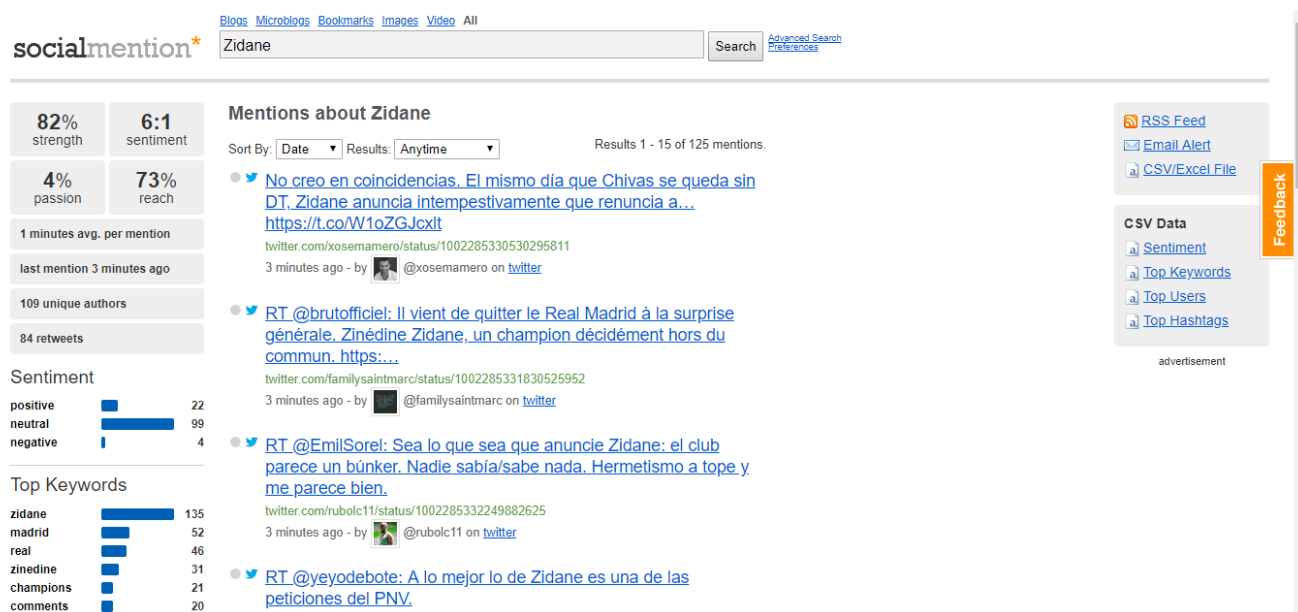


Рисунок 1.2 – Головна сторінка веб-системи SocialMention

Недоліками даної системи є все таж відсутність можливості аналізу конкретного повідомлення та аналізу окремо введеного тексту. Отримані результати аналізу не є основним контентом на сторінці, нажаль система має не інтуїтивно розроблений інтерфейс.

Веб-додаток Tweet Visualizer один з найбільш надійних, високо функціональних та безкоштовних інструментів для аналізу настроїв Twitter(рис. 1.3). Система дуже просто в використанні і розумінні принципу роботи. Необхідно просто ввести ключове слово, а програма Tweet Visualizer автоматично витягує останні твіти, давністю до одного тижня.

Система аналізує повідомлення не тільки за емоційним забарвленням, але й за великою іншою кількістю параметрів. Після чого можливо вивчити багато

варіантів візуалізації, які пропонує інструмент для твітів.

Однією з дуже корисних функцій візуалізації Tweet Visualizer є те, що надає інтерактивну можливість, щоб витягнути окремі твіти з ідентифікованих і подивитися, де вони потрапляють у емоційний спектр.



Рисунок 1.3 – Результати роботи сервісу Tweet Visualizer

Однозначно, даний сервіс є найкращим варіантом для визначення тональності повідомлень з соціальної мережі. Система надає можливість проаналізувати текст за безліччю параметрів та отримати детальні результати. Нажаль, система не використовує при аналізі методи з використанням смайлів та емотіконів.

Проаналізувавши можливості кожної з основних систем з сентімент аналізу текстів, можна зробити висновок, що існує попит на дослідження емоційного окрасу тексту, але сучасні систему не в повному обсязі задовольняють потреби користувачів.

Оскільки дійсно серйозних конкурентів, які б відповідали всім вимогам користувача до даного програмного продукту знайдено не було, то ідея про розробку даної програмної системи є прийнятною.

1.3 Постановка задачі

Метою атестаційної роботи є дослідження та розробка методів аналізу тональності текстів з емотіконами на основі повідомлень в соціальних мережах та розробка системи для автоматично аналізу тексту на основі досліджуваних методів.

Для досягнення даної мети були поставлені такі завдання:

- провести огляд існуючих методів автоматичного аналізу емоційного забарвлення текстів;
- провести дослідження текстових особливостей повідомлень в соціальних мережах в контексті розробки методів аналізу їх емоційного забарвлення;
- дослідити рівень впливу та відповідності емотіконів при визначенні загальної тональності тексту;
- розробити систему автоматичного визначення емоційного забарвлення повідомлень в соціальних мережах;
- дослідити можливе значення отриманих результатів в науковій та практичній діяльності.

При дослідженні методів та підходів для аналізу тексту, слід дослідити рівень відповідності емоційного забарвлення емотікона до забарвлення тексту загалом[4]. А також способи використання емотіконів в методах аналізу тональності тексту. Тому одним з напрямків аналізу тональності тексту є вибір методів таким чином, щоб проводити класифікацію максимально точно, враховуючи різні можливі комбінації.

Отримані результати маю бути наведені в зручному вигляді для опрацювання та аналізу. Результати отримані в ході даної роботи можуть стати кроком до створення новітніх систем для роботи не лише с текстом, а й з людьми, допоможуть краще розуміти співрозмовників, покупців, клієнтів та допомагати їм.

При розробці системи слід враховувати новітні тенденції в світі розробки програмного забезпечення та використовувати досвід попередніх, схожих проектів.

Система з аналізу тексту, як і будь-який сучасний, програмний продукт має відповідати декільком параметрами:

- безпечність;
- інтегрованість з соціальними мережами;
- відмово стійкість;
- наочність отриманих результатів;
- візуальне оздоблення повинне бути приємним.

Враховавши всі вище перераховані параметри, можна створити, стабільну та добре оптимізовану систему з аналізу повідомлень в соціальній мережі. Програма повинна оптимально використовувати ресурси, бути стійкою та безпечною, мати інтеграцію з соціальною мережею, щоб користувач маю лише одне посилення на повідомлення міг визначити його тональність, не роблячи рутині дії з копіювання повідомлень. Система, створена з урахуванням цих параметрів, буде поєднувати в собі переваги аналогічних систем, при цьому усуваючи їх недоліки. Така система буде мати високу ефективність аналізу і широкі області застосування.

2 АНАЛІЗ ЕМОЦІЙНОГО ОКРАСУ ТЕКСТУ З ЕМОТІКОНАМИ

2.1 Інтелектуальний аналіз даних

Інтелектуальний аналіз даних (ІАД), або як його ще називають – Data mining. Сучасна концепція аналізу даних, яка припускає, що дані можуть бути неточними, неповними (містити пропуски), суперечливими, різнорідними, непрямими, і при цьому мати гігантські обсяги. Тому розуміння даних в конкретних програмах вимагає значних інтелектуальних зусиль. В інтелектуальному аналізі даних застосовується математичний апарат для виявлення закономірностей і тенденцій, що існують в даних[5]. Зазвичай, такі закономірності не можна виявити при традиційному перегляді даних, оскільки зв'язки занадто складні, або через надмірні обсяги даних. Побудова моделі інтелектуального аналізу даних є частиною більш масштабного процесу, в який входять всі завдання, від формулювання питань щодо даних і створення моделі для відповідей на ці питання до розгортання моделі в робочому середовищі. Його методи запозичені з областей штучного інтелекту, машинного навчання, комп'ютерних наук, технологій баз даних та статистики.

Інтелектуальний аналіз даних дозволяє підвищити ефективність ведення бізнесу, отримати конкурентні переваги і, як результат, збільшити прибутковість компанії.

Необхідність інтелектуального аналізу даних виникла в кінці ХХ століття в результаті повсюдного поширення інформаційних технологій, що дозволяють детально протоколювати процеси бізнесу і виробництва. Великі обсяги даних, широту і різноманітність інформації привели до вибухового зростання популярності методів інтелектуального аналізу даних. Починаючи з 60-х років, з появою засобів автоматизації і текстів в електронному вигляді, набув розвитку контент-аналіз інформації з великими обсягами[6]. Під Data Mining, з погляду контент-аналізу, розуміють механізм виявлення в потоці даних нових знань, таких як моделі, конструкції, асоціації, зміни, аномалії і структурні новоутворення. Контент-аналіз – це якісно-кількісна, систематична обробка, оцінка та

інтерпретація форми і змісту тексту.

З виникненням та глобальним розповсюдженням соціальних мереж, на початку XXI століття, об'єм даних для аналізу збільшився в рази. Кожен хто має доступ до мережі Інтернет висловлює свою думку в мережі, це може бути рецензія до фільму, відгук про товар або висловлення відношення про певну подію[7].

Метою ІАД є вилучення корисної інформації або знань з будь-якого набору даних і приведення їх до зрозумілого вигляду. ІАД вирішує безліч завдань, основними з яких є:

- класифікація;
- кластеризація;
- асоціація;
- регресія;
- прогнозування;
- послідовність;
- визначення відхилень або викидів.

Розглянемо кожную задачу більш детально. Класифікація є найбільш частою завданням ІАД. По суті своїй, класифікація являє собою акт присвоєння категорії кожному об'єкту. Будь-який об'єкт містить набір ознак, які характеризують ту чи іншу категорію. Аналізуючи ознаки об'єкта, класифікатор визначає, до якої категорії його віднести. Завдання кластеризації виникає в тому випадку, коли дані потрібно згрупувати, тобто знайти природні групи об'єктів на основі будь-яких ознак. Об'єкти, що потрапляють в одну групу мають схожі ознаки. Кластеризація це задача навчання без учителя[8]. Більшість алгоритмів кластеризації будують модель за допомогою ряду ітерацій і зупиняються, коли модель сходиться, тобто коли кордони сегментів стабілізуються.

Асоціація. Дане завдання переслідує дві мети: перебування елементів, які часто з'являються разом і, відштовхуючись від цього, визначення асоціативного правила, за яким це відбувається. Прикладом даного завдання може бути покупка супутніх товарів.

Регресія. Дане завдання схоже з завданням класифікації, але замість того,

щоб шукати ознаки, які описують ту чи іншу категорію, шукаються закономірності, що визначають чисельне значення, наприклад, вік, вага, відстань тощо.

Прогнозування є важливим завданням ІАД. Мета прогнозування – пророкування майбутніх подій. В якості вхідних даних використовується послідовність цифр, що представляє собою особливості історичних даних. Спираючись на ці дані, здійснюється прогноз. Прикладом даного завдання може бути передбачення кількості продажу товару[9].

Послідовність. Дане завдання полягає в пошуку закономірностей в ланцюжку подій, пов'язаних у часі. В якомусь сенсі, це узагальнення завдання асоціації, оскільки в даному випадку знаходиться закономірність не між одночасно наступаючими подіями, а подіями що відбуваються у часі. Так, наприклад, на при покупці будинку в половині випадків протягом місяця купується нова кухонна плита, а в рамках двох тижнів 60% відсотків новоселів обзаводяться холодильниками.

Визначення відхилень або викидів. Дане завдання визначається як пошук і аналіз даних, сильно відмінних від загальної множини даних.

Крім самої тональності, текст можна оцінювати по суб'єктивності або об'єктивності судження (Opinion Mining). Якщо ця думка автора висловлювання, що містить суб'єктивну оцінку описуваного, то текст вважається суб'єктивним. І навпаки, якщо це повідомлення ЗМІ або думка, за замовчуванням розділяється учасниками діалогу, то воно вважається об'єктивним.

Основна особливість ІАД – це поєднання широкого математичного інструментарію, від класичного статистичного аналізу до нових кібернетичних методів і останніх досягнень у сфері інформаційних технологій[10]. У технології Data Mining гармонійно об'єдналися строго формалізовані методи і методи неформального аналізу, кількісний і якісний аналізи даних. Більшість аналітичних методів, які використовуються в технології Data Mining, – це відомі математичні алгоритми і методи. Новим є те, що їх можна застосовувати при рішенні тих або інших конкретних проблем. Це обумовлено новими властивостями технічних і програмних засобів.

У даній роботі розглядається завдання класифікації тексту за ознакою емоційного забарвлення. Людина оцінює світ відразу за багатьма шкалами, хороший–поганий, сильний–слабкий, великий–маленький, щасливий–нещасливий, веселий–сумний, швидкий–повільний, і шкали ці по-різному емоційно навантажені. Але для простоти можна вважати, що емоційна оцінка зводиться до шкали хороший–поганий або позитивний–негативний. Текст класифікується на позитивно або негативно забарвлений. Для більшої точності доцільно використовувати класифікацію, ще й на нейтрально забарвлені тексти. Але такий підхід є більш складним, порівняно з класифікацією на дві категорії.

2.2 Аналіз тональності текстів

Сентимент–аналіз, або аналіз тональності тексту, представляє великий інтерес для сфер та інститутів суспільства, що оперують з текстовими документами. Особливо це відноситься до сфер освіти, журналістики, культури, видавничої діяльності, ефективність яких обумовлена якістю тексту, а вміння і навички роботи з ним входять до складу професійних вимог. Емоційне забарвлення тексту в загальному випадку є багатовимірною. Таким чином, за допомогою сентимент–аналізу відгуків і листування людей на форумах пропонується автоматично оцінювати громадську думку щодо обговорюваних об’єктів[11].

Широта охоплення аудиторії в мільйони чоловік і оперативність отримання інформації дозволили отримувати недосяжні раніше результати досліджень. Якщо раніше, щоб виявити думку з будь–якого питання, потрібно було проводити опитування, то сьогодні висловлювання по величезній кількості популярних тем вже є в мережі, треба тільки виявити їх, розпізнати і оцінити.

Історично склалося так, що традиційний підхід до сентимент аналізу являє собою задачу класифікації тексту на дві–три категорії. Саме з такого завдання почав свій розвиток аналіз тональності: оцінити сентимент оціночних відгуків з

якої–небудь тематики кіно, ресторани, електроніка та ін.

Тим не менш, це не єдиний і не визначальний тип завдання, яке має вирішувати сентимент аналіз тексту. В даний час цікавить не загальна емоційна оцінка тексту, а відношення сентимент до конкретного об'єкта, про що йдеться в тексті, або відношення суб'єкта висловлювання до обговорюваного об'єкту.

Технологія аналізу знайшла широке комерційне застосування у корпорацій – власників брендів для аналізу соціальних медіа. Сучасні додатки надають можливість не тільки оцінити тональність висловлювань про бренд, а й отримати цілий ряд додаткових інструментів, що спрощують управління соціального аудиторією, яка цікавиться брендом, встановлення контактів, обмін інформацією, вплив на вирощування соціального контенту, пошук лідерів думок соціальної спільноти, постачання їх інформацією і залучення до просування бренду[12].

Тональність тексту в цілому визначається лексичною тональністю складових його одиниць і правилами їх поєднання. Тональність тексту визначається трьома факторами: суб'єкт тональності, тональна оцінка, об'єкт тональності. Суб'єктом тональності є автор тексту, об'єкт тональності – те, про що або про кого йде в тексті мова.

Нижче наведені основні проблемні місця аналізу тональності повідомлень. Саме цим слабким місця, необхідно приділити особливу увагу при роботі з текстами.

Сарказм: це одне з найскладніших почуттів для автоматичного відстеження і правильної інтерпретації. Приклад: «Яка чудова у них служба підтримки, через три дні передзвонили».

Самомилування: коли під час моніторингу соціальних мереж з'являються записи, пов'язані з вашими власними рекламними зусиллями, і їх необхідно відфільтрувати.

Нейтральні настрої: схожі на «невизначених» виборців на виборах, чиї голоси можуть вирішити результат боротьби.

Порівняльний настрій: не класичний негатив, але тим не менш, може мати негативний окрас. Приклад: «Я купив iPhone», що добре для Apple, але не для

Samsung.

