УДК 004.855.5

*ДОРОГИЙ Я.Ю.,*
*ДЕМЧИНСЬКИЙ В.В.,*
*ПОПКО А. В.*

# МЕТОД ВИЗНАЧЕННЯ ЗАБАРВЛЕНОСТІ ТЕКСТОВИХ ПОВІДОМЛЕНЬ

В даній статті розглянуто питання вибору математичних засобів для рішення задачі визначення забарвленості текстових повідомлень та реалізації програмного забезпечення за обраною моделлю.

In this article the selection of mathematical tools for solving the problem of determining the color for text messages and software implementation according to the chosen model are considered.

**Ключові слова:** класифікація текстів, машинне навчання, метод опорних векторів.

## Вступ

Обсяг тексту коментарів до публікацій в соціальних мережах практично завжди перевищує безпосередньо сам запис. Часто серед висловлювань користувачів можна зустріти досить негативні, що часто супроводжуються нецензурною лексикою та образливими виразами, які в подальшому можуть призводити до розростання конфлікту на місці публікації. Модерація коментарів вручну чи їх відключення частково вирішує проблему з сторони публікатора. Якщо ж стоїть питання фільтрації та відключення з сторони користувача, наприклад батьківський контроль, то вирішення проблеми є дещо складнішим. Наявність програмного модуля, що зміг би сортувати вхідний текст відносно невеликого обсягу принаймні на дві категорії за наявністю образ чи їх відсутності, дало б змогу автоматизувати процес ідентифікації та в подальшому створити різноманітні обгортки для використання в різних умовах.

## Аналіз існуючих рішень

В більшості випадків кожен з алгоритмів для вирішення даної задачі класифікується за двома складовими: за вибором характеристик тексту, що корелюють з його забарвленням та за методом їх обробки. Далі наводяться деякі характеристики тексту та методи їх обробки.

*Характеристики тексту.* Характеристика тексту, як правило, представляє собою деяку величину, яку є можливість обрахувати для даного тексту. Як приклад можна розглядати розподілення довжин слів у виразі. Існує велика кількість задач аналізу тексту, що потребують попереднього визначення його характеристик. До їх числа відноситься, наприклад, задача визначення авторства тексту. Ефективність використання даної характеристики оцінювалась експериментально.

В якості характеристик тексту для визначення його авторства в різні часи пропонувалось використовувати наступні [1–3]:

- розподіл довжини слів, середня довжина слова [5-7];
- розподіл сполучень символів деякої довжини (n-літерні n-грами) [8, 9]
- розподіл частин тексту та їх позиції в реченні [1-3]. Розглядався розподіл частин на декількох перших позиціях та в кінці речення;
- розподіл часто вживаних в мові слів. Існують частотні словники, де вказується кількість входжень певного слова на мільйон інших в колекції текстів [8, 9];
- словарний запас [10, 11]. В роботі [12] також запропоновані показники для тексту такі як $V_1 = \frac{V}{L}$; $V_2 = \frac{V}{\sqrt{L}}$; $V_3 = \frac{\log V}{\log L}$; $V_4 = \frac{\log V}{\log\log L}$ , де $V$ – число унікальних слів в тексті, $L$ – їх загальна кількість в тексті.

Пізніше виявилось, що більшість з характеристик тексту підходять і для рішення задачі визначення емоційного забарвлення та наявності образливих фраз в тексті.

Існують також дещо інші характеристики, які можна використовувати, якщо мова стосується коментарів соціальних мереж:

- наявність слів, написаних повністю заголовними літерами [13];
- наявність слів, що оточені різними астерисками (знаками) [13];
- наявність смайликів, посилань, хеш тегів тощо [13–15].

