**Міністерство освіти і науки України**

**Харківський національний університет радіоелектроніки**

Факультет комп’ютерних наук

Кафедра Програмної інженерії

**АТЕСТАЦІЙНА РОБОТА МАГІСТРА**пояснювальна записка

Дослідження методів аналізу емоціонального окрасу тексту з емотіконами

(Тема роботи)

Магістрант гр. ПЗСм-16-2 Пугачов Є.А.

Керівник роботи Вечур О.В.

Рецензент <прізвище, ініціали >

Рецензент <прізвище, ініціали >

Допускається до захисту  
Зав. кафедри, проф. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Дудар З.В.

2018 р.

*Харківський національний університет радіоелектроніки*

|  |  |  |
| --- | --- | --- |
| Факультет комп’ютерних наук. | | Кафедра Програмної інженерії |
| Спеціальність 121- Інженерія програмного забезпечення  Освітня програма Програмне забезпечення систем | | |
|  | ЗАТВЕРДЖУЮ:  “\_\_\_\_»\_\_\_\_\_\_\_\_20\_\_р \_\_\_\_\_\_\_\_\_\_  Зав. кафедри проф. З.В.Дудар | |

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ МАГІСТРА

магістрантові

***Пугачову Євгену Анатолійовичу***

Тема проекту (роботи) «*Дослідження методів аналізу емоціонального окрасу тексту з емотіконами»* затверджена наказом університету №\_\_\_\_\_\_\_\_ від «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ р.

1. Термін здачі студентом закінченої роботи «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_ 20\_\_\_ р.
2. Вихідні дані до проекту (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.*
3. Зміст пояснювальної записки(перелік питань, що їх належить розробити) *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.*
4. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень) *Мета завдання, обґрунтування доцільності розроблення, постановка задачі, об'єктна модель системи, базові моделі, методи й алгоритми, структура бази даних, структурно-логічна схема взаємодії даних, план захисту інформації (за необхідністю), інтерфейс програмної системи, результати тестування програмної системи.*

Зворотний бік бланку завдання

1. Консультанти з проекту (роботи) із зазначенням розділів проекту, що їх стосуються

|  |  |  |  |
| --- | --- | --- | --- |
| Розділ | Консультант | Підпис, дата | |
| Завдання видав | Завдання прийняв |
| Спецчастина | доц. Сидоров С.С. |  |  |
|  |  |  |  |

**Календарний план**

|  |  |  |  |
| --- | --- | --- | --- |
| № | Назва етапів дипломного проекту (роботи) | Термін виконання етапів проекту (роботи) | Примітка |
| 1 | Об'єктний аналіз поставленої задачі | 25-02-2018 | виконано |
| 2 | Розробка моделі взаємодії даних | 10-03-2018 | виконано |
| 3 | Розробка структури зберігання даних | 05-04-2018 | виконано |
| 4 | Створення коду програми | 13-03-2018 | виконано |
| 5 | Тестування і налагодження програми | 14-04-2018 | виконано |
| 6 | Підготовка пояснювальної записки. | 20-04-2018 | виконано |
|  | Спецчастина | 08-06-2018 | виконано |
| 7 | Підготовка презентації та доповіді | 08-06-2018 | виконано |
| 8 | Попередній захист | 12-06-2018 | виконано |
| 9 | Нормоконтроль, рецензування | 13-06-2018 | виконано |
| 10 | Занесення диплома в електронний архів | 14-06-2018 | виконано |
| 11 | Допуск до захисту у зав. кафедри | -06-2018 | виконано |

Дата видачі завдання «\_\_\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_20\_\_\_ р.

Керівник доц. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Вечур О.В.

Завдання прийняв до виконання\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Пугачов Є.А.

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 54 с., 10 рис., 7 табл., 3 додатки, 27 джерел. (НУЖЛИ ЛИ ЗДЕСЬ ФОРМУЛЫ???)