Змішаність або багато вимірність настрою: містять позитиви і негативи в одній і тій же фразі. Приклад: «Мені подобається серіал Mad Men, але мене дратують нав'язливі трейлери нових серій».

Умовний настрій: пов'язано з діями, які можуть статися в майбутньому. Приклад: покупець не рздратований зараз, але каже, що буде, якщо представники компанії не передзвонять йому.

Позитивні почуття можуть бути не пов'язані з основною темою. Наприклад, багато коментарів про акторів зосереджені на їх особистому житті, а не на їх акторській майстерності. Негативні настрої не обов'язково погані: Це пов'язано з класичною дилемою, щодо негативного висвітлення в ЗМІ. Приклад: поява відомого американського політика на одному з тв-шоу викликало багато негативних коментарів, але все ж значно збільшило загальний рейтинг. Погана реклама, також реклама.

Неоднозначні негативні слова: їх контекст повинен бути розібраний і позначений відповідним чином. Фраза «Який стрибок, з глузду з'їхати!» носить позитивний забарвлення, хоча в іншому контексті ці слова могли бути витлумачені інакше.

Існує два основні методи вирішення завдання автоматичного визначення тональності. Статистичний метод. Для нього потрібні заздалегідь розмічені по тональності колекції(корпус) текстів, на яких відбувається навчання моделі, за допомогою якої і відбувається визначення тональності тексту або фрази[13].

Метод, заснований на словниках і правилах. Для цього заздалегідь складаються словники позитивних і негативних слів і виразів. Цей метод може використовувати як списки шаблонів, так і правила з'єднання тональної лексики всередині пропозиції, засновані на граматичному і синтаксичному розборі. Крім того, іноді використовують змішаний метод, комбінацію першого і другого підходів. В даній роботі розглядається, ще один метод, заснований на емотіконах та їх емоційному забарвленні.

2.3 Емотікони

Інтернет–простір на сучасному етапі розвитку формує нові комунікативні практики, нові стратегії взаємодії учасників, що реалізуються в сучасному комунікаційному просторі за допомогою повідомлень, постів, відгуків.

Сьогодні складно уявити Інтернет–спілкування без смайликів. Багато з нас, очевидно, вже не замислюючись, ставлять в кінці свого повідомлення дужку, яка не несе жодної граматичної навантаження. Її завдання передати настрій відправника. Це дороговказ, необхідний мінімум, плоть і кров нашого віртуального спілкування. Ти посміхаєшся або посміхається? Веселий ти або засмучений? Щоб зрозуміти це, для нас недостатньо слів, обов'язково потрібен знак.

Емотікони, емограма або смайлик – це графічний символ, який використовується для вираження емоції. Зручність його в тому, що він, по–перше, просто малюється, що дозволяє легко використовувати його на листі, а по–друге, що набагато важливіше саме для Інтернету та sms, легко вставляється в друкований текст без використання будь–яких додаткових дій. Смайли призначені для того, щоб більш багато і різноманітно доповнювати зміст висловлювання, уточнювати його експресивно–інтонаційну забарвлення. Емотікони і піктограми є прикладами специфічного застосування знакового різноманітності клавіатури з метою представлення емоційного стану співрозмовника. Емотікони представляють собою послідовність з допоміжних символів і знаків пунктуації, що позначають емоції пише – позитивні чи негативні.

Більшість емотіконів – це варіації від основи :-). Проте користувачі всесвітньої мережі вдаються до синтаксичним девіацій, таких як збільшення кількості функціональних розділових знаків в графічних комплексах. Смайлик – це абстрактне кодування мимічного вираження емоційної експресії. Узагальнюючи, можна сказати, що поява і характер смайликів обумовлені наступними факторами:

- відсутністю в письмовій комунікації каналу для невербальної

інформації;

- відсутністю адекватних лінгвістичних засобів для кодування емоцій;
- необхідною швидкістю передачі інформації;
- обмеженістю обсягу повідомлень.

В активному використанні зараз налічується близько тридцяти різних смайликів, що відповідають різним відтінкам емоційної експресії. Взагалі їх існує набагато більше. Смайли, система, що динамічно розвивається, і тому вони не мають стійкого, раз і назавжди прийнятого набору знаків. З часом з'являються все нові і нові комбінації символів які характеризують певні явища, емоції. Об'єднує всі існуючі в мережі інтернет смайли їх загальне функціональне призначення – встановлювати і підтримувати контакт зі співрозмовником, більш точно і конкретно висловлювати свій емоційний стан. Набагато рідше смайли служать для позначення різних понять, абстрактних чи конкретних об'єктів, дій і станів людини[14].

Особливість класичних смайликів – в горизонтальному розташуванні вертикальних зон обличчя і тіла(рис. 2.1). Ці смайлики позначають емоції, міміку, жести, дії і стану людини, а також різних персонажів.

З часом комбінації друкованих символів, інтернет додатки стали, замінювати, на привабливе зображення, що збільшило ефект візуальності. Традиційна письмова комунікації була повільною, не інтерактивною і служила не стільки засобом спілкування, скільки засобом відчуження інформації від її носія, фіксації та трансляції в просторі і часі за умови неможливості передати її усно.

Сучасні засоби комунікації висувають жорсткі вимоги до швидкості та інтенсивності передачі інформації, її інтерактивності, онлайн–доступності, інтернаціональності, інформаційної щільності повідомлення (тобто співвідношення його інформативності до інформаційної ємності). У таких умовах письмова комунікація шукає засоби досить ємні і прості для швидкої, економної і універсальної передачі соціальних сенсів.


Емотікон	Смайл	Значення
:)		Посмішка
:D		Широка посмішка
;-)		Підморгування
:(	Смукот
:o		Здивування
:		Нейтральне
8O		Ботанік
:x		Роздратування
:P		Дражнити
:?		В сумніві

Рисунок 2.1 – Приклад зображення емотіконів

Також в Unicode є група символів «Емотікони» (1F600–1F64F), призначена для вираження емоцій.

2.4 Методи дослідження

Аналіз тональності зазвичай визначають як одну з задач комп'ютерної лінгвістики, тобто мається на увазі, що ми можемо знайти і класифікувати тональність, використовуючи інструменти обробки природної мови[15]. Зробивши велике узагальнення, можна розділити існуючі підходи на наступні категорії:

- методи на основі правил;
- методи засновані на словниках;
- машинне навчання без вчителя;
- машинне навчання;
- гібридний метод.

В таблиці 2.1 наведені дані порівняння найбільш популярних методів.

У першому випадку генеруються правила, на основі яких буде визначатися тональність тексту. Для цього текст розбивається на слова або послідовності слів. Потім отримані дані використовуються для виділення шаблонів які часто використовуються, яким присвоюється позитивна чи негативна оцінка. Наприклад,

для речення “я люблю спорт”, правило буде мати наступну структуру: якщо, присудок люблю, входить на переліку позитивних дієслів і речення не містить заперечень, то його тональність можна класифікувати як позитивну.

Таблиця 2.1 – Порівняльна характеристика методі

	Точність	Автоматизація	Дані для навчання	Простота використання	Використання в комерційних системах
Метод на правилах	найбільш точний	можлива	не потребує даних	-	+
Метод зі словником	не універсальний	можлива	потребує даних	+	-
Машинне навчання	точний	автоматична	потребує даних	+/-	+
Машинне навчання без вчителя	низька точність	автоматична	не потребує даних	+	+

Даний підхід, став популярним серед комерційних систем. Переважна більшість правил, таких систем, пов’язані з певною тематикою, наприклад політика або готельний бізнес. Основним недоліком даного підходу є те, що для хорошої роботи системі необхідно мати велику кількість правил, метод вимагає великих витрат як людських так і технічних[16]. Даний підхід не розглядається в роботі, через велику об’ємність даних необхідних для його реалізації та вузько направлену тематику правил. Тим не менш, цей підхід є найбільш точним при наявності хорошої бази правил.

При використанні підходів, заснованих на словниках, використовують так звані тональні словники для аналізу тексту. У простому вигляді тональний словник представляє з себе список слів і пропозиції, для яких відома оцінка вираженої в них тональності. Цей підхід ефективний при використанні великих словників, але процес їх складання досить трудомісткий.

Підхід машинного навчання без учителя заснований на ідеї, що найбільшу

вагу в тексті мають терміни, які найчастіше зустрічаються в цьому тексті, і в той же час присутні в невеликій кількості текстів всієї колекції. Виділивши ці терміни і визначивши їх тональність, можна зробити висновок про тональності всього тексту цілком. Машинне навчання без вчителя є, напевно, найбільш цікавим і в той же час найменш точний метод аналізу тональності. Одним із способів застосування цього методу може бути автоматична кластеризація документів.

Машинне навчання з учителем, або я його ще називають навчання по прецедентах, є найбільш поширеним методом, що використовується в дослідженнях. Його суть полягає в тому, щоб навчити машинний класифікатор на колекції заздалегідь розмічених текстів, а потім використовувати отриману модель для аналізу нових документів[17]. У цьому підході необхідна наявність навчальної колекції розмічених в рамках емотивного простору текстів, на базі якої відбуватиметься статистичний або імовірнісний класифікатор наприклад, баєсівський.

Якщо процес навчання проходить правильно, то алгоритм може узагальнити навчальні дані так, що наданні йому на вхід нові дані будуть правильно зіставлені з потрібними відповідями. Для того, щоб вирішити задачу навчання по прецедентах, необхідно виконати наступні кроки:

- визначити тип навчальних прикладів. Наприклад, це може бути одне слово уніграма, біграма чи триграма;
- зібрати навчальні дані: навчальна вибірка повинна бути репрезентативною;
- визначити уявлення вхідних ознак навченою функції, точність навченою функції строго залежить від того, як представлені вхідні об'єкти. Як правило, вхідний об'єкт перетвориться в вектор ознак, який містить ряд особливостей, що описують цей об'єкт. Кількість ознак не повинно бути занадто великим, але повинно містити достатньо інформації, щоб точно передбачити відповідь;
- визначити структуру навченою функції і відповідний алгоритму навчання;

- запустити алгоритм навчання на зібраних навчальних даних;
- оцінити точність навченою функції. Після налаштування параметрів і навчання, точність навчені функції повинна бути перевірена на тестових даних.

У деяких дослідженнях при поданні тексту всі слова проходять через процедуру стемінгу, видалення закінчення, або лематизації, приведення до початкової форми. Мета процедури – зменшення розмірності задачі, іншими словами – якщо в тексті зустрічаються однакові слова, але з різними закінченнями, за допомогою стемінгу і лематизації можна їх привести до одного виду. Однак, на практиці це зазвичай не дає відчутних результатів[17]. Причина цього в тому, що, позбавляючись від закінчень слів, ми втрачаємо морфологічну інформацію, яка може бути корисна для аналізу тональності. Наприклад, слова «хочу» і «хотів» мають різну тональність. Якщо в першому випадку тональність швидше за все позитивна, тому що автор може висловлювати надію і позитивні емоції, то у дієслова в минулому часі, тональність може бути негативною.

2.4.1 Метод на основі емотіконів

Як вже біло сказано вище, емотікони не залежать від мови і не підпорядковується граматичним правилам, будучи поняттям наднаціональним, одним з основних призначень якого – вираз невербальної інформації, емоцій. Емотікони ніколи не походять від слів, а, як правило, є спробою графічного вираження настрою або стану. Насправді, один емотікон може мати декілька варіантів відображення(рис 2.2). Його відображення може мати вигляд жовтого круглого обличчя з відповідною емоцією, або мати вигляд послідовності символів.

Зазвичай це символи двокрапки та дужки, в окремих випадках, може залишатися лише дужка. Також, деякі сервіси надають можливість використовувати текстове відображення сенсу емотікона.

Тож, даний метод повинен вміти правильно визначати емотікони в тексті та

вірно їх інтерпретувати. Наприклад, правильно визначити текст обернений в дужки від чередування двох різних за полярністю емотікона. Для такої оцінки потрібен перелік основних піктограм, що можуть бути використані для вираження емоції.










	:)	:-)	:smile:	Посмішка
	:D	:-D	:grin:	Широка посмішка
	;-)	;-)	:wink:	Підморгування
	:(:-(:(:sad:	Смуток
	:o	:-o	:eek:	Здивування
	:	:-	:neutral:	Нейтральне
	8O	8-O	:shock:	Ботанік
	:X	:-X	:mad:	Роздратування
	:P	:-P	:razz:	Дражнити
	:?	:-?	:???:	В сумніві

Рисунок 2.2 – Варіанти відображення емотіконів

Метод, визначення тональності тексту на основі емотіконів має декілька етапів та варіантів для аналізу.

На першому етапі вихідний текст перевіряється на наявність емотіконів, піктограм або послідовності друкарських знаків, що зображають емоцію. Емоційне забарвлення кожного емотікона задавалася згідно експертної оцінки автора роботи. У найпростішому випадку, якщо повідомлення містить емотікони, то тональність повідомлення визначається тональністю емотіконів. Коли всі емотікони мають однополярне забарвлення і необхідно визначити ступінь емоційного окрасу автора, слід визначити середнє значення. В іншому випадку, або, якщо повідомлення містить позитивний і негативний емотікони, програма переходить на наступний етап.

На цьому етапі слід визначити суму значень всіх позитивно і всіх емоційно негативно забарвлених емотіконів, та вирахувати з одного значення інше, таким чином знак результату буде відповідати відношенню автора до об'єкту повідомлення. У цьому випадку визначення середнього арифметичного значення

не буде дієвим способом, адже автор висловлює свої емоції не лише типом емотікона, а й їхньою кількістю. Також, варто враховувати чи є останній символ твіту позитивним емотиконом або негативним. Слід, відмітити, що результативність цього підходу зростає у поєднанні з іншими методами.

До переваг даного підходу, можна віднести його простоту та легкість програмної реалізації. До недоліків можна віднести необхідність мати та постійно підтримувати перелік емотіконів та проводити їх оцінювання і недоліком є неоднозначність при виокремленні емотіконів різних за полярністю.

Окремою проблемою є текст, який не містить в собі жодного емотікона, в будь-якому його вигляді. Визначити забарвлення такого тексту, даним методом не є можливим і слід використовувати один з нижче описаних методів для його класифікації за емоційним окрасом.

2.4.2 Метод на основі словника

При використанні словникового підходу ключову роль відіграє використання тональних словників.

Такі словники, як правило, представляють собою списки слів, які допомагають визначити ставлення автора до деякого об'єкту. Словники оціночної лексики можуть бути створені вручну або автоматично, деякі з них опубліковані і можуть використовуватися для проведення досліджень і вирішення практичних завдань[18].

Для створення словника з аналізу тональності текстів, як правило, обирають один з наступних підходів – на основі гібридного підходу, вручну і за допомогою корпусу. На практиці, зазвичай, використовують гібридний підхід. Спочатку на основі корпусу текстів деякої предметної області автоматично відбираються слова, які можуть впливати на тональність тексту. Далі експерти вручну оцінюють тональність кожного з цих слів в рамках даної галузі. Використання створених

таким чином словників дозволяє отримати результати, зіставні зі словниками, створеними автоматично, а в деяких випадках перевершити їх.

Слова, що входять до переліку словника можуть бути оціненими по бінарний шкалі, тобто слова–сентименти мають значення 1 чи -1, або використовувати більш широку шкалу оцінок і мати значення в діапазоні від -5 і до 5, чи навіть більше, де від'ємне число вказує на негативне забарвлення, а значення більше нуля говорить про позитивне забарвлення слова.

Часто разом з попереднім підходом використовується робота зі словниками слів–сентиментів. За знайденими в тексті лексичним тональностям він може бути оцінений за шкалою, що містить кількість позитивної і негативної лексики. Найпростіша оцінка – середнє арифметичне всіх значень полярності слів–сентиментів.

В деяких випадках, для більш точного зіставлення кожного слова в реченні зі словником виконуються синтаксичний аналіз і лематизація. Лематизація – це перетворення слів в лемму, тобто приведення їх до первісної словникової форми.

Основною проблемою словникових методів вважається процес складання словника: щоб отримати метод, що класифікує документ з високою точністю, терміни словника повинні мати вагу, адекватний предметної області документа. Наприклад, слово «великий» по відношенню до обсягу пам'яті жорсткого диска є позитивною характеристикою, але негативною по відношенню до розміру мобільного телефону.