*Методи обробки характеристик тексту.* Одними з перших методів для визначення емоційного забарвлення тексту, були методи, які засновані на правилах (Rule-based methods). Їх робота полягає в пошуку в тексті характерних синтаксичних сполучень, конструкцій, які можуть з високою ймовірністю вказати на емоційне забарвлення чи наявність образи. Наприклад, вживання частки «ні» перед позитивно забарвленими словами, що характеризують особистість тощо, змінює значення на негативне. Більш складні правила можуть враховувати структуру речення, наприклад, наявність заперечних часток чи сполучників в складносурядних чи складнопідрядних реченнях: «…ти молодець, але (проте)...». Якість таких методів напряму залежить від якості та кількості запропонованих правил та потребує роботи лінгвіста. Автоматичний синтаксичний розклад речення є досить складною задачею. Приклад даного підходу розглянуто в роботі [16]. Прості елементи даного підходу використовують разом в сполученні з іншими методами, наприклад, методами машинного навчання.

В деяких випадках в задачах класифікації текстів використовують різноманітні статистичні методи і критерії. В роботі [17] текст розглядають як Марківський ланцюг символів. В роботі [12] запропоновано Хі-квадрат – статистику як міру визначення схожості текстів. Теоретично, такі методи можна використовувати і для рішення задачі визначення наявності образ в коротких текстах (коментарях), проте надійних експериментальних результатів в даному напрямку поки що немає.

Найчастіше на практиці використовують методи машинного навчання для досягнення результату. На відміну від методів, заснованих на правилах, вони дозволяють враховувати більшу кількість характеристик тексту. Даний підхід дозволяє абстрагуватись від прив'язаності до конкретної мови та є автоматизованим.

## Мета роботи

Метою даною роботи є розробка та аналіз методу дослідження емоційного забарвлення тексту, а також програмного комплексу, який реалізує даний метод.

## Метод опорних векторів для аналізу тексту

Нехай $Y = \{1, -1\}$. Метод опорних векторів дає можливість побудувати класифікатор наступного вигляду (1):

$$a(x, w, w_0) = \text{sign}(\sum_{i=1}^{n} w_i \xi_i - w_0) \quad (1)$$

де $w = (w_1, \dots, w_n) \in \Re^n$, $w_0 \in \Re$ – параметри алгоритму. Рівняння $\langle w, x \rangle = w_0$ описує гіперповерхню в просторі $\Re^n$. Ідея методу заключається в побудові оптимальної в деякому розумінні гіперповерхні, що буде розділяти класи $y = 1$ та $y = -1$. Вибірка лінійно подільна, якщо існує така гіперповерхня, коли об'єкти різних класів находяться по різні сторони даної поверхні. Розділяюча гіперповерхня називається оптимальною, якщо відстань від неї до найближчого об'єкта максимальна у порівнянні с іншими розділяючими поверхнями. Ці вимоги записуються системою (2):

$$\begin{cases} \langle w, w \rangle \to \min; \\ y_i(\langle w, x_i \rangle - w_0) \geqslant 1, \ \ i = 1 \dots l. \end{cases} \quad (2)$$

Вибірки на практиці зазвичай є лінійно нероздільними. Метод узагальнюється на цей випадок шляхом дозволу похибок на об'єктах. В такому випадку вимоги будуть записуватись системою (3):

$$\begin{cases} 0.5\langle w, w \rangle + C \sum_{i=1}^{k} e_i \to \min; \\ y_i(\langle w, x_i \rangle - w_0) \geqslant 1 - e_i, \ \ i = 1 \dots l; \\ e_i \geqslant 0, \ \ i = 1 \dots l. \end{cases} \quad (3)$$

Тут $e_i$ – розмір похибки на об'єкті $x_i$, C – параметр алгоритму, оптимальне значення якого знаходиться експерементально і регулює співвідношення між шириною смуги і сумарною похибкою на об'єктах.

