Міністерство освіти і науки, молоді та спорту
Національний технічний університет України
“Київський політехнічний інститут”

Факультет інформатики та обчислювальної техніки
Кафедра автоматизованих систем обробки інформації та управління

ЛАБОРАТОРНА РОБОТА №2

з дисципліни “Комп’ютерна лінгвістика”

на тему: “Аналіз тональності тексту за допомогою Баєсівого класифікатора”

Виконав: Турко М. В.
ст. гр. ІС-73
Перевірив: Фіногенов О. Д.

Київ 2020

Лабораторна робота №2

Мета: вивчення методів класифікації при визначенні тональності тексту.

Варіант завдання: було обрано дві категорії анекдотів на сайті anekdot.ru

- ➢ Погода;
- ➢ Політика;

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Визначення тональності тексту є однією з типових задач, що має велике практичне значення. Найбільше розповсюдження дана задача отримала в галузі електронної комерції при аналізі відгуків на товари або послуги в соціальних мережах або інтернет-магазинах, визначенні вподобань цільової аудиторії та її відгук на політичні події, рекламні та передвиборчі "слогани" тощо.

Дана задача відноситься до задач класифікації в галузі лінгвістики з наперед заданим набором класів.

Одним з найпростіших методів оцінювання тональності тексту є метод машинного навчання на основі наївного баєсіва класифікатора (НБК) [3].

В основі НБК лежить теорема Баєса (2.1)

$$P(c \mid d) = \frac{P(c) \cdot P(d \mid c)}{P(d)} \qquad (2.1)$$

де,

– P(c|d) — імовірність, що документ d належить класу c (саме цю імовірність необхідно розрахувати);

– P(d|c) — імовірність зустріти документ d серед усіх документів класу c;

– P(c) — безумовна імовірність зустріти документ класу c в корпусі документів;

– P(d) — безумовна імовірність документа d в корпусі документів.

Метою класифікації є встановлення приналежності документу до якогось класу. Для цього необхідно не визначення безпосередньо імовірності, а визначення найбільш імовірного класу.

$$c_{map} = \arg\max_{c \in C} \frac{P(c) \cdot P(d \mid c)}{P(d)} \tag{2.2}$$

Тобто, необхідно обрахувати імовірності для кожного класу і обрати той, який має найбільшу імовірність. Так як знаменник (P(d)) є константою, то він не впливає на ранжування класів за імовірністю та може не враховуватися при обчисленнях (2.3).

$$c_{map} = \arg\max_{c \in C} \left[P(c) \cdot P(d \mid c) \right] \tag{2.3}$$

Для реалізації Баєсівского класифікатора необхідно сформувати навчальну вибірку, в якій співставленні текстові документи та їх класи. На етапі класифікації буде необхідна наступна статистика з вибірки:

- відносна частота класів в корпусі документів (як часто зустрічаються документи того, чи іншого класів);
- сумарна кількість слів в документах кожного класу;
- відносна частота слів а межах кожного класу;

– розмір словника вибірки (кількість унікальних слів у вибірці).

Модель класифікатора – це сукупність даної інформації. На етапі класифікації необхідно для кожного класу обчислити значення (2.11) та обрати максимальне з них (2.10).

$$\log\frac{D_c}{D}+\sum_{i=1}^{n}\log\frac{W_{ic}+1}{k+\sum_{j=1}^{k}(W_{jc})} \qquad (2.11)$$

де:

– D_c – кількість документів в навчальній вибірці, що належать класу c;
– D – загальна кількість документів в навчальній вибірці;
– k – кількість унікальних слів у всіх документах навчальної вибірки;
– $\sum_{j=1}^{k}(W_{jc})$ – сумарна кількість слів в документах класу c в навчальній вибірці;
– W_{ic} – скільки разів, і-е слово зустрічається в документах класу c в навчальній вибірці.