2.4.3 Наївний Баєсівський метод

Мета наївного Баєсівського методу полягає в тому щоб зрозуміти до якого класу належить документ, тому нам потрібна не сама ймовірність, а найбільш ймовірний клас.

Баєсівський підхід до класифікації заснований на теоремі, яка стверджує, що якщо щільності розподілу кожного з класів відомі, то шуканий алгоритм можна виписати в явному аналітичному вигляді. На практиці щільності розподілу класів,

як правило, не відомі. Їх доводиться оцінювати (відновлювати) за навчальною вибіркою. В результаті баєсівський алгоритм перестає бути оптимальним, так як відновити щільність по вибірці можна тільки з деякою погрешністю. Чим коротше вибірка, тим вище шанси підігнати розподіл під конкретні дані і зіткнутися з ефектом перенавчання. Баєсівський підхід до класифікації є одним з найстаріших, але до сих пір зберігає міцні позиції в теорії розпізнавання[19]. Він лежить в основі багатьох досить вдалих алгоритмів класифікації.

В основі наївного Баєсова класифікатора лежить теорема Баєса (формула 2.1). яка описує ймовірність події, спираючись на обставини, що могли би бути пов'язані з цією подією.

Для наївного баєсівського класифікатора визначено істотне припущення – передбачається, що всі ознаки x_1, x_2, \dots, x_n документа B незалежні один від одного. Через це допущення модель і отримала назву «наївна». Це дуже серйозне спрощує допущення і, в загальному випадку, воно не так, але наївна Баєсова модель демонструє непогані результати, незважаючи на це.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (2.1)$$

де $P(A|B)$ – ймовірність що документ B належить класу A , саме її і треба розрахувати, $P(B|A)$ – ймовірність зустріти документ B серед всіх документів класу A , $P(A)$ – безумовна ймовірність зустріти документ класу A в корпусі документів, $P(B)$ – безумовна ймовірність документа B в корпусі документів.

Передбачається також, що позиція терміна в реченні не важлива. Теорема Баєса дозволяє переставити місцями причину і наслідок. Знаючи з якою ймовірністю причина призводить до якогось події, ця теорема дозволяє розрахувати ймовірність того що саме ця причина призвела до нинішнього події[20].

Для реалізації Баєсова класифікатора необхідна навчальна вибірка в якій проставлені відповідності між текстовими документами і їх класами. Потім нам

необхідно зібрати наступну статистику з вибірки, яка буде використовуватися на етапі класифікації:

- відносні частоти класів в корпусі документів. Тобто, як часто зустрічаються документи того чи іншого класу;
- сумарна кількість слів у документах кожного класу;
- відносні частоти слів у межах кожного класу;
- розмір словника вибірки. Кількість унікальних слів у вибірці.

Сукупність цієї інформації називається моделлю класифікатора. Потім на етапі класифікації необхідно для кожного класу розрахувати значення наступного виразу(формула 2.2.) і вибрати клас з максимальним значенням.

$$\log \frac{D_C}{D} + \sum_{i \in Q} \log \frac{W_{ic} + 1}{|V| + L_C}, \quad (2.2)$$

де D_c – кількість документів в навчальній вибірці належать класу C , D – загальна кількість документів в навчальній вибірці, $|V|$ – кількість унікальних слів у всіх документах навчальної вибірки, L_c – сумарна кількість слів у документах класу c в навчальній вибірці, W_{ic} – скільки разів i -е слово зустрічалось в документах класу c в навчальній вибірці, Q – безліч слів класифікуємого документа, включаючи повтори.

З точки зору програмування, вище наведена формула матиме вигляд формули 2.3. Існує невелика проблема, пов'язана з цією формулою. Якщо в тестовому наборі зустрінеється слово, яке не зустрічається в наборі навчальних документів, то ймовірність слова для будь-якого з класів буде дорівнює нулю.

$$\log \frac{D_C}{D} + \text{foreach}(\text{word}) \{ \log \frac{W_{ic} + 1}{|V| + L_C} \} \quad (2.3)$$

Основна відмінність полягає в тому, що для даного документа розглядається

не кількість входжень слова, а тільки їх наявність або відсутність. Незважаючи на найвний вигляд і, безсумнівно, дуже спрощені умови, найвні баєсівські класифікатори часто працюють набагато краще в багатьох складних життєвих ситуаціях.

Перевагою найвного баєсівського класифікатора є мала кількість даних для навчання, необхідних для оцінки параметрів, необхідних для класифікації.

2.4.4 Метод опорних векторів

Метод опорних векторів (Support Vector Mashine, SVM), запропонований В.М. Вапніком, відноситься до групи граничних методів класифікації. Він визначає приналежність об'єктів до класів за допомогою кордонів областей.

Класифікація даних – завдання машинного навчання, в цьому напрямку інтенсивно застосовуються методи оптимізації та аналітичної геометрії. Така класифікація має досить широке застосування: від розпізнавання образів до створення спам-фільтрів. Завдання класифікації полягає у визначенні до якого класу з, як мінімум, двох спочатку відомих належить цей об'єкт. Зазвичай таким об'єктом є вектор в n -вимірному просторі. Координати вектора описують окремі атрибути об'єкта. Наприклад, колір c , заданий в моделі RGB, є вектором в тривимірному просторі: $c = (\text{red}, \text{green}, \text{blue})$.

Якщо класів всього два, спам – не спам, червоне – чорне, то завдання називається бінарної класифікації. Якщо класів кілька – багато класова класифікація. Також можуть бути зразки кожного класу – об'єкти, про які заздалегідь відомо до якого класу вони належать. Такі завдання називають навчанням з учителем, а відомі дані називаються навчальною вибіркою. Якщо класи спочатку не задані, то перед нами завдання кластеризації. Даний метод вимагає навчання. Щоб показати SVM, що таке класи, використовується набір даних – тільки після цього він виявляється здатний класифікувати нові дані. Даний

метод спочатку відноситься до бінарним класифікаторів, хоча існують способи змусити його працювати і для завдань мульти класифікації[21].

Роботу методу зручно проілюструвати на простому прикладі: дані точки на площині, розбиті на два класи (рис. 2.3).

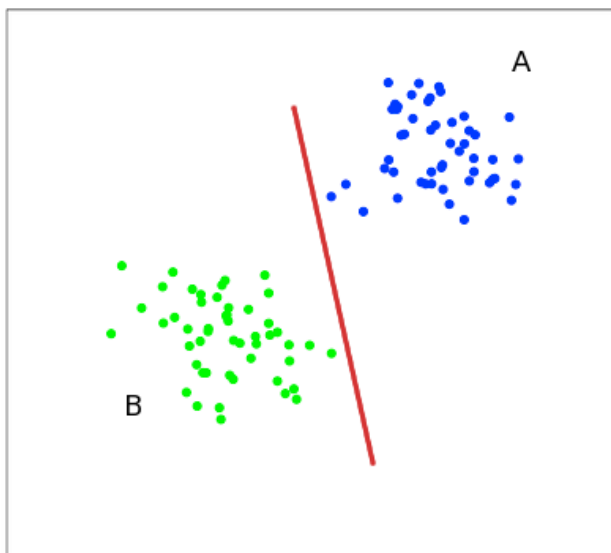


Рисунок 2.3 – Розділення вибірки на два класи

Проведена лінія, розділяє ці два класи. Далі, всі нові точки, не з навчальної вибірки, автоматично класифікуються наступним чином – точка вище прямої потрапляє в клас А, точка нижче прямої – в клас В.

Таку пряму називають – роздільною прямою. Однак, в просторах великих розмірностей пряма вже не буде розділяти наші класи, так як поняття «нижче прямої» або «вище прямої» втрачає будь-який сенс. Тому замість прямих необхідно розглядати гіперплощини – простору, розмірність яких на одиницю менше, ніж розмірність початкового простору. В, наприклад, гіперплощина – це звичайна двовимірна площина.

У даному прикладі, існує кілька прямих, які поділяють вибірку на два класи, хоча і роблять це з однаковим результатом, вони матимуть велике значення при розподілі вже нових даних (рис. 2.4).

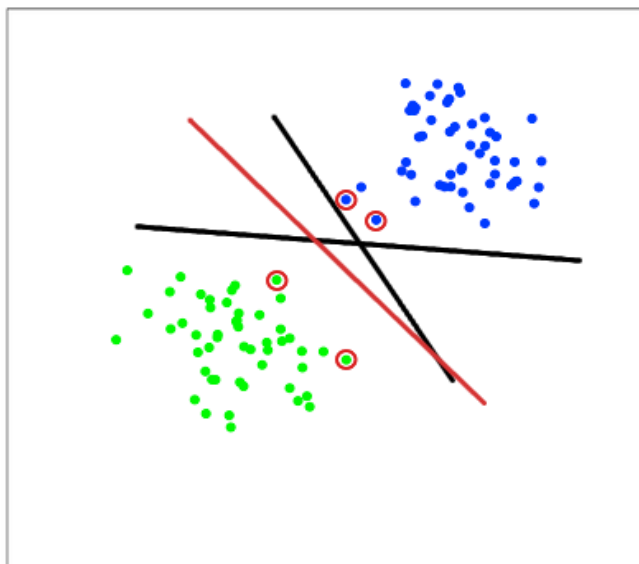


Рисунок 2.4 – Варіанти поділу вибірки

З точки зору точності класифікації найкраще вибрати пряму, відстань від якої до кожного класу максимально велике. Іншими словами, виберемо ту пряму, яка розділяє класи найкращим чином. Така пряма, а в загальному випадку – гіперплощина, називається оптимальною роздільною гіперплощиною. Метод знаходить роздільну смугу максимальної ширини, що дозволяє в подальшому здійснювати більш впевнену класифікацію[22].

Необхідно додати, що слабкими сторонами цього методу є необхідність вибору ядра і погана інтерпретованість, метод чутливий до шумів і стандартизації даних. Є безліч реалізацій SVM. Найпопулярніші – це `scikit-learn`, `MATLAB` і зрозуміло `libsvm`.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Джерело вхідних даних

Автоматична класифікація емоційного забарвлення текстів з кожним роком стає все більш актуальним завданням і з теоретичної і з практичної точок зору. В першу чергу, це пов'язано з розвитком інтернету і зміною формату комунікацій в сучасному світі – для переважної більшості людей соціальні мережі стали займати лідируючу позицію серед інших джерел інформації та майданчиків для дискусій. Таким чином, користувачами соціальних мереж щодня генеруються значні обсяги текстової інформації.

Для отримання вхідних даних була обрана соціальна мережа Twitter, аудиторія якої, на кінець 2017 року, становила більш ніж 330 мільйонів активних користувачів. Англійська мова є найбільш часто використовуваною в Twitter. Це не дивно, оскільки сучасна соціальна мережа Twitter користується найбільшою популярністю в США і Великобританії. Японська мова являється другою мовою після англійської. Далі йдуть португальська та іспанська мови. Що стосується інших мов, то їх присутність на просторах Twitter не перевищує трьох відсотків. За результатами опитування, найбільша кількість користувачів Twitter – 107 мільйонів – живуть в США. У три рази менше резидентів з Бразилії. Далі, згідно з даними статистики, з невеликим розривом розташовуються Японія, Великобританія і Індонезія. Ще меншою популярністю Twitter користується в Індії і Мексиці. Україна та Росія входять до двадцятки країн за кількістю користувачів мережею. На основі вище наведених даних, можна зробити висновки, що для аналізу повідомлень, більше простору та варіантів дає англomовний сегмент соціальної мережі.

Оскільки в даній роботі розглядається аналіз тональності англomовних висловлювань в Twitter, то доцільно буде розглянути особливості таких висловлювань. Робота з англomовними твітами має деякі проблеми для обробки природної мови. Текстів в соціальних мережах більш характерний розмовний стиль

мовлення, ніж літературний. Так як твіти обмежені довжиною в 280 символів, то зазвичай вони містять в собі лише одну фразу, або пропозицію.

3.2 Архітектура програмної та структура даних

Обравши соціальну мережу, в якості джерела даних для аналізу, необхідно реалізувати взаємодію між додатком та Twitter.

Основна проблема полягає в тому, як отримувати ту інформацію, яка доступна в Twitter, і як саме її можливо використовувати для того, щоб зібрати корисну статистику. Для цього треба спочатку створити систему як зможе спілкуватися з соціальною мережею то отримувати від неї необхідні дані.

Твіттер дозволяє розробникам збирати дані за допомогою Twitter REST API та API потоку. Twitter має численні правила та граничні ліміти, накладені на його API, і з цієї причини він вимагає, щоб усі користувачі мали зареєструвати обліковий запис і надавати інформацію про автентифікацію, коли вони запитують API(рис. 3.1). Ця реєстрація вимагає від користувачів вказати адресу електронної пошти та номер телефону для підтвердження, після того як обліковий запис користувача буде підтверджено, користувачеві буде видано детальну інформацію про автентифікацію, яка дає доступ до API. Основними даними є публічний та приватний ключі, які є необхідними для верифікації запиту до API сервісу соціальної мережі. Такий підхід використовують для підтвердження ролі в системі, більшість систем які надають доступ до власного API, і Твіттер не виняток.

Використання соціальної мережі Twitter, в якості площадки для аналізу повідомлень, має свої переваги. Твіттер надає можливість проаналізувати як одне конкретне повідомлення так і виконати пошук за ключовими словами.

На жаль, в нативній реалізації, API Twitter експортує дані у форматі JSON, який потрібно перетворювати в формат зручний для аналізу або зберігання в базах даних. Для більш швидкої та зручної роботи доцільно використовувати стороні

реалізації бібліотек для взаємодії з TwitterAPI. Комбінація мови програмування C#, .Net Framework 4.6.2. та Nuget пакета linqtotwitter допомагають отримувати дані для аналізу не залучаючи додаткових ресурсів. Весь процес збору даних може бути повністю автоматизований за допомогою середовища розробки програмного забезпечення Visual Studio Community 2017, яке є безкоштовним, повнофункціональним інтегрованим середовищем розробки для учнів та студентів.

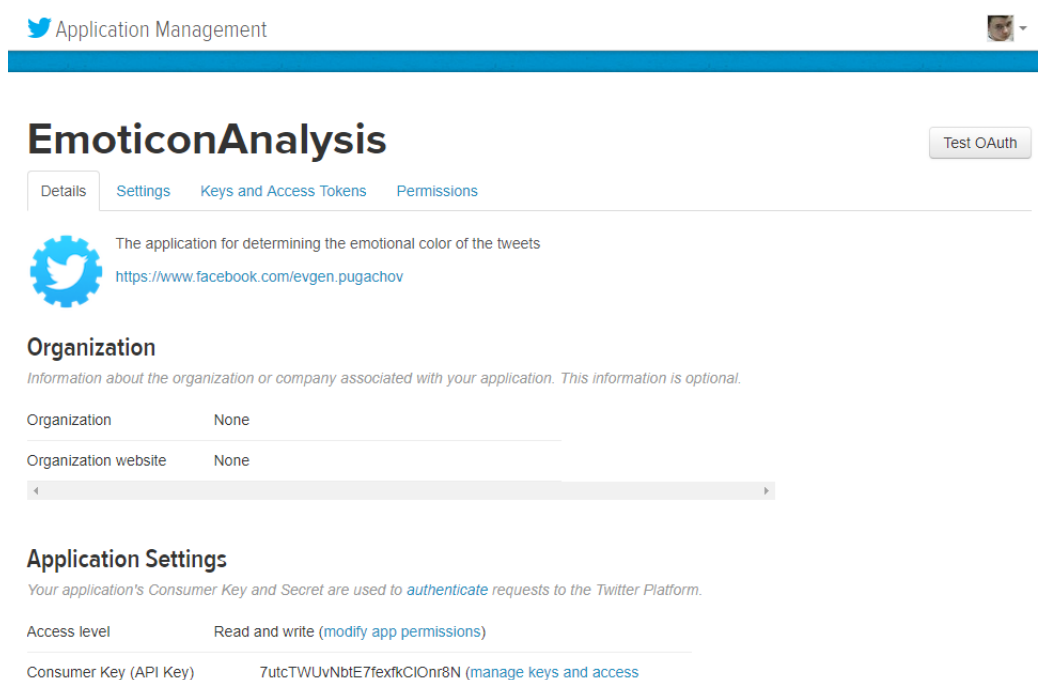


Рисунок 3.1 – Сторінка додатка для взаємодії с Twitter

Для написання цього продукту використовувалась мова програмування C# 6, яка є дуже зручна для написання прикладних програм та веб-сайтів та має дуже простий та зручний синтаксис, що значно полегшує розробку та підтримку програмного продукту.