У випадку коли вибірка лінійно неподільна зазвичай використовується «ядровий трюк» (*kernel trick*). Його суть в наступному. Нехай дано відображення $\varphi: \Re^n \to H$, де $H$ – простір із скалярним добутком $\langle a_1, a_2 \rangle_H$. Якщо цей простір має більш високу розмірність, то ймовірно, що вибірка в ньому буде роздільною. Рівняння розділяючої гіперповерхні в просторі $H$ буде мати вигляд $\langle \varphi(X), w \rangle H = w_0$. Відповідь на питання «Якою буде поверхня та як буде виглядіти в просторі» залежить від $H$. На практиці беспесередньо відображення нас не цікавить, достатньо лише знати ядро – функцію $K(x_1, x_2) = \langle \varphi(x_1), \varphi(x_2) \rangle_H$.

Приклади ядер, що можуть бути використані, наступні:

- $K(x_1, x_2) = \langle (x_1), (x_2) \rangle_{\Re^n}$ – лінійне ядро, розділяюча поверхня матиме вид гіперповерхні;

- $K(x_1, x_2) = (\langle (x_1), (x_2) \rangle_{\Re^n})^d$ – поліноміальне ядро, розділяюча поверхня буде мати вид поверхні порядку $d$;
- $K(x_1, x_2) = \mathrm{e}^{\frac{|x_1 - x_2|^2}{2\sigma^2}}$ – гаусове ядро з параметром σ;
- $K(x_1, x_2) = th(k_0 + k_1 \langle (x_1), (x_2) \rangle_{\Re^n})$ – сигмоїдне ядро з параметрами $k_0, k_1$.

Існують правила, за якими можна побудувати свої ядра, проте в більшості випадків на практиці використовують один з вказаних вище ядер. Для модифікації алгоритму с урахуванням ядер достатньо всюди в програмі зробити заміну звичайного скалярного добутку на функцію ядра.

Даний метод широко використовується в різноманітних задачах машинного навчання. Можливість варіювати ядро $K$ та параметр $C$ забезпечує методу необхідну гнучкість. В загальному випадку складність алгоритму є поліноміальною від вибірки. Для лінійного ядра існують методи, що дозволяють найти близьку до оптимальної гіперповерхню за лінійний проміжок часу.

### Запропонований метод рішення

В даній роботі висунута наступна гіпотеза: використання літерних *n*-грам в якості характеристик тексту у зв'язці з методами машинного навчання дає гарний кінцевий результат класифікації тексту за наявністю образливих оборотів в коментарях до записів з соціальних мереж. Одним із найчастіше вживаних методів для класифікації текстів є використання звичайних *n*-грам в поєднанні з будь-яким методом машинного навчання. Зазвичай в ролі останнього використовують метод опорних векторів SVM (*support vector machine*). Проте, в силу властивостей та специфіки соціальних мереж, вхідний текст має невелику кількість слів. Це негативно впливає на якість розпізнавання та віднесення до певного класу при використанні *n*-грам. Останні успішно використовуються для класифікації текстів за їх авторством [12] у зв'язці з використанням статичних критеріїв. Пропонується наступний алгоритм. Кожному тексту ставиться у відповідність чисельний вектор характеристик, що буде складатись з частоти зустрічання в ньому літерних *n*-грам при $n \leqslant 3$ або ж індикаторів наявності літерних *n*-грам при $n > 3$. Для уникнення великої кількості характеристик є варіант використання відстані між розподіленнями. Для цього необхідно для кожного заданого n розрахувати розподілення літерних n-грам на даному тексті. Після цього розраховується відстань від даного розподілення до розподілення літерних n-грам на множині текстів з позитивним забарвленням і відсутністю в ньому образ та з фразами, що їх мають. В якості відстані взяті наступні функції (4):

$$L_1(\mathrm{p}, \mathrm{q}) = \sum_i |p_i - q_i|;$$
$$L_2(\mathrm{p}, \mathrm{q}) = \sum_i (p_i - q_i)^2; \quad (4)$$
$$D_{KL}(\mathrm{p}, \mathrm{q}) = \sum_i \ln(\frac{p_i}{q_i}) p_i.$$