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

Об'єктом дослідження є методи аналізу тональності тексту, які застосовуються до аналізу тональності англомовного 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 основуються на наївному баєсовському класифікаторі, методі опорних векторів з урахуванням емотиконів, методі в основі якого лежить словник та методі визначення оцінки основаному 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 6](#_Toc513585063)

[1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ 8](#_Toc513585064)

[1.1 Аналіз предметної області 8](#_Toc513585065)

[1.2 Постановка задачі 10](#_Toc513585066)

[2 АНАЛІЗ ЕМОЦІЙНОГО ОКРАСУ ТЕКСТУ З ЕМОТІКОНАМИ 12](#_Toc513585067)

[2.1 Інтелектуальний аналіз даних 12](#_Toc513585068)

[2.2 Аналіз тональності текстів 15](#_Toc513585069)

[2.3 Емотікони 17](#_Toc513585070)

[2.4 Методи дослідження 17](#_Toc513585071)

[2.4.1 Метод на основі емотиконів 21](#_Toc513585072)

[2.4.2 Метод на основі словники 21](#_Toc513585073)

[2.4.3 Наївний Баєсівський метод 22](#_Toc513585074)

[2.4.4 Метод опорних векторів 25](#_Toc513585075)

[3 ПРОГРАМНА РЕАЛІЗАЦІЯ 26](#_Toc513585076)

[3.1 Джерело вхідних даних 26](#_Toc513585077)

[3.2 Архітектура програмної та структура даних 26](#_Toc513585078)

[3.3 Опис програмної системи 26](#_Toc513585079)

[3.4 Результати 26](#_Toc513585080)

[4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ 28](#_Toc513585081)

[ВИСНОВКИ 30](#_Toc513585082)

[ПЕРЕЛІК ПОСИЛАНЬ 31](#_Toc513585083)

[Додаток А Програмний код 33](#_Toc513585084)

[Додаток Б Слайди презентації 34](#_Toc513585085)

[Додаток В Електронні матеріали (CD) 35](#_Toc513585086)

Додаток В Апробація результатів роботи..

# 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 абревіатурні 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 англомовного 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 основуються на наївному баєсовському класифікаторі, методі опорних векторів з урахуванням емотиконів, методі в основі якого лежить словник та методі визначення оцінки основаному на значенні забарвлення смайла.

Отримані результати дозволяють стверджувати, що емотікони є відносно точним показником для визначення ставлення автора до об’єкту висловлювання. Врахування емотиконів при аналізі тексту за допомогою інших методі 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 IT-індустрії.

Серед найцікавіших і популярних методів цього широкого наукового напряму 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. 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. 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 [2]. Все вищесказане робить актуальним завдання аналізу тональності тексту.

## 1.2 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 емотиконів при визначені 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 емотикона до забарвлення тексту загалом. А також способи використання емотиконів 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 ім.

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.

Врахувавши 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 посилання 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. 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 XX 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. 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.

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 кожну 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. 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.

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. 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 т.п.), 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 текти. 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.

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.

Тональність тексту в цілому визначається лексичної 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.

Існує два основні методи вирішення цього 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.

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, емограмма 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 мімічного 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.

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.



Рисунок2.1 **–** Прикладзображенняемотиконів**.**

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.

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. Зробивши велике узагальнення, можна розділити існуючі підходи на наступні категорії:

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

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.

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. Тим не менш, цей підхід є найбільш точним при наявності хорошої бази правил.

При використані 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. 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 виповнити 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 формі. Мета цієї 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. 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, якщо автор висловлює жаль.

### 2.4.1 Метод на основі емотиконів

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 походить 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.

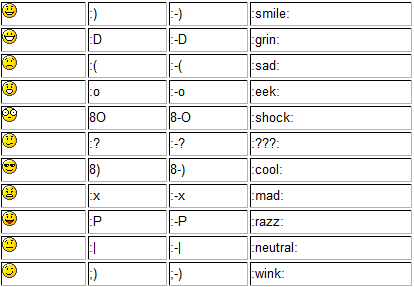


Рисунок 2.2 – Варіанти відображення емотіконів

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.

Метод, визначення тональності тексту на основі має декілька етапів та варіантів 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 най простішому випадку, якщо повідомлення містить емотікони, то тональність повідомлення визначається тональністю емотіконів. Коли всі емотікони мають одно-полярне 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 емотиконів, 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 способів, адже автор висловлює свої емоції не лише типом емотікона, а й їхньою кількістю. Слід, відмітити, що результативність цього підходу зростає у поєднані 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 емотиконів та проводити їх оцінювання і недоліком є неоднозначність при виокремлені емотиконів різних за полярністю.

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

### 2.4.2 Метод на основі словника

При використанні словникового підходу ключову роль відіграє використання тональних словників.

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.

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.

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.