Що стосується технології, то використовувалась технологія ASPMVC, яка дозволяє відокремити представлення від моделі та бізнес-логіки та розробляти або змінювати усі ці компоненти окремо, не переймаючись за те, що зміна одного з них спричинить несправність інших компонентів.

Реалізувавши шар доступу до даних для аналізу, настав час задуматися про їх обробку, та приведення до загального вигляду. Як вже було відмічено раніше,

способи спілкування в соціальних мережах сильно відрізняються від норм літературної мови. Зважаючи на те. Що такі повідомлення можуть містити в собі багато сленгових слів, скорочень, посилань, відміток користувача і помилок, Тому виникає потреба в попередній обробці даних. Відповідно, при попередній обробці твітів з соціальної мережі повідомлення необхідно попередньо підготувати. Посилання замінюються на рядок формату @link, згадки користувачів замінюються на @username. Це дозволить зробити текст повідомлення більш незалежними від зовнішніх факторів які можуть вплинути на кінцевий результат. Також варто замінити повторювані символи, послідовності однакових символів слід замінити на послідовність з двох таких же символів.

Перед обробка складається з:

- видалення знаків пунктуації;
- видалення цифр;
- видалення зайвих пробілів;
- видалення посилань;
- видалення символу хештег;
- видалення інших символів;
- приведення всіх слів до нижнього регістру.

Таким чином повідомлення *"We want to finish as well as we can and build a stronger #MUFC for next season."* – @AnderHerrera. Перетворюється в – *"we want to finish as well as we can and build a stronger mufc for next season."* – @username.

Після попередньої обробки повідомлень, проводиться обробка тексту з метою виділення інформації важливої для аналізу, яку можна розділити на 3 етапи:

- токенизація – це процес виділення з тексту окремих слів, чисел і знаків пунктуації;
- стемінг – це процес знаходження основи слова;
- обробка заперечень.

Мета стемінгу – приведення слів, що мають однакову основу до єдиної форми. Негативним моментом, є те що, після стемінгу втрачається частина морфологічної інформації, тому, як показали результати дослідження, використання стемінгу в

поєднані з наївним Баєсівським класифікатором для аналізу тональності не збільшує точність.

Вище описані методи обробки тексту, варто застосовувати і для навчальних вибірок, що допоможе підвищити точність результатів.

Досліджувані методи визначення тональності описані в даній роботі мають свою реалізацію і в додатку розробленому в рамках атестаційної роботи.

На рисунку 3.2 зображена внутрішня структура проекту. Дане рішення розроблено на основі трьох шарової архітектури, Хоча в даному випадку відсутній рівень Доступу до даних. Так як додаток не передбачає зберігання даних до бази даних, його наявність не є обов'язковою.

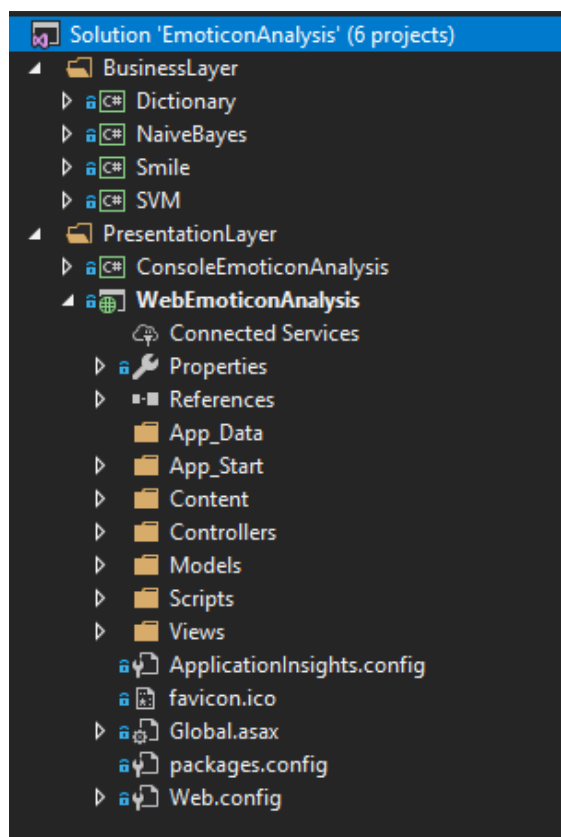


Рисунок 3.2 – Внутрішня структура додатку

Рівень даних, по суті, є сервером, що зберігає всі дані програми. Шар даних містить таблиці бази даних, файли XML та інші засоби зберігання даних програми. Бізнес-рівень працює як міст між рівнем даних і рівнем уявлення. Всі дані проходять через бізнес-рівень перед їх передачею рівню уявлення. Бізнес-рівень –

сума шару бізнес-логіки, шару доступу до даних, та інших компонентів, що використовуються для додавання бізнес-логіки. Рівень представлення – рівень, на якому користувачі взаємодіють з додатком. Рівень представлення містить загальний код інтерфейсу користувача.

Архітектура програми – це система рівнів, що забезпечує взаємодію внутрішніх функцій і процесів програми для досягнення необхідного результату.

Даний архітектурний підхід робить можливим паралельну розробку і тестування кожного з шарів, дозволяє змінювати структуру, а іноді навіть повністю замінювати один з шарів, не змінюючи при цьому інші шари. Можливість поділу місць функціонування шарів на фізичному рівні. Повна інтеграція 3х-шарової архітектури додатку з моделлю клієнт-сервер.

Такі архітектури більш розумно розподіляють модулі обробки даних, які в цьому випадку виконуються на одному або декількох окремих серверах.

Як видно з рисунку 3.2 шар Представлення має декілька проектів, консольний та веб. Консольний проект необхідний для отримання результуючих даних які будуть використовуватися для аналізу можливості використання отриманих результатів у науковій і практичній діяльності.

Веб-проект розроблений з метою надання можливості аналізу повідомлень з соціальної мережі Твіттер, якомога більшої кількості людей. Дане рішення представляє собою ASP.NET MVC рішення з використанням .Net Framework 4.6.2.

ASP.NET MVC Framework – фреймворк для створення веб-додатків, який реалізує шаблон Model-View-Controller. Згідно з яким веб-додаток ділиться на компоненти Model, View, та Controller. При цьому модель додатки, призначений для користувача інтерфейс і взаємодія з користувачем розділені таким чином, щоб модифікація одного з компонентів надавала мінімальний вплив на інші. Такий поділ полегшує управління окремими частинами програми, що спрощує їх розробку, зміна і тестування.

Завдяки обраним технологіям та архітектурним рішенням користувач взаємодіє з системою за допомогою веб сторінок а все, що відбувається далі скрито від нього. Бізнес логіка додатку, в даному випадку реалізація методів аналізу

тонального забарвлення тексту, знаходиться на окремому рівні і зовсім не залежить від того як і хто її використовує.

Повернувшись до рисунку 3.2 бачимо, що шар Бізнес логіки містить чотири проекти, відповідно до кількості досліджуваних методів, і кожен з них є самостійною одиницею та не залежить від інших.

Наприклад, метод на основі емотіконів, може бути використаним для визначення окрасу тексту, та навіть не здогадуватися про те, що інші методи зробили те саме. Даний метод надзвичайно простий в реалізації, в його основі лежить перелік емотіконів(рис 3.3), що необхідний для визначення тональності.

```
{
  "name": "slightly_frowning_face",
  "emoji": "😞",
  "polarity": -1
},
{
  "name": "slightly_smiling_face",
  "emoji": "😏",
  "polarity": 1
},
{
  "name": "smile",
  "emoji": "😄",
  "polarity": 2
},
{
  "name": "smile_cat",
  "emoji": "😸",
  "polarity": 2
},
}
```

Рисунок 3.3 – Частина емотіконів для аналізу в форматі JSON

Кожен емотікон має свою назву, символ яким його позначають в тексті та оцінку. Варто відмітити, що емотікони можуть позначатися по різному але мати одне й теж саме значення, ім'я. Загальна кількість емотіконів в переліку сягає 119 і з часом буде збільшуватися. Саме на основі імені емотікона і визначалась його полярність, що і допомогло об'єктивно оцінити емотікони. Шкала за якою класифікувалися об'єкти, розроблена автором роботи, та має лише цілі значення в

діапазоні від -3 до 3 включно.

Після того як метод отримує текст для аналізу, він завантажує список всіх доступних емотіконів для аналізу та намагається відшукати їх в тексті, врахувавши їх кількість та полярність метод побить висновок про загальне забарвлення тексту.

У випадку коли текст не містить жодного емотікона результатом буде повідомлення, що тональність визначити не вдалося.

В період розробки та відладки методу, виникла проблема, пов'язана з тим як відрізняти текст взятий в круглі дужки від послідовності з сумного та веселого смайлика.

Результатом стало рішення основане на аналізі загально прийнятих правил написання текстів, а саме, якщо за відкриваючою круглою смужкою, відразу слідує текст і першою наступною дужкою є закриваюча, і вона разом з сусідніми символами не входить до жодного з емотіконів, то вважається, що такий текст взятий в дужки і вони не впливають на кінцевий результат.

Метод на основі словників дуже схожий на метод в основі якого лежить пошук смайликів. Відмінністю, цього методу є те, що він враховує всі відомі для нього слова, аби зробити висновок про забарвленість тексту. Для аналізу метод використовує два словника, один зі словами, що мають позитивний окрас, інший містить негативно забарвлені слова. Словники негативних і позитивних слів містять 1597 і 879 слів негативно(рис. 3.4). Кожен словник оцінює слово від 0 до 5, де п'ять означає, що слово максимально негативне чи позитивне, в залежності від словника.

Кожне слово має свою оцінку яка враховується при визначенні кінцевого результату. Коли слово зустрічається в будь-якому зі словників, його оцінка додається до відповідної змінної, яка відповідає за підрахунок значень певної тональності.

//positive	//negative
true,2	loser,3
trust,1	losing,3
trusted,2	loss,3
unbiased,2	lost,3
unequaled,2	lowest,1
unified,1	lugubrious,2
united,1	lunatic,3
unmatched,1	lunatics,3
unstoppable,2	lurk,1
untarnished,2	lurking,1
useful,2	lurks,1
usefulness,2	mad,3
vested,1	maddening,3
vibrant,3	made-up,1
vigilant,3	madly,3
vindicate,2	madness,3

Рисунок 3.4 – Слова, що включає в себе словник позитивних слів

Метод опорних векторів та метод наївного Баєса засновані на машинному навчанні з учителем. Це означає, що для роботи цього типу алгоритмів нам буде потрібен учитель, який і буде вчити наш алгоритм. В даному випадку ми вже маємо вибірку реальних твітів(рис. 3.5) кількістю 17650 штук. Де кожен запис має свій порядковий номер, джерело з якого було взято текст та його класифікація, де 0 – означає, що текст має негативне забарвлення, а 1 – позначає текст як позитивно забарвлений.

Отримавши дані для навчання кожен з алгоритмів намагається їх проаналізувати. Базуючись на загальній оцінці тексту з нього необхідно виокремити певну приховану закономірність, яка відповідальна за розподіл даних в тренувальній вибірці, Завдяки віднайденню такої закономірності система зможе її використати для ефективного прогнозування відповідей на тестувальній вибірці.

Процес навчання є дуже ресурсомістким заняттям і потребує відносно великих часових і програмних затрат. На етапі планування системи постало питання, як зменшити кількість необхідних ресурсів та зробити систему менш залежною від процесу навчання. Було прийнято, що оптимальним виходом є варіант коли навчання відбувається лише одного разу, на момент першого

звернення до методу який оснований на машинному навчанні.

```
168,0,Sentiment140, ..had to turn back to North due to body on line. So missing Design Council Design 4 Tech Transfer event
169,0,Sentiment140, ..y everything is so hard? :/
170,0,Sentiment140, .bueno good bye. good night
171,1,Sentiment140, : Beach day with Scoobs. =( : Still no phone.
172,0,Sentiment140," : Let's have fun,. When I give ya what I give ya."
173,0,Sentiment140, :[ I don't want to move!!!
174,1,Sentiment140," :-D )))))))...WHAT AN AMAZING NIGHT,DAY & NIGHT AGAIN!! HI TWITS! I MISS U GUYS"
175,1,Sentiment140, @ canaveral national seashore
176,1,Sentiment140, @ progressing in the production department with The Uprizing
177,1,Sentiment140, @ taylorrhicks enjoy chicago..b new venue is always cool. You shine everywhere you go. I know who you are.
178,0,Sentiment140, @ the train crash in DC .....
```

Рисунок 3.5. – Вхідні дані для машинного навчання

Так як, система має декілька шарів, і в майбутньому, може надавати публічний API, оптимізацію необхідно виконувати на рівні, який не залежить від джерела запиту, а саме на рівні бізнес логіки.

Для вирішення проблеми оптимізації, необхідне рішення яке, дасть нам можливість мати одночасно лише один об'єкт та мати до нього глобальний доступ. Ідеальним рішенням для цього випадку є патерн – одинак (від англ. Singleton)(рис 3.6).

Даний шаблон проектування, відноситься до класу твірних шаблонів. Він гарантує, що клас матиме тільки один екземпляр, і забезпечує глобальну точку доступу до цього екземпляра.

В даному випадку, глобальна змінна не вирішує такої проблеми, бо не забороняє створити інші екземпляри класу, що неодмінно потягне за собою процес навчання.

Рішення полягає в тому, щоб сам клас контролював свою «унікальність», забороняючи створення нових екземплярів, та сам забезпечував єдину точку доступу. Це є призначенням шаблону одинак.

Розглянувши рисунок 3.6 більш детально можна побачити, що для відкладеної ініціалізації синглтона використовується статичний конструктор. CLR автоматично викликає конструктор типу при першому зверненні до типу, при цьому забезпечуючи безпеку щодо синхронізації потоків. Конструктор типу автоматично генерується компілятором і в ньому відбувається ініціалізація всіх

полів типу.

```
public class SmileMethod
{
    private static readonly Lazy<SmileMethod> instance = new Lazy<SmileMethod>(() => new SmileMethod());
    private static List<EmoticonModel> emoticons;

    private SmileMethod()
    {
        emoticons = LoadEmoticons();
    }

    public static SmileMethod Instance
    {
        get { return instance.Value; }
    }

    public string Analyze(string message) ...

    private List<EmoticonModel> LoadEmoticons() ...
}
```

Рисунок 3.6 – Варіант реалізації патерну одинак

Кожен з чотирьох методів реалізує даний підхід, незалежно від того пов’язан він з машинним навчанням чи ні. Наприклад, імплементація патерну в методі на основі словника, дає змогу виконувати завантаження словників лише один раз, що позитивно впливає на працездатність системи. Відповідно, метод наївного Баєса використовує даний підхід, аби виконувати процес навчання лише один раз, в момент першого звертання.

Одним з варіантів, який може вирішити проблему повторного навчання є спосіб при якому вже навчена модель буде зберігати до бази даних та братися звідти при необхідності. Але такий підхід вимагає наявності бази даних та виконання запитів до для кожного тексту. Цього можна позбутися використовуючи, вже вище описаний патерн один.

На рисунку 3.7 відображається час затрачений на виконня функції з аналізу тексту всіма доступними методами. Детально процес аналізу та можливості систему будуть розглянуті в наступному розділі. Ці данні є суб’єктивними, адже в залежності від середовища виконання та інтернет з’єднання результати можуть змінюватися.

Числа на рисунку 3.7 представляють собою час, в мілісекундах, необхідний

для аналізу текстів. Перший запит на аналіз тексту, зайняв майже втричі більше часу ніж всі наступні. В цей час ініціалізувалися є класи для аналізу, завантажувалися дані, виконувався процес машинного навчання з учителем та відбувався сам процес аналізу текстів. Завдяки вибраному рішенню всі наступні рази, в процесі навчання вже не було необхідності, кожен раз використовувалась же навчена модель, що зменшує час роботи програми втричі.