Тут $p$, $q$ – дискретний розподіл з ймовірністю $i$-го значення $p_i$ і $q_i$ відповідно, $D_{KL}$ – відстань Кульбака-Лейблера між двома розподілами, яка запропонована для використання при класифікації текстів [18]. Таким чином, якщо використати $m$ відстаней при заданому $n$ у нас буде можливість поставити у відповідність $2m$ відстаней: $m$ відстаней до розподілу $n$-грам на множині текстів, що не мають образ та $m$ відстаней до множини текстів, які їх включають.

Для отримання кращого результату при обчисленні характеристик використані тільки *n*-грами, які зустрічаються в наборі коментарів більше заданого числа разів. Даний підхід дозволяє враховувати наявність друкарських помилок в тексті.

Також, в даній задачі запропоновано використовувати ряд синтаксичних характеристик тексту, які успішно використовуються в рішеннях задач визначення авторства тексту [10]. Виходячи з чого було складено наступні комплекти характеристик:

1. Синтаксичні:
   - довжина речення в словах;
   - середня довжина слова в символах;
   - середня довжина виразів (речень) в словах;
   - середня довжина виразів (речень) в символах;
   - розподіл кількості слів заданої довжини (від 1 до 20);
   - розподілення частин мови;
   - розподілення частин мови, що стоять на першій та другій позиціях в реченні.
2. Літерні *n*-грами. Їх можна розглядати як набір характеристик при одному *n*, так і поєднання декількох наборів характеристик при різних *n*. Так, наприклад, доцільним буде розглядати 2-грами та 3-грами разом в силу їх відносно невеликої кількості. В такому випадку

їм буде відповідати набір характеристик, що складається з частоти зустрічання літерних 2-, 3-грам.

3. Відстань між розподіленнями літерних *n*-грам в даному тексті і розподіленні на множині всіх текстів одного класу.

### Оцінка якості роботи алгоритму

Для визначення ефективності роботи алгоритму необхідно мати декілька метрик, які відображатимуть якість віднесення висловлювань до однієї з двох категорій на тестовому наборі фраз:

$tp_x$ – кількість об'єктів класу *x*, які віднесені алгоритмом до класу *x;*

$fp_x$ – кількість об'єктів не класу *x*, які віднесені алгоритмом до класу *x*;

$fn_x$ – кількість об'єктів класу *x*, які віднесені алгоритмом не до класу *x*;

$tn_x$ – кількість об'єктів не класу *x*, які віднесені алгоритмом не до класу *x*;

$S$ – множина всіх класів, до яких відносяться об'єкти вибірки.

Для основних метрик якості запропоновані наступні:

- Precision – частка об'єктів, які віднесені алгоритмом до конкретного класу та які дійсно відносяться до нього (5):

$$P = \frac{tp_x}{tp_x + fp_x} \quad (5)$$

- Macro Precision – середнє значення точності за всіма класами:

$$\text{Macro\_P} = \frac{1}{|S|} \sum_{x \in S} \frac{tp_x}{tp_x + fp_x} \quad (6)$$

- Recall – частка об'єктів конкретного класу, які вірно класифіковані алгоритмом і віднесені до нього (7):

$$R = \frac{tp_x}{tp_x + fn_x} \quad (7)$$

- Macro Recall – середнє значення за повнотою серед усіх класів (8):

$$\text{Macro\_R} = \frac{1}{|S|} \sum_{x \in S} \frac{tp_x}{tp_x + fn_x} \quad (8)$$

- F1-measure – середнє гармонічне для метрик Precision та Recall (9):

$$F1 = \frac{2PR}{P+R} \quad (9)$$

- Macro F1-measure – середнє за метрикою F1-measure серед усіх класів (10):

$$\text{Macro\_F1} = \frac{1}{|S|} \sum_{x \in S} F1_x \quad (10)$$

- Accuracy – частка вірно класифікованих об'єктів серед усіх об'єктів за всіма класами (11):

$$\text{Accuracy} == \frac{1}{|S|} \sum_{x \in S} \frac{tp_x + tn_x}{tp_x + fn_x + tn_x + fp_x} \quad (11)$$

