**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.

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 8](#_Toc513585066)

[2 АНАЛІЗ ЕМОЦІЙНОГО ОКРАСУ ТЕКСТУ З ЕМОТІКОНАМИ 9](#_Toc513585067)

[2.1 Інтелектуальний аналіз даних 9](#_Toc513585068)

[2.2 Аналіз тональності текстів 9](#_Toc513585069)

[2.3 Емотікони 9](#_Toc513585070)

[2.4 Методи дослідження 9](#_Toc513585071)

[2.4.1 Метод на основі емотиконів 9](#_Toc513585072)

[2.4.2 Метод на основі словники 9](#_Toc513585073)

[2.4.3 Наївний Баєсівський метод 9](#_Toc513585074)

[2.4.4 Метод опорних векторів 10](#_Toc513585075)

[3 ПРОГРАМНА РЕАЛІЗАЦІЯ 11](#_Toc513585076)

[3.1 Джерело вхідних даних 11](#_Toc513585077)

[3.2 Архітектура програмної та структура даних 11](#_Toc513585078)

[3.3 Опис програмної системи 11](#_Toc513585079)

[3.4 Результати 11](#_Toc513585080)

[4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ 12](#_Toc513585081)

[ВИСНОВКИ 13](#_Toc513585082)

[ПЕРЕЛІК ПОСИЛАНЬ 14](#_Toc513585083)

[Додаток А Програмний код 15](#_Toc513585084)

[Додаток Б Слайди презентації 16](#_Toc513585085)

[Додаток В Електронні матеріали (CD) 17](#_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. Розробити систему 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 їх можливі комбінації, для отримання більш точних результатів. Отримані результати маю бути наведені в зручному вигляді для опрацювання та для подальшого аналізу.

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.

# 2 АНАЛІЗ ЕМОЦІЙНОГО ОКРАСУ ТЕКСТУ З ЕМОТІКОНАМИ

## 2.1 Інтелектуальний аналіз даних

Текст

## 2.2 Аналіз тональності текстів

текст

## 2.3 Емотікони

Текст

## 2.4 Методи дослідження

Існують два основні підходи до вирішення завдання аналізу тональності тексту: на основі словників і на основі машинного навчання. У першому підході використовуються словники, що містять слова і пропозиції, для яких відома оцінка вираженої в них тональності. Цей підхід ефективний при використанні великих словників, але процес їх складання досить трудомісткий. Другий підхід полягає у створенні автоматичного класифікатора, який використовує колекцію навчальних текстів. В основі цього підходу лежать статистичні методи. Підхід ефективний при наявності великої колекції навчальни

### 2.4.1 Метод на основі емотиконів

Текст

### 2.4.2 Метод на основі словники

текст

### 2.4.3 Наївний Баєсівський метод

текст

### 2.4.4 Метод опорних векторів

# 3 ПРОГРАМНА РЕАЛІЗАЦІЯ

## 3.1 Джерело вхідних даних

текст

## 3.2 Архітектура програмної та структура даних

текст

## 3.3 Опис програмної системи

Текст

## 3.4 Результати

текст

# 4 МОЖЛИВІСТЬ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ У НАУКОВІЙ І ПРАКТИЧНІЙ ДІЯЛЬНОСТІ

текст

# ВИСНОВКИ

# 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.

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)