В деяких випадках, для більш точного зіставлення кожного слова в реченні зі словником виконуються синтаксичний аналіз і лематизація.

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.

Байєсівський 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 погрішністю. Чим коротше вибірка, тим вище шанси підігнати розподіл під конкретні дані і зіткнутися з ефектом перенавчання. Байєсівський підхід до класифікації є одним з найстаріших, але до сих пір зберігає міцні позиції в теорії розпізнавання. Він лежить в основі багатьох досить вдалих алгоритмів класифікації.

В основі наївного Баєсова класифікатора лежить теорема Байеса(формула 2.1).

|  |  |
| --- | --- |
| , | (2.1) |

де P(A|B) - ймовірність що документ B належить класу A, саме її і треба розрахувати, P (B|A) - ймовірність зустріти документ В 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.

Для наївного баєсівського класифікатора визначено істотне припущення - передбачається, що всі ознаки x1, x2, ..., xn 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 [10, 11]. 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.

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.

|  |  |
| --- | --- |
| , | (2.2) |

де, Dc - кількість документів в навчальній вибірці належать класу С, D - загальна кількість документів в навчальній вибірці, |V| - кількість унікальних слів у всіх документах навчальної вибірки, Lc - сумарна кількість слів у документах класу c в навчальній вибірці, Wic - скільки разів i-е слово зустрічалося в документах класу c в навчальній вибірці, Q - безліч слів классифицируемого документа, включаючи повтори.

З точки зору програмування, вище наведена формула матиме вигляд як формула 2.3

|  |  |
| --- | --- |
| , | (2.3) |

Існує невелика проблема, пов'язана з цією формулою. Якщо в тестовому наборі зустрінеться слово, яке не зустрічається в наборі навчальних документів, то ймовірність цього слова для будь-якого з класів буде дорівнює нулю.

Основна відмінність полягає в тому, що для даного документа розглядається не кількість входжень слова, а тільки їх наявність або відсутність. Незважаючи на наївний вигляд і, безсумнівно, дуже спрощені умови, наївні баєсовські класифікатори часто працюють набагато краще в багатьох складних життєвих ситуаціях.

Перевагою наївного 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 (machine learning), 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 = (red, green, blue).

Якщо класів всього два, спам - не спам, червоне - чорне, то завдання називається бінарної класифікації. Якщо класів кілька - многоклассовая 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.

Роботу методу зручно проілюструвати на простому прикладі: дані точки на площині, розбиті на два класи (рис. 2.2).

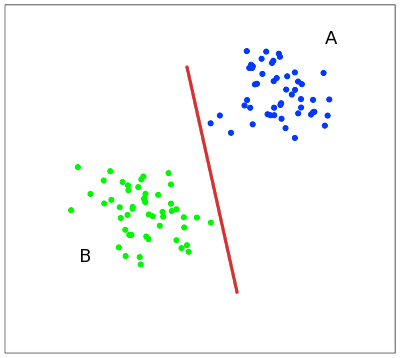


Рисунок 2.2 – Розділення вибірки на два класи.

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

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. 2.3).

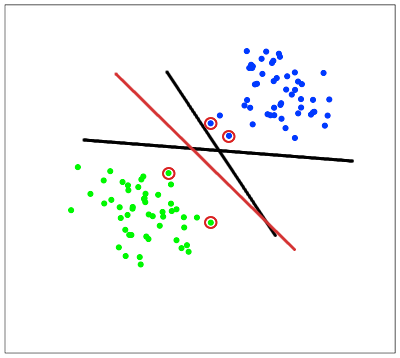


Рисунок 2.3 – Варіанти поділу вибірки.

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;

Необхідно додати, що слабкими сторонами цього методу є необхідність вибору ядра і погана интерпретируімість, 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

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 англомовний сегмент соціальної мережі.

Оскільки в даній роботі розглядається аналіз тональності англомовних висловлювань в Twitter, то доцільно буде розглянути особливості таких висловлювань. Робота з англомовними твітами має деякі проблеми для обробки природної мови. Так як твіти обмежені довжиною в 280 символів, то зазвичай вони містять в собі лише одну фразу, або пропозицію. Як правило, такі повідомлення містять в собі масу орфографічних помилок, сленгу, скорочень, url-посилань, хеш=тегів і звернень типу @ ім`я користувача. 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 зробите 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.