### Опис програмного забезпечення

Для демонстрації функціонування запропонованого підходу та підтвердження гіпотези розроблено програмне забезпечення, яке дає інструментарій для класифікації коротких текстів за наявністю в них образливих виразів та словосполучень. Текст, який подається на вхід застосування після опрацювання буде віднесений до однієї з двох категорій: «образливий» та «не образливий». На рис.1 схематично зображений алгоритм роботи програмного застосування.

Мовою для кодування обрано *Python* – найпопулярнішу для реалізації застосувань, пов'язаних з нейромережами та машинним навчанням. Це дозволило використати наявні безкоштовні бібліотеки, що значно скоротило час на реалізацію алгоритмів.

Основними частинами застосування є: модуль, який визначає характеристики тексту та модуль, що використовує метод опорних векторів для навчання та подальшої класифікації.

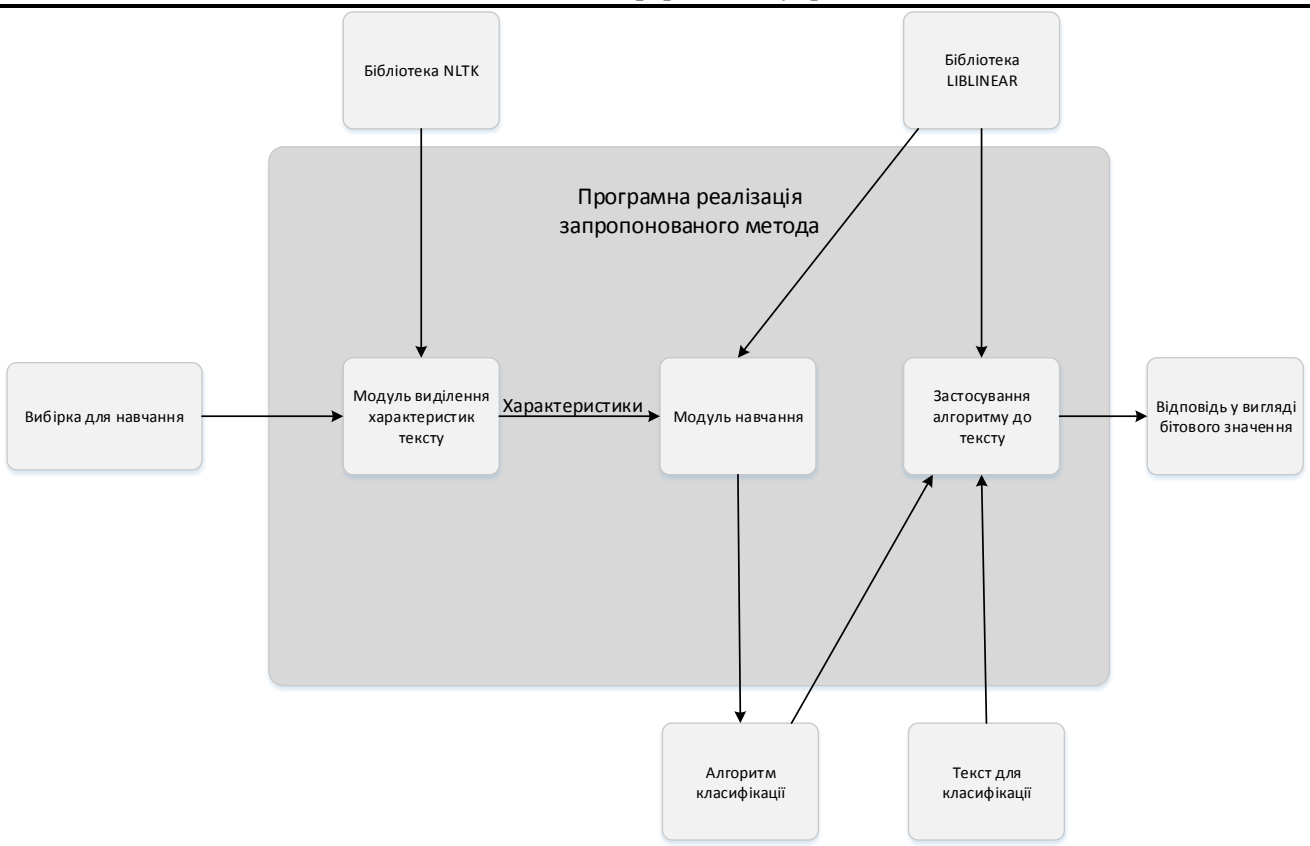


***Рис. 1. Схема роботи програми***

Вхідні дані для навчання представлені у вигляді файлу csv (*Comma-Separated Values*) з порталу змагань «Kaggle Data Science. Так як змагання призначені для учасників зі всього світу, всі вирази складені англійською мовою. Кількість записів у файлі – біля 4000. Представлену вибірку необхідно обробити перед проведенням подальшого аналізу. Так наприклад існують випадки, де фраза *"<font color="#8f8f88">goon it's about you </font>"* повинна бути приведена до вигляду *"goon it's about you"*. Таке приведення тексту необхідне для використання більш «чистих» даних для навчання. Для виконання розбору фраз та викорінення спеціальних символів використаний підхід, побудований на регулярних виразах. Весь текст також приведений до нижнього регістра.