35,01
10,14
10,3
10,35
10,51
11,29

Рисунок 3.7 – Час виконання аналізу текстів

Незважаючи на описані досягнення в реалізації даної системи, вона має і певні недоліки. Постійно потрібно підтримувати список емотіконів та словники в актуальному вигляді.

3.3 Опис програмної системи

Розроблена система, на основі досліджуваних методів, представляє собою веб–додаток з можливістю аналізу тексту з різних джерел та різними методами(рис 3.8).

Веб–система Emoticon Meaning створена з урахуванням всіх сучасних вимог до дизайну, інтерфейсу та взаємодії з користувачами, аби нові користувачі могли з легкістю оволодіти системою.

Дизайн системи передбачає використання світлих та яскравих кольорів, які роблять сайт приємним для сприйняття, та не відволікають від основного контенту

додатка.

Emoticon Meaning

Search tweet
star wars

Tweet Url
1002898877597577217

Message
It was a good day

Just smile ☐ Analysys only by smile

Dictionary ☒ Use dictionary method

SVM ☒ Use SVM method

Naive Bayes ☒ Use Naive Bayes method

Start

© 2018 - Emoticon Analysis

Рисунок 3.8 – Головна сторінка програми

Система надає можливість користувачу, обрати якими саме способами необхідно проаналізувати текст. Серед них метод опорних векторів, метод наївного Баєса, методи на основі словників і емотіконів, кожен з них можна підключити або відключити за допомогою відповідного чекбокса.

Дані для аналізу можуть бути отримані декількома різними способами.

Один з них безпосередньо пов'язаний з використанням Twitter API. Система дозволяє виконувати пошук в соціальній мережі за пошуковим запитом введеним в поле Search tweet. Система звертається до Twitter API, щоб отримати статуси які містять в собі пошукову строку. Варто відмітити, що Твіттер використовує слова статус замість звичного нам твіт. Завдяки можливості виконувати пошук серед твітів та робити їх аналіз, дана система може використовуватися для визначення відношення людей до певного об'єкта, чи то ресторан, фільм або публічна персона.

Окремо аналізуватися може і певний твіт з соціальної мережі. Для цього необхідно лише скопіювати посилання на нього з адресної строки браузера або просто його унікальний ідентифікатор та вставити до відповідного поля.

Існує можливість проаналізувати окремо текст повідомлення, без будь-якої

прив'язки до соціальної мережі.

Після заповнення всіх полів для початку аналізу необхідно натиснути кнопку Start. Після чого з'явиться сторінка з результатами роботи аналізатора(рис. 3.9).

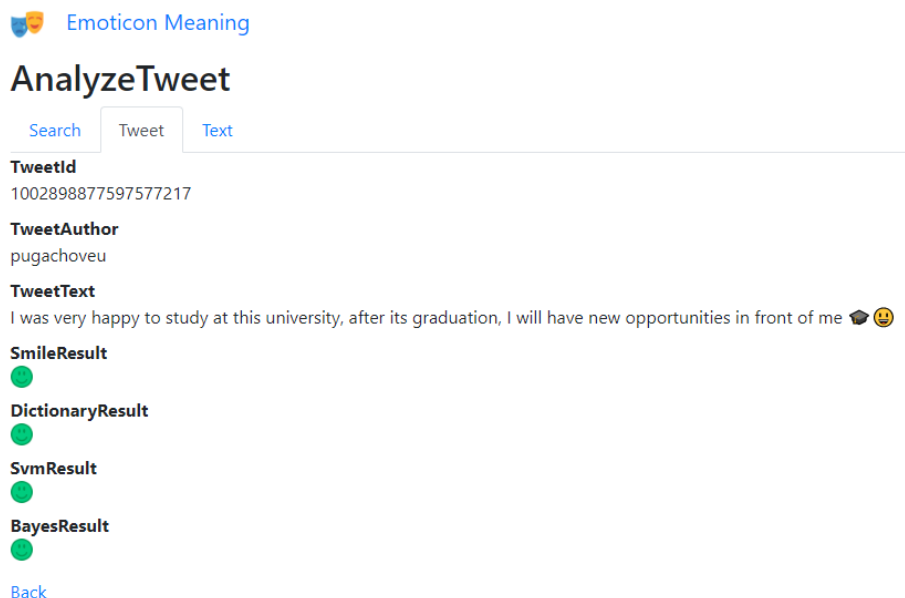


Рисунок 3.9 – Результати аналізу повідомлення

Сторінка результатів містить три вкладки, відповідно до кількості джерел текстів, що використовувалися для аналізу. На рисунку 3.8 зображені результати аналізу повідомлення з соціальної мережі за його унікальним номером. В результаті можливо дізнатися хто є автором повідомлення його текст та оцінку за кожним методів дослідження.

Оцінка тексту представлена за допомогою емотікона. Зелений смайл з посмішкою, говорить, що текст є позитивно забарвленим. Червоний сумний смайл вказує на негативне забарвлення повідомлення, а жовтий емотікон, використовується лише в випадку з нейтрально забарвленим висловлюванням. Нейтральний окрас можливий лише в результаті роботи метода з емотіконами.

Сторінка результатів аналізу вхідного тексту на забарвленість майже повністю ідентична сторінці на рисунку 3.8 і детально не розглядається.

На рисунку 3.10 приводяться результати аналізу твітів отриманих в результаті пошуку за ключовою фразою.

Кожне повідомлення зі знайденого переліку містить в собі підстроку яка відповідає тій, що була введена раніше. Всі записи, також, мають автора та результати оцінки за різними методами.

 Emoticon Meaning

AnalyzeTweet





























Search	Tweet	Text				
Tweet Author	Text	Smile	Distionary	SVM	Naive Bayes	
thepodcouple	RT @NerdyThingsPod: 🌟 NEW EPISODE🌟 This week we talk Pokémon news, #TWD, AND OUR MAAAAIN TOPIC is Solo: A Star Wars Story!! We go talk what...					
BrothersBinge	RT @kiersmclean: REVIEW! #SoloAStarWarsStory is a blast from beginning to end! @RealRonHoward and everyone involved did a great job. #StarW...					
one_abud	RT @heroichollywood: 'Star Wars': Ron Howard Retweets 'Solo' Praise That Bashes 'Last Jedi' https://t.co/J5EkmMcebM https://t.co/BOPx51P6J5					
jedifarfy	"Here's this detail that everyone noticed 10 years ago when this book/movie/song was released and I just noticed!" = 10k upvotes. Fandoms on reddit, or why I've had to unsub from things like Harry Potter and Star Wars. :)					
ManuelDuarte24	Won't even dare to check the comments on the latest post on the Star Wars page.					
T1meTraveler01	RT @Lost789Boy: Ways the world will implode due to an overdose of haterade: 1) The Fountainhead directed by Zack Snyder nominated & earnin...					
Warmustbeend123	RT @heroichollywood: 'Star Wars': Ron Howard Retweets 'Solo' Praise That Bashes 'Last Jedi' https://t.co/J5EkmMcebM https://t.co/BOPx51P6J5					

Рисунок 3.10 – Результати аналізу твітів з результатів пошуку

Оцінки аналізу, представляють собою зображення емотіконів. Для кожного повідомлення, також, визначається загальний результат дослідження, який враховує результати всіх досліджуваних методів.

Пошук твітів за наявністю в них певних слів, може бути корисним для маркетологів, політиків та тих кому важливо знати відношення людей до якогось об'єкта чи події.

В результаті, розроблена програма, має легкий та зручний інтерфейс, та водночас має всі основні функції для того аби стати успішним продуктом на ринку програм для аналізу тональності.

3.4 Результати

Дослідження методів сентимент-аналізу емоційного забарвлення текстів на прикладі даних з соціальної мережі Твіттер призвело до реалізації прототипу веб-системи для аналізу тональності повідомлень. Тональність повідомлення визначалася за двома категоріями: позитивна та негативна.

Для тестування алгоритмів визначення тональності текстів рецензій був використаний метод кросс-валідації, або, по-іншому, перехресної перевірки. Процедура кросс-валідації виконується наступним чином:

- Фіксується безліч розбиття навчальної вибірки на власне навчальну та тестову;
- Для кожного розбиття відбувається навчання алгоритму на навчальній підвибірці і тестування на тестовій;
- Результатом перехресної валідації алгоритму є середні значення оцінок ефективності для тестових підвибірок.

На базі навчальної вибірки з 17650 англомовних повідомлень з соціальної мережі було проведено аналіз 400 повідомлень. Тестові повідомлення не використовувалися в процесі навчання. Значення співвідношення позитивних і негативних твітів у навчальній і тестовій вибірці було порівняно однакове. Відхилення складало не більше 20 твітів, в будь-яку зі сторін, що на такій кількості навчальних і тестових значень не є суттєвим фактором.

В результаті тестування програми на тестовій вибірці, була зібрана велика кількість інформації, аналіз якої допоможе краще зрозуміти, залежності між використанням емотіконів в повідомленнях та справжнім забарвленням тексту. Також є змога визначити комбінацію підходів для сентимент аналізу яка дає найбільш точний результат

В таблиці 3.1 наведені результати роботи класифікатора з аналізу емоційно забарвлених повідомлень. У кожного повідомлення є реальна тональність тексту та загальна. Загальна тональність тексту визначається сумою тональності всіх методів. Метод визначає негативну тональність як -1 , позитивну як $+1$, нейтральна тональність, відповідно має значення 0 . Три з чотирьох методів класифікують текст, лише, як позитивний або негативний. Метод на основі емотіконів може

визначити тональність тексту як нейтральну, у випадку, коли ньому не було знайдено жодного емотікона. В результаті, визначення загальної тональності повідомлення може відбуватися за допомоги складання результатів всіх методів.

Таблиця 3.1 – Результати роботи класифікатора

#	Повідомлення	Визначена загальна тональність	Реальна тональність
1	not feeling good...sick	–	–
2	just finished mowing the lawn...feeling human again!	–	+
3	Is excited to start packing tonight!!	+	+
4	Why wont the chatroom work :'(–	–
5	Is there any problem to access the website or it's just me? I can't go to Flickr for 3 days in a row!	–	–

Як видно з вище наведеної таблиці, результати роботи класифікатора мають дуже близький результат до істини. Лише в другому випадку класифікатор зробив помилку, та визначив загальний окрас тексту як негативний. Помилка може бути в'язана з роботою словників та методів машинного навчання.

Для отримання більш точних результатів роботи класифікатора, у якості тестових даних використовувалась вибірка з 5000 повідомлень, кожне з яких було взято з корпусу розмічених твітів і не використовувалося в процесі навчання. з соціальної мережі Twitter.

Загальні результати роботи аналізатора наведені на рисунку 3.11 у вигляді діаграми. Всі повідомлення піддавались аналізу кожним з методів і для кожного з них визначалась загальна тональність. Найкращий результат показав метод на основі словників. Результати, близькі до найкращого, також показали методи машинного навчання без учителя. Хоча значення і далекі від ідеальних. Такі результати можна пояснити вмістом орфографічних помилок і сленгу.

Поєднання результатів методів, для отримання єдиного результату,

однозначно має позитивний вплив. Адже, як видно з діаграми, загальний результат має більш високу точність в порівнянні з іншими методами.

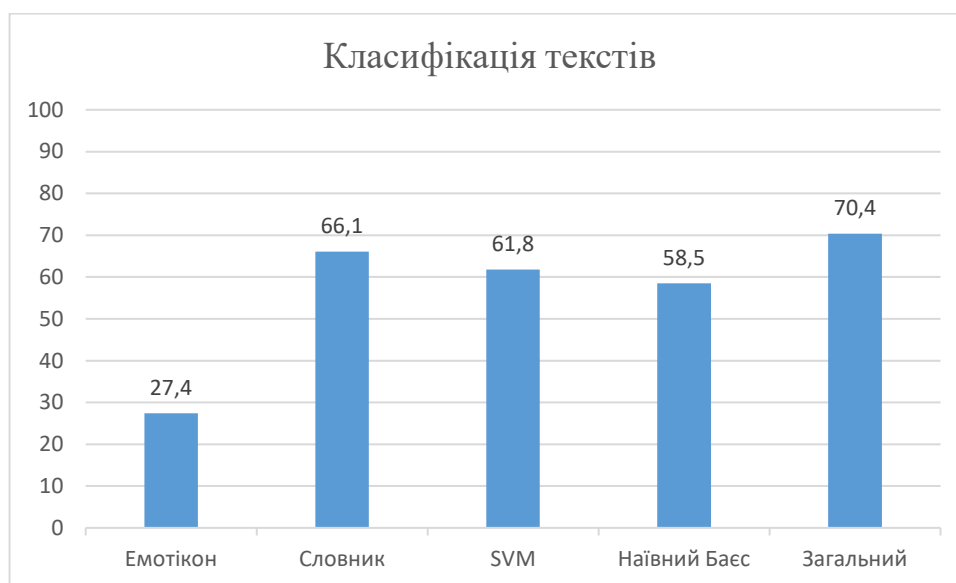


Рисунок 3.11 – Результати роботи аналізатора

Нажаль, метод з використанням емотіконів, має найгірший результат з усіх. Значення точності визначення тональності тексту майже вдвічі гірше, порівняно з наївним Баєсом. Такі дані можна пояснити особливостями роботи алгоритму та вхідних даних. Тестові дані мають лише два можливих значення, серед яких позитивний та негативний окрас. А метод класифікації на основі емотіконів може визначити тональність тексту як нейтральну, у випадку коли текст не містить жодного смайла. Таким чином метод, робить хибний висновок, хоча фактично цей результат був отриманий без використання емотіконів.

Для отримання коректних результатів роботи алгоритму слід використовувати тестову вибірку, в якій кожен запис містить емотікон. В результаті, була сформована вибірка з 500 записів, кожен з яких є коректним для перевірки роботи метода на основі емотіконів.

Результати аналізу такої вибірки є схожими з даними отриманими в результаті попередніх тестувань. Метод на основі емотіконів показує кращий результат в порівнянні з іншими трьома методами (рис. 3.12). В результаті, також, покращився результат методу визначення загального результату. Інші

класифікатори залишилися на попередніх значеннях, з невеликим відхиленням

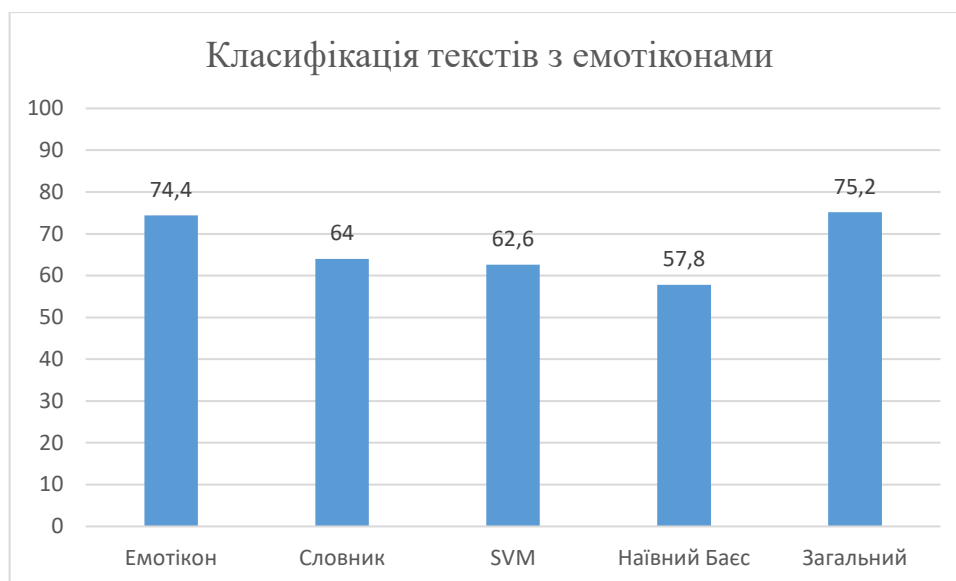


Рисунок 3.12 – Результати аналізу текстів з емотіконами

Розглянувши результати більш детально, можна зробити висновок про залежність між використанням емотіконів і тонального забарвлення тексту.