## 3.2 Архітектура програмної та структура даних

Основна проблема полягає в тому, як витягнути багату інформацію, яка доступна в Twitter, і як вона може бути використана для того, щоб витягнути корисну статистику. Для цього треба спочатку побудувати точний аналізатор настроїв для твітів, і саме це має на меті досягти цього рішення. В якості програмного забезпечення для аналізу даних можна використовувати SAS Text Miner, SAS Visual Analytics або інші інструменти. Завдання полягає в тому, щоб отримати спеціальні Tweets та очистити дані перед будь-яким текстом або символом видобутку. SAS Visual Analytics дозволяє безпосередньо імпортувати дані Twitter, але використовувати SAS Text Miner та інші інструменти, дані повинні бути завантажені та конвертовані. Твіттер дозволяє розробникам збирати дані за допомогою Twitter REST API [20] та API потоку [21]. 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. 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. На жаль, API Twitter експортує дані лише у форматі JSON, який потрібно перекласти на читання для баз даних або аналітичного формату програмного забезпечення. Комбінація API Twitter, скрипти для перетворення JSON в CSV [22], SAS Macro [23] або Excel Macro [24] можуть бути використані для вилучення інформації з Twitter та створення набору даних для аналізу. Весь процес збору даних може бути повністю автоматизований шляхом планування виконання Visual Basic for Applications (VBA) або макросів SAS. Оскільки думки мають цілі, подальша попередня обробка та фільтрація зібраних даних може бути зроблена за допомогою @ twitter\_names та #hashtags як цілі, як описано в [20]. Цей метод є більш точним та забезпечує кращий результат, ніж інші підходи до видобутку тексту.

Предварительная обработка текста В работе была произведена предварительная обработка текста, которую можно разделить на 3 этапа: 1) токенизация — это процесс выделения из текста отдельных слов, чисел и занков пунктуации; 2) стемминг2 — это процесс нахождения основы слова для задан- ного исходного слова; 3) обработка отрицаний. Цель стемминга — приведение слов, имеющих одинаковую основу к единой форме (так же за счёт этого происходит уменьшение раз- мерности задачи). После стемминга теряется часть морфологической информации, поэтому, как показали результаты ислледований, при- менение стемминга при использвании NB для анализа тональности текста на русском языке, как и в английском языке, не увеличивает точность. Токенизация производится на основе регулярного выражения.

Текст

СИНГЛТОН

## 3.3 Опис програмної системи

Текст

## 3.4 Результати

Для тестирования алгоритмов определение тональности текстов рецензий был использован метод кросс-валидации (или, по-другому, перекрестной проверки). Процедура кросс-валидации выполняется следующим образом:

1. Фиксируется множество разбиений обучающей выборки на собственно обучающую и тестовую.
2. Для каждого разбиения происходит обучение алгоритма на обучающей подвыборке и тестирование на тестовой.
3. Результатом кросс-валидации алгоритма являются средние значения оценок эффективности для тестовых подвыборок.

Для 800 положительных и 800 отрицательных кинорецензий было сделано 4 подразбиения (1200 рецензий – обучающая выборка, 400 рецензий – тестовая). В каждой обучающей и тестовой подвыборке содержалось одинакое количество положительных и отрицательных рецензий. Для каждой группы рассчитывались оценки эффективности, а затем высчитывались их средние значения. Таким образом, получились усредненные оценки эффективности. Результаты усредненных оценок эффективности для каждого из алгоритмов представлены ниже, в таблицах 3 и 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Векторная модель** | **Униграммы** | | **Биграммы** | |
| **Веса векторов** | **Точность, %** | **Полнота, %** | **Точность, %** | **Полнота, %** |
| **Бинарные векторы** | 84,5 | 88,25 | 85,3 | 91,5 |
| **Частотные векторы** | 85,2 | 87,4 | **86,2** | **89,4** |

Наивный Байесовский классификатор показал очень хорошие результаты, наиболее эффективной конфигурацией в плане точности оказалось сочетание биграмм и частотной функции взвешивания – точность 86%. Хотя, на рассматриваемой выборке, для разных параметров классификации, оценка точности варьируется слабо – поэтому эффективным можно считать любое сочетание.