Наступним етапом роботи програмного застосування є аналіз текстів. Тут обраховуються характеристики кожного запису з вибірки та середні характеристики для класу. Для виділення характеристик тексту використано бібліотеку під назвою NLTK (*Natural Language Toolkit*). Вона підходить для морфологічного, семантичного аналізу, стемінгу (скорочення слова до основи), лематизації (покращений варіант стемінгу) для різних мов.

Далі формується вибірка для навчання у вигляді набору векторів характеристик для кожного тексту з файлу.

Для машинного навчання використана реалізація методу опорних векторів, яка представлена бібліотекою «Liblinear» і яка була обрана для використання в застосуванні. Після успішного навчання застосування здатне класифікувати інший текст за наявністю в ньому образливих зворотів.

### Експериментальні дослідження

Для даної задачі протестовані набори характеристик, які описані вище. В якості вибірки для тестування взято тестову вибірку з вище вказаного порталу «Kaggle Data Science». Дані мають той же формат, що і для навчання. Проте тепер нам відомі правильні відповіді. Так як набір достатньо великий (близько двох тисяч коментарів), у нас є можливість більш точно оцінити показники класифікації.

Для оцінки якості класифікації використовувались метрики *MacroPrecision, MacroRecall, Macro_F1, Accuracy*. В табл. 1 представлені дані відповідно до варіантів модифікації запропонованого методу.

Найпростішим варіантом для класифікації взятий Baseline – алгоритм, який відносить усі

об'єкти до найбільш популярного класу. Так як вибірка достатньо нерівномірна (близько 26% коментарів відносяться до образливих, решта – до не образливих), показник *Macro_F1* є більш об'єктивним, що і видно за показниками.

Алгоритм з використанням синтаксичних признаків показав той самий результат, що і *BaseLine* алгоритм. Цей факт свідчить про те, що отримані числові представлення текстів неінформативні і оптимальною поверхнею виявилась та, по одну сторону від якої опинились всі вибірки.