Користувачі соціальної мережі, використовують смайли з метою вираження своїх емоцій. Згідно отриманих результатів в, майже, 75 відсотках випадків забарвлення використаних емотіконів в тексті, повністю відображає його тональне забарвлення.

Тестуючи розроблену програмну систему на різних вибірках, були зроблені заміри часу роботи системи з різною кількістю даних. Результати наведені на рисунку 3.13. Час витрачений на аналіз повідомлень вимірювався в мілісекундах. Одна секунда містить в собі одну тисячу мілісекунд. Варто відмітити, що вимірювався час затрачений лише на аналіз тексту, включаючи його попередню обробку, і не враховувався час необхідний на отримання тестового корпусу текстів. Моделі які використовуються в методах з використання машинного навчання, були створені та навчанні попередньо. Також, не враховувався час на відображення результатів аналізу, адже це може бути консольний вивід даних, збереження до бази даних, запис до файлу в файловій системі, відображення на веб-сторінці.

Час витрачений на аналіз перших 100 текстів близький до двадцяти

мілісекунд. Класифікація п'яти ста текстів зайняла менше ніж сто мілісекунд. Десять тисяч записів аналізувалися майже дві з половиною секунди.

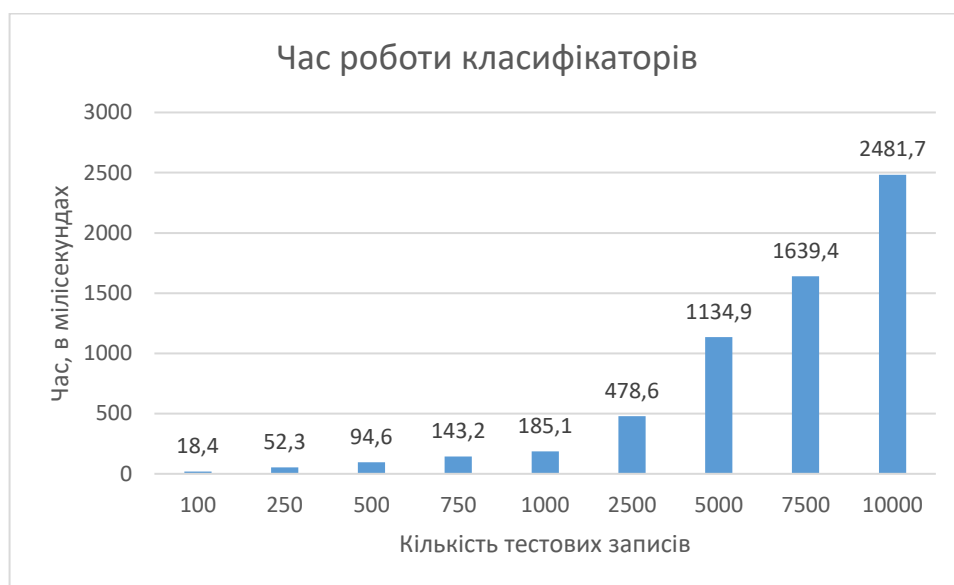


Рисунок 3.13 – Час роботи класифікаторів

Проаналізувавши дані з таблиці вище, можна виділити закономірність, кожні 100 повідомлень потребують приблизно двадцять мілісекунд. Це значення може коливатися в залежності від довжини повідомлень в соціальній мережі Твіттер, адже максимальна довжина таких повідомлень може бути 240 символі, середня кількість символів в повідомленнях складає близько ста.

Проведене тестування системи з різними за кількістю та змістом повідомленнями, виявило, що точність роботи алгоритму становить до 75% при аналізі повідомлень, що містять емотіконі і 70% при довільно вибраних повідомленнях. Для тестування точності визначення емоційного забарвлення повідомлень було проведено понад 10000 запитів до джерел, що містить як позитивно, так і негативно забарвлені повідомлення. Отримані результати, безумовно, не вирішують проблему визначення емоційного забарвлення текстів, так як завдання визначення тональності неймовірно складна, сильно залежна від предметної області і внаслідок цього об'ємна. Однак, дослідження проблем сентимент-аналізу в соціальних мережах може вплинути на вдосконалення систем соціологічних досліджень, фільтрації небажаного контенту, пошуку повідомлень,

що містять погрози й т.д.

Дослідження на основі даних соціальної мережі Твіттер особливо цікаві, тому що мережа є однією з найбільших в світі з аудиторією близько 330 млн. користувачів. Методи аналізу даних з соціальної мережі можуть також стати кроком до створення принципово нових автоматизованих соціологічних і маркетингових досліджень тональності в конкретній предметній області.

Для підвищення ефективності програми, можливо, слід доповнити існуюче рішення елементами лінгвістики. Лінгвістичні підходи дозволяють визначати тональність частин тексту, тим самим істотно збільшуючи точність класифікації для окремого тексту – адже, часто, в повідомлення згадуються і позитивні і негативні сторони об'єкта тональності.

Обробка заперечень Для збільшення точності був використаний алгоритм обробки заперечень, описаний Десом і Ченом. Суть його полягає в наступному: при появі частки «not» до кожного слова між цією частиною і подальшим знаком пунктуації або іншою частку "not" приписується приставка "not_". Наприклад, фраза «Мені не сподобався цей фільм.», перетвориться до виду: «Мені не not_сподобався not_цей not_фільм.»

Крім цього, варто виключити зі словника слова, що не мають емоційного забарвлення, тобто незначущі. Цілком можливо, що кращі результати дасть витяг ключових слів. Далі, для зменшення словника, також варто проводити стемінг, приведення слів до їх основ.

4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ

Соціальні мережі, такі як "Twitter" в наші дні стали одним з основних засобів спілкування. Велика кількість інформації, що міститься на цих сторінках цих ресурсів, робить їх привабливим джерелом даних для аналізу думки та аналізу настроїв. Більшість текстових методів аналізу не можуть бути корисними для аналізу настроїв у цих випадках. Для досягнення значного прогресу нам потрібні нові ідеї. Використання імен твіттера та хеш-тегів для збору тренувальних даних може забезпечити кращі результати. Також додавання аналізу символів за допомогою смайлів та символів емоції може значно підвищити точність розпізнавання емоцій. Найбільш успішними алгоритмами буде, мабуть, інтеграція методів обробки природних мов та аналізу символів.

Широта охоплення аудиторії в мільйони чоловік і оперативність отримання інформації, яка доступна практично в режимі реального часу, дозволили отримувати недосяжні раніше результати досліджень. Якщо раніше, щоб виявити думку з будь-якого питання, потрібно було проводити опитування, то сьогодні висловлювання по величезній кількості популярних тем вже є в мережі, треба тільки виявити їх, розпізнати і оцінити.

Технологія сентимент аналізу знайшла широке комерційне застосування у корпорації – власників брендів для аналізу соціальних медіа. Сучасні програми надають можливість не тільки оцінити тональність висловлювань про бренд, а й отримати цілий ряд додаткових інструментів, що спрощують управління соціальною аудиторією, яка цікавиться брендом, встановлення контактів, обмін інформацією, вплив на вирощування соціального контенту, пошук лідерів думок соціальної спільноти, постачання їх інформацією і залучення до просування бренду[23].

Серед клієнтів, особливо це стосується клієнтів технологічних фірм, багато бажаючих залишатися на зв'язку з виробниками продуктів, якими вони

користуються. У їх числі є такі, до думки яких прислухаються всі інші клієнти. Саме їх треба виявити і підтримати. Є попит – є пропозиція. На базі платформи, вирішує завдання визначення тональності тексту, будуються системи, що виконують цілий ряд прикладних задач, таких як моніторинг соціальних медіа, визначення майданчиків, на яких обговорюється бренд, оцінка того, яку думку виражається, аналіз змісту цих розмов, а також засоби управління мережевий активністю в соціальних медіа. Таким чином, цілий ряд рішень забезпечує не тільки оцінку тональності, а й підтримку клієнтів, зв'язок з соціальною громадськістю, дослідження ринку і вимір результативності маркетингових кампаній.

Особливий інтерес для компаній представляє залучення лідерів думок – тобто людей, які користуються особливим впливом в соціальних мережах. Як правило, це люди з активною життєвою позицією, яким подобається бути не просто слухачами, а активними учасниками дискусій. Це люди з широкою мережею контактів в офлайновій і онлайн-середовищі, вони люблять вчитися і знайомитися з новими технологіями і продуктами, використовують різні джерела інформації, щоб бути в курсі всіх подій, формують свою власну думку. Ці люди не тільки навчаються самі, а й виявляють зацікавленість в поширенні своїх знань і корисних порад. Для представників брендів дуже важливо залучати подібних людей. Наприклад, надавши лідеру думки можливість самостійно випробувати новий товар або послугу, можна очікувати, що інформація про нього буде донесена до великої кількості людей. Негативні відгуки лідерів теж можуть бути дуже корисні для власників брендів, для удосконалення продукту, і досить небезпечні для них, якщо критика має форму не поради щодо вдосконалення продукту, висловленого в рамках приватної бесіди, а публічної скарги. Тому дуже важливо встановлення діалогу і управління процесом донесення оцінки лідерів думок до середовища соціальних медіа. Лідери думок можуть виступати як генератори новацій і вдосконалень продукту, і тривала робота з ними може бути досить плідною.

На даний момент, було розглянуто лише один з аспектів аналізу збору інформації на базі соціальних медіа. В принципі, інформація від клієнтів і

потенційних і наявних споживачів продуктів компанії, інтегрована з іншими корпоративними системами, є основою для підтримки прийняття рішень в корпорації на різних рівнях управління та маркетингової стратегії компанії.

Аналізу тональності тексту може застосовуватися, навіть, для вирішення такої актуальної проблеми, як забезпечення безпеки користувачів в Інтернеті. За допомогою sentiment аналізу можливо визначити потенційно небезпечних осіб з числа користувачів соціальної мережі[24].

Чимало критики викликає і вторгнення подібних систем в особисте життя. Якщо система, якимось чином, буде спостерігати і аналізувати особисту переписку користувачів, це може викликати не тільки їх занепокоєння, але і привести до судових позовів. Хоча системи і так використовують лише публічно доступну інформацію в аналізі.

Чудовим прикладом використання автоматизованого sentiment – аналізу в реальному часі може бути аналіз президентської передвиборної кампанії. Так, під час передвиборної гонки 2012 року в США, щоб продемонструвати роботу аналізатора емоцій, відстежувалися всі твіти, в яких згадувалися основні кандидати в президенти. Однозначно, результати роботи такого інструменту виходять за рамки простої оцінки «позитивно» або «негативно». Такий аналізатор дає детальний розбір емоційного профілю досліджуваної аудиторії та допомагає зрозуміти розподіл голосів виборців

Для перевірки працездатності розроблених методів аналізу тональності повідомлень, виконувалась перевірка в режимі реального часу. В якості предмету аналізу був вибраний фінал Ліги Чемпіонів 2018 року та відношення вболівальників кожної з команд до подій на полі. Під час перевірки аналізувалися твіти на їх тональний окрас. Головним питанням було те як, правильно розпізнати вболівальником якої команди є автор повідомлення, а вже потім аналізувати його окрас.

Для отримання даних в режимі реального часу, під час фінального матчу, виконувалися запити до Twitter API з інтервалом в п'ять хвилин. Цього часу було достатньо аби проаналізувати попередньо отримані твіти та зберегти статистику.

Відповідно за цей час користувачі встигали висловити свої думки в соціальній мережі відносно подій на стадіоні.

Для пошуку повідомлень, що відносяться до футбольного матчу і пов'язані з певною командою, кожні п'ять хвилин виконувалося два запити для пошуку твітів, кожен з яких мав на меті отримати твіти вболівальників певної з команд. Які потім аналізувалися. Кожен запит мав ключові слова які стосувалися фіналу, аби отримати твіти, що відносяться до фіналу Ліги Чемпіонів. Запити відрізнялися лише ключовими словами які містили назву команди. Таким чином формувалися дві колекції повідомлень, автори яких висловлювали думку про одну з команд. Всі отримані тексти є англomовними. Їх аналіз допоможе відтворити хід матчу(рис. 4.1).

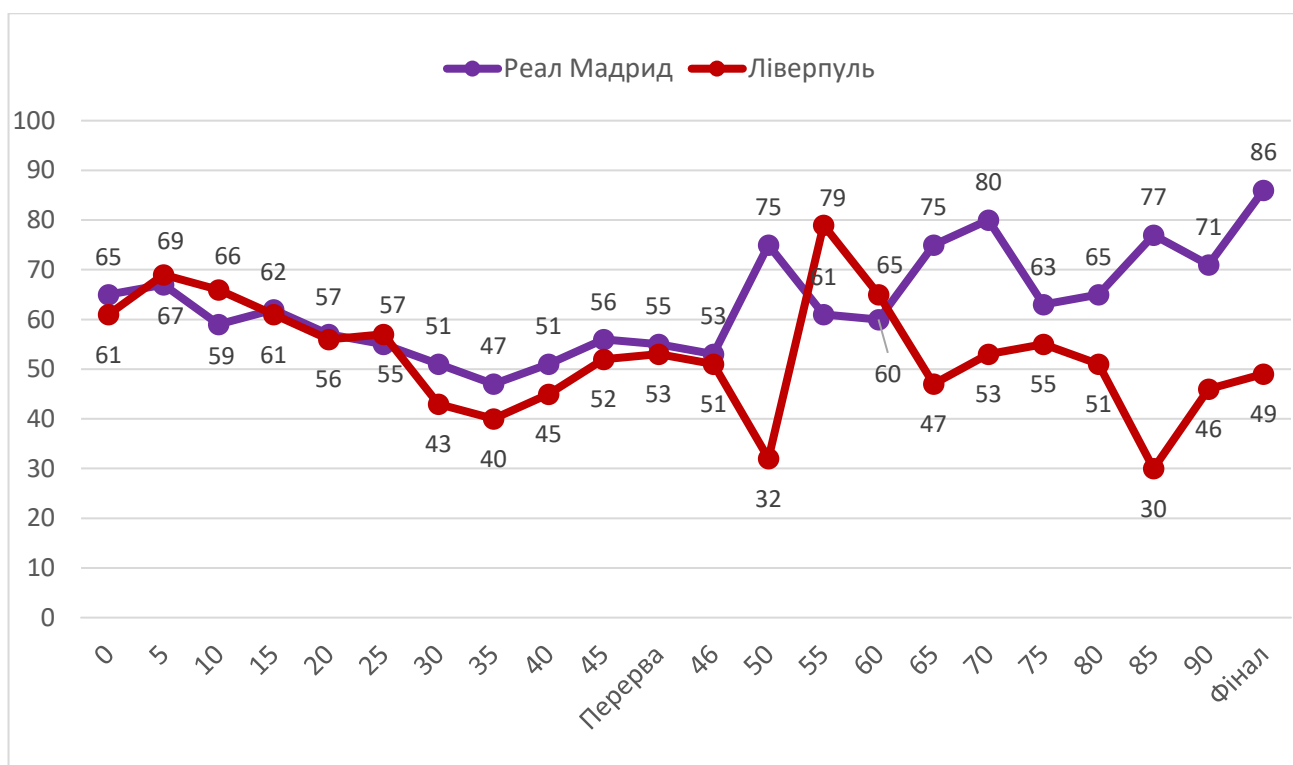


Рисунок 4.1 – Аналіз емоцій, виражених в Твіттер під час фінального матчу UEFA Champions League 2018

Дивлячись на графік, легко визначити емоції вболівальників, які спостерігають за ходом гри Реал Мадрид – Ліверпуль.

Для того щоб оцінити можливість застосування методів і точність роботи

системи в реальному часі, цікаво порівняти графік на рис. 4.1 і динаміку матчу:

- 30-та хвилина матчу – Травма гравця футбольного клубу Ліверпуль;
- 35-та хвилина матчу – Заміна травмованого гравця команди Реал Мадрид;

- 50-та хвилина матчу – Реал Мадрид виходить вперед 1–0;
- 55-та хвилина матчу – Ліверпуль зрівнює рахунок 1–1;
- 60-та хвилина матчу – Забиває Реал Мадрид 2–1;
- 85-та хвилина матчу – Гарет Бейл робить дубль 3–1;
- Фінал – Реал Мадрид виграв з рахунком 3–1.