# 4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ

Была установлена зависимость точности распознавания от объё- ма обучающей выборки. Для этого количество отзывов в обучающей выборке последовательно увеличивалось от 2000 до 70000 с шагом в 2000, а количество отзывов в тестовой выборке оставалось неизмен- ным и составляло 70000. На графике можно видеть, что точность возрастает при росте обу- чающей выборки (см. рис. 2). Это объясняется тем, что классифика- тор восстанавливает функцию плотности распределения по заданной конечной обучающей выборке, и при увеличении размера обучающей выборки, найденная плотность распределения более точно соответ- ствует действительной4 .

Сделать графики, віводи и пример на поиске твитов про фильм

ДАЛЬНЕЙШИЕ ПЛАНы

Для повышения эффективности программы, возможно, следует дополнить существующее решение элементами лингвистики. Лингвистические подходы позволяют определять тональность частей текста, тем самым существенно увеличивая точность классификации для отдельного текста – ведь, как правило, в рецензиях упоминаются и положительные и отрицательные стороны объекта тональности.

Кроме этого, стоит исключать из составляетмого словаря слова, не имеющие эмоциональной окраски, т.е. незначащие. Вполне возможно, что лучшие результаты даст извлечение ключевых слов (признаков).

Далее, для уменьшения словаря, также стоит проводить стемминг (приведение слов к их стеммам - основам). Также, можно высчитывать важность слов: например, согласно нотации TF.IDF. Важными словами для конкретного текста будут считаться слова, часто встречающиеся в данном тексте и редко – в остальных текстах выборки. Малозначимые слова можно убирать из словаря, тем самым, опять же, уменьшая его.

Обработка отрицаний Для увеличения точности был использован алгоритм обработки отрицаний, описанный Десом и Ченом [4]. Суть его заключается в следующем: при появлении частицы «не» к началу каджого слова между этой частицой и последующим знаком препинания либо дру- гой частицой «не» приписывается приставка «not\_»3 . Пример. Предложение: «Мне не понравился этот фильм.» Преобразуется к виду: «Мне не not\_понравился not\_этот not\_фильм.» Обработка отрицаний позволила улучшить точность, но на зна- чительно меньшую величину, по сравнению с англоязычными теск- тами [5

# ВИСНОВКИ

В результате осуществления проекта была создана программа для автоматического анализа тональности сообщений в русскоязычном сегменте социальной сети Twitter на основе методов машинного обучения. В ходе работы был создан байесовский классификатор с мультиномиальной моделью распределения, выбраны соответствующие метрики и произведены расчеты эффективности классификации путем тестирования методом кроссвалидации. Установлено, что точность классификации сопоставима с точностью современных аналогов.

# ПЕРЕЛІК ПОСИЛАНЬ

1. Feldman R. Techniques and Applications for Sentiment Analysis / / Communications of the ACM. 2013. Vol. 56, № 4. P. 82-89.

2. Kuznetsov S.O., Obiedkov S.A. Comparing Perfomance of Algorithms for Generating Concept Lattices / / Journal of Experimental and Theoretical Artificial Intelligence. 2002. Vol. 14.

3. Liu B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

4. Refaeilzadeh P., Tang L., Liu H. Encyclopedia of Database Systems / / Springer, US. 2009.

. Sebastiani F. Machine learning in Automated Text Categorization / / ACM Computing Surveys. 2002. Vol. 34. P. 1-47.

6. Автоматичне породження гіпотез в інтелектуальних системах / під ред. В. К. Фінна.- М.: Ліброком, 2009. - 528 с.

. ДСМ-метод автоматичного породження гіпотез / під ред. О. М. Аншакова.- М.: Ліброком, 2009. - 432 с.

. Кожунова О. С. Технологія розробки семантичного словника системи інформаційного моніторингу / / Автореферат дисертації на здобуття наукового ступеня кандидата технічних наук.- М., 2009. - 21 с.

. Котельников Є. В. Розпізнавання емоційної складової в текстах: проблеми та підходи / Є. В. Котельников, М. В. Клековкіна, Т. А. Пескішева, О. А. Пестов; під. ред. С. М. Окулова.- К.: Вид-во ВятГГУ, 2012. - 103 с.

. Котельников Є. В., Пескішева Т. А., Пестов О. А. Паралельний вибір параметрів класифікатора для аналізу тональності текстів / / Питання сучасної науки і практики. Університет ім. В.І. Вернадського. Тамбов: ГОУ ВПО ТДТУ, 2012. С. 67-74.