Проведено дослідження впливу значення *n* на точність класифікації. В усіх випадках значення точності *Accuracy* не дуже високі. Це зумовлено перевагою текстів без образ над текстами, які їх мають. Значення *Macro_F1* із збільшенням *n* зростає та при *n*=6 стає найбільшим. Те саме стосується й інших метрик. Тобто кількість текстів, хибно віднесених до категорії образливих, скорочується як і кількість тих, які були пропущені й залишились не розпізнаними.

***Табл. 1. Результати тестування***

| Алгоритм | Метрика | | | |
|---|---|---|---|---|
| | **Macro_P** | **Macro_R** | **Macro_F1** | **Accuracy** |
| BaseLine | 0,422 | 0,50 | 0,47 | 0,84 |
| SVM і синтаксичні характеристики | 0,422 | 0,50 | 0,47 | 0,84 |
| SVM і відстань між розподілами літерних 2-,3-грам | 0,830 | 0,55 | 0,58 | 0,53 |
| SVM і літерні 2-,3-грами | 0,81 | 0,53 | 0,61 | 0,53 |
| SVM і літерні 4-грами | 0,80 | 0,56 | 0,58 | 0,55 |
| SVM і літерні 5-грами | 0,62 | 0,69 | 0,61 | 0,69 |
| SVM і літерні 6-грами | 0,66 | 0,72 | 0,67 | 0,79 |

## Висновки

В даній роботі представлені результати дослідження методів і засобів для вирішення задачі на визначення забарвленості змісту короткого тексту (повідомлення, коментарю). В рамках даного дослідження розглянуті деякі існуючі способи класифікації математичними методами. Запропоновано власний варіант для досягнення позитивного результату, а також визначено деякі можливі його модифікації для подальшого визначення кращого в рамках поставленої мети. Запропоновано і досліджено характеристики текстів, а саме літерні *n*-грами та ряд синтаксичних признаків. Реалізовано програмне забезпечення, яке здатне визначати характеристики тексту, які описують емоційну складову та образливі звороти.

Попередньо текст піддається обробці для видалення певного «шуму», який не несе ніякої інформації. Представлено опис методу машинного навчання та показаний спосіб його застосування для аналізу отриманих характеристик. Запропонований метод та програмне застосування пройшли тестові випробування, в тому числі й на реальних коментарях.

В якості майбутніх напрямків досліджень та вдосконалення моделі планується наступне:

1. Розгляд та дослідження методів автоматичного складання словників для слів, які явно будуть вказувати на образу. Знайдене, так зване «стоп-слово» або комбінація таких слів дозволить з більшою ймовірністю віднести даний коментар (текст) до категорії образливих. З урахуванням специфіки коментарів в соціальних мережах є сенс враховувати друкарські помилки при пошуку таких. Як варіант, їх необхідно буде порівнювати та визначати ступінь схожості з вже існуючими.
2. Дослідження можливості ідентифікувати сарказм. На даний момент реалізація моделі, яка описана в цій роботі не здатна ідентифікувати даний аспект. Проте за допомогою NLP проектів, таких як «Stanford CoreNLP» можна зрозуміти структуру висловлювання та в певній мірі його зміст. І, в подальшому, навчитись ідентифікувати сарказм в тексті автоматично.
3. Розширення кількості характеристик тексту, що в певній мірі повинно покращити якість класифікації текстів.