Маючи інформацію про результати матчу в часі, можна припустити сплески емоцій фанатів, які будуть відображатися в Твіттері. Ми бачимо, що іспанські вболівальники проявили найбільше емоцій в кінці гри. Очевидно, що автоматизований аналіз Twitter повідомлень дає дуже чітку картину матчу. Так що твіти, як і інші пости соціальних медіа, можуть досить точно відображати настрої і думки аудиторії. За допомогою Twitter-потоків можна отримувати дані в реальному часі і використовувати їх для прийняття оперативних бізнес і політичних рішень.

Соціальні мережі є місцем де люди можуть висловити свою думку, про будь-яку подію в своєму житті. Це може бути подорож в іншу країну, завершення читання книги, покупка нового телевізора або похід до нового ресторану. Все це, однозначно цікаво для тих людей, котрі працюють з думкою людей, наприклад маркетологів. Часто обираючи черговий фільм для перегляду, ми базуємо свій вибір на відгуках тих людей які вже його переглянули. Це може бути відгук на кіно форумі або коментар на сайті прокатної компанії. Але більшу кількість рецензій на фільм, завжди, можна знайти на сторінках соціальних мереж.

Так, наприклад, вирішуючи чи варто йти до кінотеатру на перегляд нової стрічки, можна скористатися результатами досліджень проведених в цій роботі та розробленим програмним забезпеченням. Виконавши пошук за назвою фільму "Solo: A Star Wars Story", можна отримати останні повідомлення в яких фігурує назва фільму. Визначивши їх тональність, можна стверджувати про те як аудиторія відреагувала на фільм. Такий підхід може бути корисним не лише глядачам, а й

творцям фільму. В таблиці 4.1 наведенні результати аналізу твітів які містять назву фільму.

Таблиця 4.1 – Результати аналізу відгуків про фільм

Кількість повідомлень	Позитивні	Негативні
156	113	43

На основі вище наведених даних, можна сказати, що більше семи десяти відсотків респондентів оцінили фільм позитивно, і лише мала частина висказала своє невдоволення. Дані для аналізу отримувалися за допомогою пошуку повідомлень які містять назву фільму, на наступний день після прем'єри.

Підсумувавши всю вище наведену інформацію та результати досліджень, можна зробити висновок, що використання сентимент аналізу в масштабах соціальних мереж дає нові, великі можливості для визначення думок користувачів. Однозначно, сучасні рішення не повністю відповідають тим вимогам які до них висуваються і їх результати можуть бути суперечливими, але експерти покладають чималі надії на нейронні мережі і машинне навчання.

ВИСНОВКИ

В ході дослідження методів емоційного аналізу тексту з емотіконами були досліджені методи класифікації тексту за допомогою машинного навчання та на основі словників. Був запропонований метод визначення тональності, що базується на емотіконах.

В результаті виконання атестаційної роботи була створена програма для автоматичного аналізу тональності повідомлень в англomовному сегменті соціальної мережі Twitter з урахуванням емотіконів.

Всі компоненти програмного продукту розроблені з використанням мови програмування C#, технологій .NetFramework 4.6.2, ASP.NET MVC та середовища розробки Microsoft Visual Studio 2017. Розроблена система має інтеграцію з соціальною мережею. Дозволяє виконувати пошук твітів або проаналізувати певне повідомлення.

Інтерфейс задовольняє всім принципам usability зручний та простий у використанні.

Попри виконання поставленої задачі, в проекті ще є напрями для подальшої розробки. Перш за все, необхідно розширити перелік мов, з якими може працювати аналізатор. Для методів які використовують машинне навчання слід додати можливість збереження вже навчених моделей. В майбутньому необхідно створити мобільну версію клієнту, для найбільш популярних мобільних операційних систем. Також потрібно додати можливість інтеграції з іншими популярними соціальними мережами.

В ході роботи було створено класифікатор який базується на використанні всіх досліджуваних методів. Вибрані відповідні метрики і проведені розрахунки ефективності класифікації шляхом тестування методом кросс-валідації. Встановлено, що точність класифікації порівнянна з точністю сучасних аналогів. Визначено, що в більшості випадків, емоційне забарвлення використаних емотіконів прямо пропорційне до загальної тональності тексту.

Проте, залишається ряд невирішених питань, такі як опрацювання помилок, використання хеш тегів та сленгових слів. Також проблемою є обмеженість використовуваного емотивного простору. Зазвичай використовуються частина лексики, добре–погано плюс сила емоційності. Таким чином, якісне поліпшення запропонованого методу визначення тональності потребує подальших фундаментальних дослідженнях не тільки в галузі лінгвістики, а й в області когнітивних наук, таких як психологія, психо і нейролінгвістика.

Високий інтерес до відкритих тестуванням систем в області аналізу тональності підтверджує актуальність вирішуваних завдань і їх затребуваність в системах обробки інформації. Вектор розвитку при вирішенні завдання аналізу тональності лежить в більш детальному аналізі текстів про об'єкти і їх атрибутах, обліку структури зв'язного тексту, а також побудові систем, які будуть стійкі при перенесенні на різні предметні області.

ПЕРЕЛІК ПОСИЛАНЬ

1. Feldman R. Techniques and Applications for Sentiment Analysis Communications of the ACM. – 2013. Vol. 56, № 4.
2. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, – 2012 – 96pp.
3. Sebastiani F. Machine learning in Automated Text Categorization ACM Computing Surveys – 2002. Vol. 94. – p. 54-58.
4. Автоматичне породження гіпотез в інтелектуальних системах – М.: Ліброком, 2009. - 528 с.
5. Кожунова О. С. Технологія розробки семантичного словника системи інформаційного моніторингу – М., 2009. - 21 с.
6. Котельников Є. В. Розпізнавання емоційної складової в текстах: проблеми та підходи. Є. В. Котельников, М. В. Клековкіна, Т. А. Пескішева, О. А. Пестов - К.: Вид-во ВятГГУ, – 2012. – 103 с.
7. Котельников Є. В., Пескішева Т. А., Пестов О. А. Паралельний вибір параметрів класифікатора для аналізу тональності текстів СП.: Символ –Плюс, 2010. – 225с.
8. Nugumanova A., Bessmertnyi I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts – 2013. – 154с.
9. Позельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке – Москва, РГГУ, – 2011. – 522с.
10. Ермаков С.А., Ермакова Л.М. Методы оценки эмоциональной окраски текста – Вестник Пермского университета. – 2012, № 1. – С. 85-90.
11. Минаков И.А. Анализ эмоциональной тональности текста и его применение для повышения качества переходов по релевантным объявлениям // Вестник Самарского государственного технического университета – 2013. –241с.
12. Добросклонская Т.Г. Вопросы изучения медиатекстов. Опыт исследования современной английской медиаречи. М.: УРСС Эдиториал, 2005. 288с.
13. Максименко О.И., Зверева П.П. Современные направления лингвистических исследований имиджа страны и её жителей – М.: Вестник. – 2013, № 6. – С. 25–30.
14. Peter Turney Thumbs Up or Thumbs Down. Semantic Orientation Applied to Unsupervised Classification of Reviews . Proceedings of the Association for Computational Linguistics. – 2002. – С. 417–424.
15. А. Антонова, А Соловьев, Использование метода условных случайных полей для обработки текстов на русском языке. – М.: Изд-во РГГУ, 2013. – С.27-44.
16. Ю. В. Рубцова. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы, 2015, №1(109). – С.72-78.

17. Ушинский К. Д. Компьютерная лингвистика и интеллектуальные технологии. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». Сб. научных статей том 2. –С. 40-50.
18. García-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R.: Retrieving product features and opinions from customer reviews. IEEE Intelligent Systems. 2013, 28(3). – p.19–27.
19. Паттерн Singleton(Одиночка) [Электронный ресурс]/Russian Software Developer Network: Режим доступа: [www/URL: http://rsdn.org/article/patterns/singleton.xml](http://rsdn.org/article/patterns/singleton.xml) - 30.05.2018. – Паттерн Singleton(Одиночка)
20. Меньшиков И. Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке М.: Знания. – 2012. – С. 140-143.
21. Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval. – 2008. – Т. 2. – № 1-2. – С. 1-135.
22. Zimbra D., Ghiassi M., Lee S. Brand-related Twitter sentiment analysis using feature engineering and the dynamic architecture for artificial neural networks - IEEE, 2016. – С. 1930-1938.
23. Reynar J. C., Ratnaparkhi A. A maximum entropy approach to identifying sentence boundaries //Proceedings of the fifth conference on Applied natural language processing. - Association for Computational Linguistics, 1997. – С. 16-19.
24. Rajadesingan A., Zafarani R., Liu H. Sarcasm detection on twitter: A behavioral modeling approach //Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. - ACM, 2015. – С. 97-106.



Дослідження методів емоціонального окрасу тексту з емотіконами

Науковий керівник:
доцент
Вечур О.В.

Виконав:
ст. гр. ПЗСм-16-2
Пугачов Є.А.

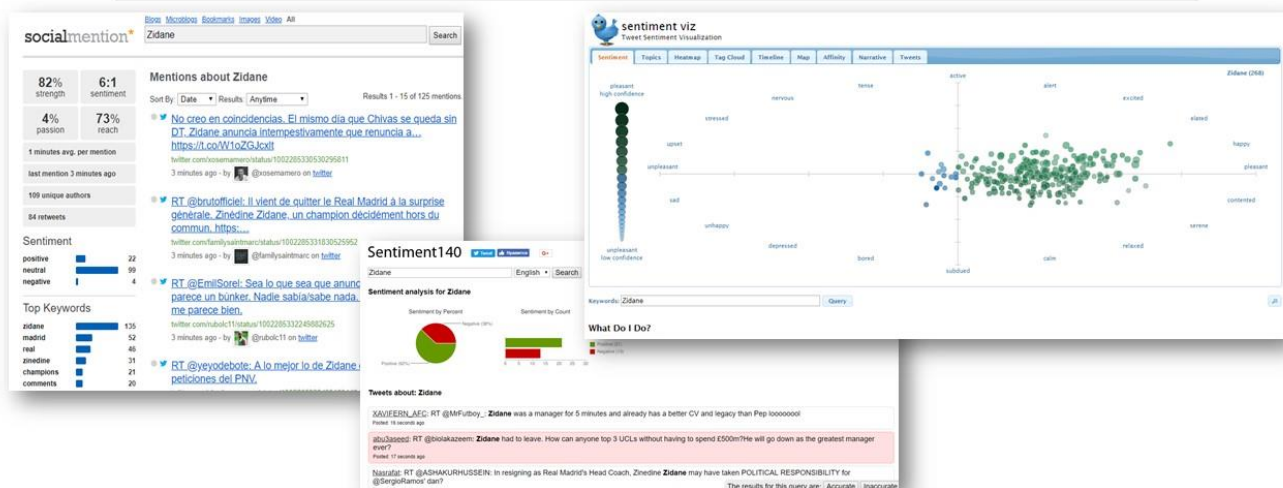


Мета роботи

- ❖ Дослідити методи емоціонального окрасу тексту з емотіконами
- ❖ Визначити метод для аналізу тональності тексту на основі емотиконів
- ❖ Оцінити практичне значення отриманих даних



Системи з аналізу тональності



3



Постановка задачі

- ❖ Проаналізувати методи які використовуються для аналізу емоційного окрасу тексту
- ❖ Розробити систему для визначення тонального забарвлення тексту
- ❖ Веб-система повина надавати змогу аналізувати повідомлення з соціальної мережі Twitter
- ❖ Твіти повині аналізуватися за декількома методами

4



Сентімент аналіз

Під словами аналіз тональності слід розуміти область комп'ютерної лінгвістики, що займається вивченням думок і емоцій в текстових документах. Аналіз тональності використовується для знаходжень думок і визначення їх властивостей, відносно вхідного тексту.

5



Сентімент аналіз

😊	:)	:-)	:smile:	Посмішка
😄	:D	:-D	:grin:	Широка посмішка
😉	;-)	;-)	:wink:	Підморгування
😞	:(:-)	:sad:	Смукот
😮	:o	:-o	:eek:	Здивування
😐	:	:-	:neutral:	Нейтральне
😱	8O	8-O	:shock:	Ботанік
😡	:x	:-x	:mad:	Роздратування
😝	:P	:-P	:razz:	Дражнити
😕	:?	:-?	:???:	В сумніві

Емотікони, емограмма або смайлик - це графічний символ, який використовується для вираження емоції.

6



Метод на основі словника

При використанні словникового підходу ключову роль відіграє використання тональних словників.

Такі словники, як правило, представляють собою списки слів, які допомагають визначити ставлення автора до деякого об'єкту.

Слова, що входять до переліку словника оціненими мають значення в діапазоні від -5 і до 5.

//positive	//negative
true,2	loser,3
trust,1	losing,3
trusted,2	loss,3
unbiased,2	lost,3
unequaled,2	lowest,1
unified,1	lugubrious,2
united,1	lunatic,3
unmatched,1	lunatics,3
unstoppable,2	lurk,1
untarnished,2	lurking,1
useful,2	lurks,1
usefulness,2	mad,3
vested,1	maddening,3
vibrant,3	made-up,1
vigilant,3	madly,3
vindicate,2	madness,3

7



Метод наївного Баєса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

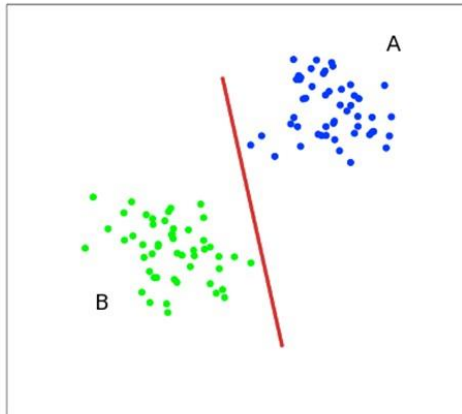
$$\log \frac{D_C}{D} + \text{foreach}(\text{word}) \left\{ \log \frac{W_{ic} + 1}{|V| + L_C} \right\} \quad (2)$$

де D_C – кількість документів в навчальній вибірці, що належать класу C , D – загальна кількість документів в навчальній вибірці, $|V|$ – кількість унікальних слів у всіх документах навчальної вибірки, L_C – сумарна кількість слів у документах класу C в навчальній вибірці, W_{ic} – скільки разів i -е слово зустрічалось в документах класу C в навчальній вибірці.

8



SVM метод



Метод опорних векторів (Support Vector Mashine, SVM).

Визначає приналежність об'єктів до класів за допомогою кордонів областей.

Один з найбільш популярних методів навчання по прецедентах.

Може використовуватися для розпізнавання образів або створення спам фільтрів.

Даний метод відноситься до бінарних класифікаторів, хоча може використовуватися для задач мультікласифікації.



Метод на основі емотіконів

```
{
  "name": "slightly_frowning_face",
  "emoji": "😞",
  "polarity": -1
},
{
  "name": "slightly_smiling_face",
  "emoji": "😊",
  "polarity": 1
},
{
  "name": "smile",
  "emoji": "😄",
  "polarity": 2
},
{
  "name": "smile_cat",
  "emoji": "😸",
  "polarity": 2
},
}
```

Емотікони не залежать від мови і не підпорядковуються граматичним правилам, будучи поняттям наднаціональним, одним з основних призначень якого – вираз невербальної інформації, емоцій.

Один емотікон може мати декілька варіантів відображення.