2015, Том 15, № 1

Р.В. Посевкин, И.А. Бессмертный

2. Nugumanova A., Bessmertnyi I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts // Communications in Computer and Information Science. 2013. V. 394. P. 92-101. doi: 10.1007/978-3-642-41360-5\_8

3. Позельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Тезисы докладов Международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2011». Москва, РГГУ, 2011. С. 510-522.

4. Cruz F.L., Troyano J.A., Pontes B., Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels // Expert Systems with Applications. 2014. V. 41. N 13. P. 5984-5994. doi: 10.1016/j.eswa.2014.04.005

5. Ермаков С.А., Ермакова Л.М. Методы оценки эмоциональной окраски текста // Вестник Пермского университета. Серия: математика, механика, информатика. 2012. № 1. С. 85-90.

6. Parau P., Stef A., Lemnaru C., Dinsoreanu M., Potolea R. Using community detection for sentiment analysis // Proc. IEEE 9th Int. Conf. on Intelligent Computer Communication and Processing (ICCP 2013). 2013. P. 51-54. doi: 10.1109/ICCP.2013.6646080

7. Chiru C.-G., Hadgu A.T. Sentiment-based text segmentation // Proc. 2nd Int. Conf. on Systems and Computer Science (ICSCS 2013). 2013. P. 234-239. doi: 10.1109/IcConSCS.2013.6632053

8. Минаков И.А. Анализ эмоциональной тональности текста и его применение для повышения качества переходов по релевантным объявлениям // Вестник Самарского государственного технического университета. Серия: технические науки. 2013. № 1 (37). С. 58-63.

Добросклонская Т.Г. Вопросы изучения медиатекстов. Опыт исследования со- временной английской медиаречи. М.: УРСС Эдиториал, 2005. 288 с.

2. Максименко О.И., Зверева П.П. Со- временные направления лингвисти- ческих исследований имиджа страны и её жителей // Вестник Московского государственного областного универ- ситета. Серия: Лингвистика. 2013. № 6. С. 25–30.

1. 3. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на рус- ском языке // Компьютерная лингви- стика и интеллектуальные технологии: «Диалог-2011  
   Bo Pang, Lillian Lee, Shivakumar Vaithyanathan [Thumbs up? Sentiment Classification using Machine Learning Techniques](http://www.cs.cornell.edu/home/llee/papers/sentiment.home.html) // — 2002. — С. 79–86.
2. Peter Turney [Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews](http://acl.ldc.upenn.edu/P/P02/P02-1053.pdf) // Proceedings of the Association for Computational Linguistics. — 2002. — С. 417–424. — arΧiv: LG/0212032
3. Анна Антонова и Алексей Соловьев, [Использование метода условных случайных полей для обработки текстов на русском языке.](http://www.dialog-21.ru/digests/dialog2013/materials/pdf/AntonovaAJ.pdf) Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». Сб. научных статей / Вып. 12 (19).- М.: Изд-во РГГУ, 2013.– С.27-44.
4. [Sentiment Analysis Track at ROMIP-2012. Chetviorkin I.I.,Loukachevitch N.V.](https://istina.msu.ru/publications/article/4706991/) Компьютерная лингвистика и интеллектуальные технологии. Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2013». Сб. научных статей том 2, с. 40-50.
5. Анна Пазельская и Алексей Соловьев, [Метод определения эмоций в текстах на русском языке. Компьютерная лингвистика и интеллектуальные технологии.](http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/50.pdf) Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Сб. научных статей / Вып. 11 (18).- М.: Изд-во РГГУ, 2011.– С.510-523.
6. Tarasov D.S. Deep [Recurrent Neural Networks for Multiple Language Aspect-Based Sentiment Analysis](http://www.dialog-21.ru/digests/dialog2015/materials/pdf/TarasovDS.pdf) // Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference “Dialogue-2015”, Issue 14(21), V.2, pp. 65-74 (2015).
7. García-Moya, L., Anaya-Sanchez, H., Berlanga-Llavori, R.: [Retrieving product features and opinions from customer reviews.](http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6509374) IEEE Intelligent Systems 28(3), 19–27 (2013)

Меньшиков И. Л., Кудрявцев А. Г. Обзор систем анализа тональности текста на русском языке // Молодой ученый. — 2012. — №12. — С. 140-143. — URL https://moluch.ru/archive/47/5951/ (дата обращения: 09.05.2018).

# Додаток А Програмний код

# Додаток Б Слайди презентації

# Додаток В Електронні матеріали (CD)