**Список посилань**


1. The advanced theory of language as choise and chance / Herdan Gustav // Kommunikation und Kybernetik in Einzeldarstellungen. – 1966. – С. 14–437.
2. The Statistical Study of Literary Vocabulary / Udny Y. G. // Shoe String Press Inc. U.S. – 1968.
3. Shakespeare, fletcher, and the two noble kinsmen / Ledger G., Merriam T. // Literary and Linguistic Computing. – 1994. – С.235–248.
4. Общий взгляд на машинное обучение: классификация текста с помощью нейронных сетей и TensorFlow [Електронний ресурс] – Режим доступу: https://tproger.ru/translations/text-classification-tensorflow-neural-networks.
5. A mechanical solution of a literary problem. / The Popular Science Monthly, с.97-105. [Електронний ресурс] / Mendenhall T. C. // Popular Science Monthly. – 1901. – №. 60. – Режим доступу: https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_60/December_1901/A_Mechanical_Solution_of_a_Literary_Problem.
6. Isaiah and the computer: A preliminary report [Електронний ресурс] / Radday Y. // Computers and the Humanities. – 1970. – №. 5. – С.65–73. – Режим доступу: https://link.springer.com/article/10.1007/BF02402282
7. Authorship attribution [Електронний ресурс] / Juola P. // Foundations and Trends in Information Retrieval. – 2006. – № 1. – С. 233—334. – Режим доступу: https://books.google.com.ua/books/about/Authorship_Attribution.html?id=_B2zDLdqe60C&redir_esc=y
8. N-gram-based author profiles for authorship attribution [Електронний ресурс] / Keselj V., Peng F., Cercone N., Thomas C. – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.7388&rep=rep1&type=pdf.
9. Scientific and Engineering Problem-Solving with the Computer / Bennett W. R. // Prentice Hall Series in Automatic Computation. – 1976. – №1.
10. Author identification: Using text sampling to handle the class imbalance problem [Електронний ресурс] / Stamatatos E. // Information Processing and Management: an International Journal. – 2008. – № 44. – С.790–799. – Режим доступу: https://dl.acm.org/citation.cfm?id=1347518.
11. The battle of The Quiet Don: Another pilot study [Електронний ресурс] / Kjetsaa G. // Computers and the Humanities. – 1977. – №11. – С.341-346. – Режим доступу: https://www.jstor.org/stable/30205004?seq=1#page_scan_tab_contents.
12. Quantitative authorship attribution: An evaluation of techniques [Електронний ресурс] / Grieve J. // Literary and Linguistic Computing. – 2005. – №22. – С.251-270. – Режим доступу: https://academic.oup.com/dsh/article-abstract/22/3/251/951481?redirectedFrom=fulltext.
13. Experiments with mood classification in blog posts [Електронний ресурс] / Mishne G. // 1st workshop on stylistic analysis of text for information access. – 2005. – Режим доступу: http://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.111.2693&rep=rep1&type=pdf
14. Mining newsgroups using networks arising from social behavior [Електронний ресурс] / Agrawal R., Rajagopalan S., Srikant R., Xu Y. // Proceedings of the 12th international conference on World Wide Web. – 2003. – С. 529–535. – Режим доступу: https://dl.acm.org/citation.cfm?id=775227.
15. Twitter sentiment analysis: The good the bad and the omg! [Електронний ресурс] / Kouloumpis E., Wilson T., Moore J. // Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. – 2011. – С. 538–541. – Режим доступу: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/2857/3251.
16. Rule-based approach to sentiment analysis at ROMIP 2011 [Електронний ресурс] / Kan D. // AlphaSense Inc.– 2011. – Режим доступу: http://www.dialog-21.ru/media/1393/138.pdf.
17. Определение авторства тексту с использованием буквенной и грамматическй информации [Електронний ресурс] / Кукушкина О.В., Поликарпов А. А., Хмелев Д. В. // Проблемы передачи информации. – 2001. – №2. – С. 96–109. – Режим доступу: http://www.philol.msu.ru/~lex/articles/grco_r.htm.
18. Using kullback-leibler distance for text categorization [Електронний ресурс] / Bigi B. // Proceedings of the 25th European conference on IR research. – 2003 – С. 305—319. – Режим доступу: https://ai2-s2-pdfs.s3.amazonaws.com/f9c7/4b45203266abf92f2f40e4b268aaf3274d38.pdf.
19. Математические методы обучения по прецедентам (теория обучения машин) [Електронний ресурс] / Воронцов К. В. // Курс лекций «Машинное обучение» – Режим доступу: http://www.machinelearning.ru/wiki /images/6/6d/Voron-ML-1.pdf.