Головна сторінка веб-системи

Emoticon Meaning

Search tweet

Tweet Url

Message

Just smile

☒ Analysis only by smile

Dictionary

☒ Use dictionary method

SVM

☒ Use SVM method

Naive Bayes

☒ Use Naive Bayes method

Start

© 2018 - Emoticon Analysis

11



Сторінка результатів

Emoticon Meaning

AnalyzeTweet

Search Tweet Text

TweetId
1002898877597577217

TweetAuthor
pugachoveu

TweetText
I was very happy to study at this university, after its graduation, I will have new oport

SmileResult
●

DictionaryResult
●

SvmResult
●

BayesResult
●

Back

Emoticon Meaning

AnalyzeTweet

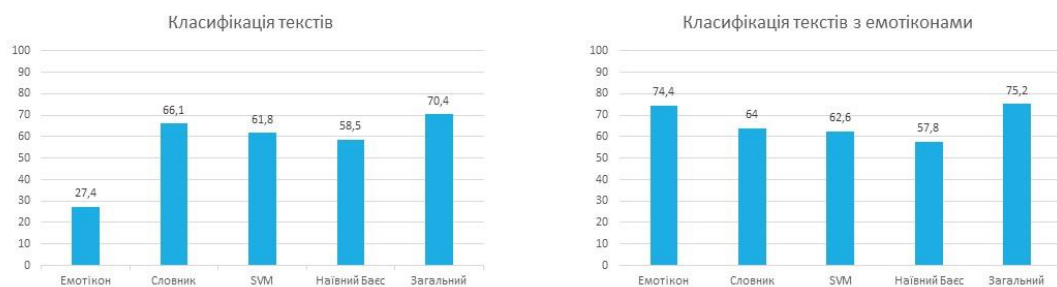
Search Tweet Text

Tweet Author	Text	Smile	Dictionary	SVM	Naive Bayes
thepodcouple	RT @NerdyThingsPod: ✨ NEW EPISODE ✨ This week we talk Pokémon news, #TWD, AND OUR MAAAAIN TOPIC is Solo: A Star Wars Story!! We go talk what...	😊	●	●	●
BrothersBinge	RT @kiersmclean: REVIEW! #SoloAStarWarsStory is a blast from beginning to end! @RealRonHoward and everyone involved did a great job. #StarW...	😊	●	●	●
one_abud	RT @heroichollywood: 'Star Wars': Ron Howard Retweets 'Solo' Praise That Bashes 'Last Jedi' https://t.co/5EkmMcebM https://t.co/BOPx51P6j5	😊	●	●	●
jedifarfy	"Here's this detail that everyone noticed 10 years ago when this book/movie/song was released and I just noticed!" = 10k upvotes. Fandoms on reddit, or why I've had to unsub from things like Harry Potter and Star Wars. :)	😊	●	●	●
ManuelDuarte24	Won't even dare to check the comments on the latest post on the Star Wars page.	😊	●	●	●
T1meTraveler01	RT @Lost789boy: Ways the world will implode due to an overdose of haterade: 1) The Fountainhead directed by Zack Snyder nominated &earnin...	😊	●	●	●
Warmustbeend123	RT @heroichollywood: 'Star Wars': Ron Howard Retweets 'Solo' Praise That Bashes 'Last Jedi' https://t.co/5EkmMcebM https://t.co/BOPx51P6j5	😊	●	●	●

12



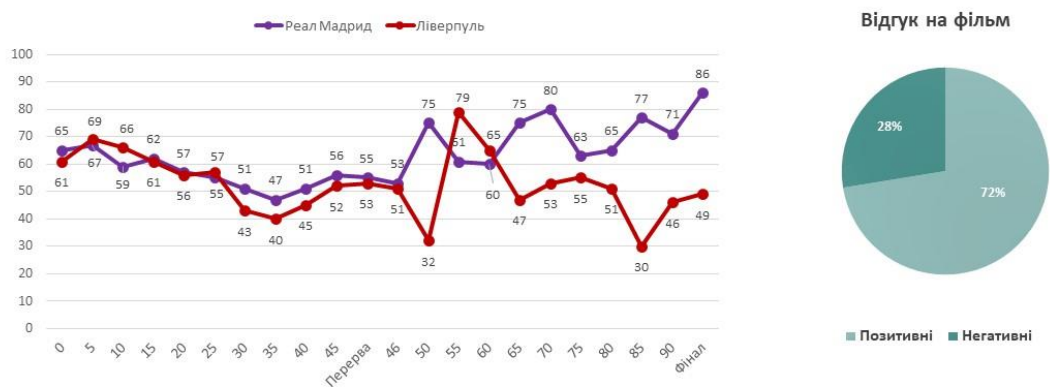
Результати



13



Прикладне використання



14



Висновки

- ❖ Досліджені методи які використовуються для аналізу емоційного окрасу тексту.
- ❖ Спроектowana і розроблена система для визначення тонального забарвлення тексту. Веб-система надає можливість аналізувати повідомлення з соціальної мережі Twitter.
- ❖ Знайдено практичне використання отриманих результатів роботи
- ❖ Шляхи до удосконалення:
 - ❖ Розширення інформаційної бази, яка використовується для аналізу текстів
 - ❖ Використання для аналізу біграм та уніграм
 - ❖ Додання інтеграції з іншими соціальними мережами

АНАЛІЗ ТОНАЛЬНОСТІ ТЕКСТА З ЕМОТІКОНАМИ

Пугачов Є.А.

Науковий керівник – к.т.н., доц. Вечур О.В.

Харківський національний університет радіоелектроніки

(61166, Харків, просп. Науки, 14, тел. (057) 702-13-06)

e-mail: yevhen.puhachov@nure.ua

Recognition of emotional coloring is also called sentiment analysis. Sentiment analysis - the area of - computer linguistics, dealing with the allocation of texts from emotionally colored vocabulary or emotional assessment of the author. It is one of the tasks of classifying texts. Also, sentiment analysis is an important part of the tasks artificial intelligence. The goal of the work is to implement a system that will be able to determine the emotional color of the message text with emoticons. The system will be self-learning based on real data from the social network with the use of machine learning from the teacher on the basis of the Support Vector Machine method.

Під словами «Аналіз тональності» слід розуміти область комп'ютерної лінгвістики, що займається вивченням думок і емоцій в текстових документах. Аналіз тональності використовується для знаходжень думок і визначення їх властивостей, відносно вхідного тексту[1]. Властивості, що визначаються при аналізі можуть бути різними, наприклад:

- автор - суб'єкт висловлює думку соціологія
- тема - об'єкт про яких йде мова,
- тональність - ставлення автора до теми тексту;

На практиці аналіз тональності знаходить застосування різних областях: маркетинг, соціологія, психологія, політологія.

У даній статті розглядаються способи визначення тональності текстів відгуків і коротких повідомлень які містять в собі емотікони.

Емотікон - це піктограма або послідовність друкованих знаків, що відображає емоцію.

Способи спілкування в соціальних мережах сильно відрізняються від норм літературної мови. І характеризуються використанням сленгових слів, авторської пунктуації, помилок і більш частим використанням смайлів. Для отримання вхідних даних була обрана соціальна мережа Twitter, аудиторія якої, на кінець 2017 року, становила більш ніж 330 мільйонів активних користувачів. А обробка повідомлень(твітів) є неможливою в ручному режимі і вимагає автоматизації.

Перед початком аналізу, повідомлення необхідно попередньо підготувати. Посилання замінюються на рядок формату @link, згадки користувачів замінюються на @username. Це дозволить зробити текст повідомлення більш незалежними від зовнішніх факторів які можуть вплинути на кінцевий результат.

Також варто замінити повторювані символи, послідовності однакових символів слід замінити на послідовність з двох таких же символів.[2]

Емотікони, в більшості випадків, однозначно вказують на емоційне забарвлення тексту і ставлення автора до теми висловлювання. Отже, якщо, текст містить емотікони тільки з позитивною або негативною тональністю, емоційна оцінка тексту буде відповідати значенню смайла. При цьому важливо враховувати кілька ознак, які впливають на ставлення автора до теми повідомлення:

- число позитивних емотіконів,
- число негативних емотіконів,
- чи є останній символ твіту позитивним емотіконом або негативним;

У разі коли текст містить емотікони протилежні за тональністю, значення розраховується на основі машинного навчання з учителем на основі методу опорних векторів (Support Vector Machine), з урахуванням наявності емотіконів у вхідному тексті. Суть підходу полягає в тому, щоб навчити класифікатор визначати тональність тексту на основі раніше запропонованих йому варіантів[3]. Для навчання використовується корпус твітів Twitter Sentiment Analysis Dataset, що містить 1 578 627 записів позначених міткою яка вказує на позитивний або негативний характер повідомлення. Ще одним можливим способом визначення забарвленості тексту є використання словників слів з заданим для кожного з них значенням рівня тональності[4].

Таким чином, система, дозволяє визначити тональність тексту з емотіконами в англomовному сегменті соціальної мережі Twitter. Для контролю використовувалися 500 повідомлень з вищевказаного корпусу повідомлень, які не були задіяні в процесі машинного навчання. Запропоновані в даній статті, методи щодо аналізу тональності текстів показали результати які можна порівнювати з сучасними аналогами. Для поліпшення роботи системи можливо також додаткове використання тональних словників.

Список використаних джерел

1. Пазельская А. Г., Соловьев А. Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2011». М.: Изд-во РГГУ, 2011. С. 510-522.
2. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения // РОМИП. 2011.
3. К.В. Воронцов. Математические методы обучения по прецедентам (теория обучения машин) 2010 М. 174 с.
4. Ермаков С.А., Ермакова Л.М. Методы оценки эмоциональной окраски текста // Вестник Пермского университета. Серия: математика, механика, информатика. 2012. № 1. С. 85–90.

Додаток В Код програми

```

using LinqToTwitter;
using System;
using System.Collections.Generic;
using System.Linq;
using System.Web.Mvc;
using WebEmoticonAnalysis.Models;
using Smile;
using Dictionary;
using SVM;
using NaiveBayes;

namespace WebEmoticonAnalysis.Controllers
{
    public class HomeController : Controller
    {
        public ActionResult Index()
        {
            return View();
        }
        [HttpPost]
        public async System.Threading.Tasks.Task<ActionResult>
AnalyzeTweetAsync(TweetAnalyzeViewModel model)
        {
            SmileMethod smileAnalyzer = SmileMethod.Instance;
            DictionaryMethod dictionaryAnalyzer =
DictionaryMethod.Instance;
            SvmMethod svmAnalyzer = SvmMethod.Instance;
            NaiveBayesMethod naiveBayesAnalyzer =
NaiveBayesMethod.Instance;
            if (String.IsNullOrEmpty(model.InputTweet))
            {
                ViewBag.Error = "Input is empty";
                return View("~/Views/Home/Index.cshtml");
            }
            if (!model.IsSmile && !model.IsDictionary && !model.IsSvm &&
!model.IsBayes)
            {
                ViewBag.Error = "Select any checkbox";
                return View("~/Views/Home/Index.cshtml");
            }
            var result = new ResultModel();
            ulong tweetId = 0;
            var tweetUrlArray = model.InputTweet.Split('/');
            try
            {
                tweetId = Convert.ToUInt64(tweetUrlArray.Last());
            }
            catch (FormatException e)
            {
                ViewBag.Error = "Enter correct tweet path or id";
                return View("~/Views/Home/Index.cshtml");
            }
        }
    }
}

```

```

var auth = new ApplicationOnlyAuthorizer
{
    CredentialStore = new InMemoryCredentialStore()
    {
        ConsumerKey = "7utcTWUvN3btE7fexfkClOnr8N",
        ConsumerSecret =
            "n5QBwTBeYsw1BsJ7K9M22St2yeGdErBTRK0ZoVLwXMPkjL5DFk3"
    }
};
await auth.AuthorizeAsync();
var twitterCtx = new TwitterContext(auth);
var currTweet = new Status();
var status =
    await
        (from tweet in twitterCtx.Status
         where tweet.Type == StatusType.Show &&
              tweet.ID == tweetId &&
              tweet.TweetMode == TweetMode.Extended &&
              tweet.IncludeAltText == true
         select tweet)
        .ToListAsync();
//Analyze tweet
if (status != null)
{
    currTweet = status.First();
    var message = currTweet.FullText;
    var smileResult = model.IsSmile ?
smileAnalyzer.Analyze(message) : 0;
    var dictionaryResult = model.IsDictionary ?
dictionaryAnalyzer.Analyze(message) : 0;
    var svmResult = model.IsSvm ? svmAnalyzer.Analyze(message)
: 0;
    var bayesResult = model.IsBayes ?
naiveBayesAnalyzer.Analyze(message) : 0;
    var smile = String.Empty;
    switch (smileResult)
    {
        case -1:
            smile = "negative";
            break;
        case 0:
            smile = "neutral";
            break;
        case 1:
            smile = "positive";
            break;
    }
    var total = String.Empty;
    switch (smileResult + dictionaryResult + svmResult +
bayesResult)
    {
        case -1:
            total = "negative";
            break;
        case 0:
            total = "neutral";
            break;
        case 1:

```

```

        total = "positive";
        break;
    }
    result.Tweet = new AnalyzeTweetResultViewModel
    {
        TweetAuthor = currTweet.User.ScreenNameResponse,
        TweetId = currTweet.StatusID,
        TweetText = currTweet.FullText,
        SmileResult = smile,
        DictionaryResult = dictionaryResult > 0 ? "positive" :
"negative",
        SvmResult = svmResult > 0 ? "positive" : "negative",
        BayesResult = bayesResult > 0 ? "positive" :
"negative",
        TotalResult = total
    };
}
//Analyze text
if (!String.IsNullOrEmpty(model.TextTweet))
{
    var message = model.TextTweet;
    var smileResult = model.IsSmile ?
smileAnalyzer.Analyze(message) : -99;
    var dictionaryResult = model.IsDictionary ?
dictionaryAnalyzer.Analyze(message) : -99;
    var svmResult = model.IsSvm ? svmAnalyzer.Analyze(message)
: -99;
    var bayesResult = model.IsBayes ?
naiveBayesAnalyzer.Analyze(message) : -99;

    var smile = String.Empty;
    switch (smileResult)
    {
        case -1:
            smile = "negative";
            break;
        case 0:
            smile = "neutral";
            break;
        case 1:
            smile = "positive";
            break;
    }
    var total = String.Empty;
    switch (smileResult + dictionaryResult + svmResult +
bayesResult)
    {
        case -1:
            total = "negative";
            break;
        case 0:
            total = "neutral";
            break;
        case 1:
            total = "positive";
            break;
    }
    result.Text = new AnalyzeTextResultViewModel

```

```

{
    TweetText = model.TextTweet,
    SmileResult = smile,
    DictionaryResult = dictionaryResult > 0 ? "positive" :
"negative",

    SvmResult = svmResult > 0 ? "positive" : "negative",
    BayesResult = bayesResult > 0 ? "positive" :
"negative",

    TotalResult = total
};
}
//Analyze tweet search
if (!String.IsNullOrEmpty(model.SearchTweet))
{
    ViewBag.SearchString = model.SearchTweet;
    Search searchResponse =
await
(from search in twitterCtx.Search
 where search.Type == SearchType.Search &&
   search.Query == model.SearchTweet &&
   search.IncludeEntities == true &&
   search.SearchLanguage == "en" &&
   search.TweetMode == TweetMode.Extended
 select search)
.SingleOrDefaultAsync();
if (searchResponse?.Statuses != null)
{
    result.Tweets = new
List<AnalyzeSearchResultViewModel>();
    foreach (var tweet in searchResponse.Statuses)
    {
        var message = tweet.FullText;
        var smileResult = model.IsSmile ?
smileAnalyzer.Analyze(message) : 0;
        var dictionaryResult = model.IsDictionary ?
dictionaryAnalyzer.Analyze(message) : 0;
        var svmResult = model.IsSvm ?
svmAnalyzer.Analyze(message) : 0;
        var bayesResult = model.IsBayes ?
naiveBayesAnalyzer.Analyze(message) : 0;
        var smile = String.Empty;
        switch (smileResult)
        {
            case -1:
                smile = "negative";
                break;
            case 0:
                smile = "neutral";
                break;
            case 1:
                smile = "positive";
                break;
        }
        var total = String.Empty;
        switch (smileResult + dictionaryResult + svmResult
+ bayesResult)
        {
            case -1:

```

```

        total = "negative";
        break;
    case 0:
        total = "neutral";
        break;
    case 1:
        total = "positive";
        break;
    }
    result.Tweets.Add(new AnalyzeSearchResultViewModel
    {
        TweetAuthor = tweet.User.ScreenNameResponse,
        TweetId = tweet.StatusID,
        TweetText = tweet.FullText,
        SmileResult = smile,
        DictionaryResult = dictionaryResult > 0 ?
"positive" : "negative",
        SvmResult = svmResult > 0 ? "positive" :
"negative",
        BayesResult = bayesResult > 0 ? "positive" :
"negative",
        TotalResult = total
    });
    }
}
ViewBag.IsSmile = model.IsSmile;
ViewBag.IsDictionary = model.IsDictionary;
ViewBag.IsSvm = model.IsSvm;
ViewBag.IsBayes = model.IsBayes;

return View(result);
}
}
}

```