ХІД ВИКОНАННЯ

1. **Формування початкового датасету**

   Було взято по 50 анекдотів(частота класу в корпусі = 0,5) з двох категорій варіанту завдання (погода, політика). У таблиці 1.1. наведено вигляд файу .csv.

   Таблиця 1.1 - датасет

	Номер	Текст	Категорія
0	0	В хорошую погоду меня от окна даже за уши не о...	0
1	1	Как правило "левые" не правы.	1
2	2	Количество новостей про осадки превысило месяч...	0
3	3	Глупые народы, прикрывая свою глупость, воюют ...	1
4	4	Как объяснить знакомому испанцу, что если идёт...	0
...	...	...	...
95	95	Депутат, которого обвинили в получении второго...	1
96	96	А вот вы задумывались, что притопывая и подпры...	0
97	97	В день юбилея Ленина Зюганов заходит в бар и г...	1
98	98	Прогноз погоды - это такой вид новостей, к кот...	0
99	99	- Найди и принеси мне то, чего на белом свете,...	1

100 rows × 3 columns

2. **Дослідження корпусу документів**

   По двом заданим категоріям проводиться кількісний аналіз слів по документам класу категорії. На рисунка 2.1-2.2 приводяться гістограми кількісних частот слів класу, включно зі стоп словами.

Рис. 2.1 - гістограма кількості вживаних слів у категорії “Погода”

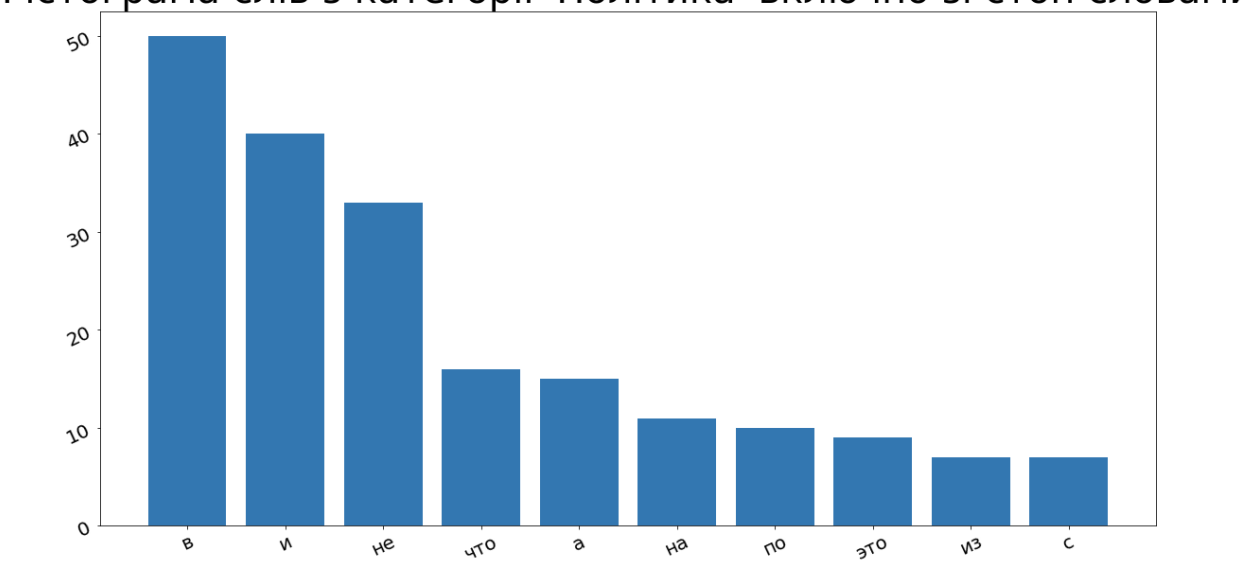

Рис. 2.2 - гістограма кількості вживання слів у категорії “Політика”

На рисунках 2.3-2.4 показані гістограми кількості вживання слів двух категорій без стоп-слів.

Рис. 2.3 - гістограма кількості вживання слів у категорії “Погода”

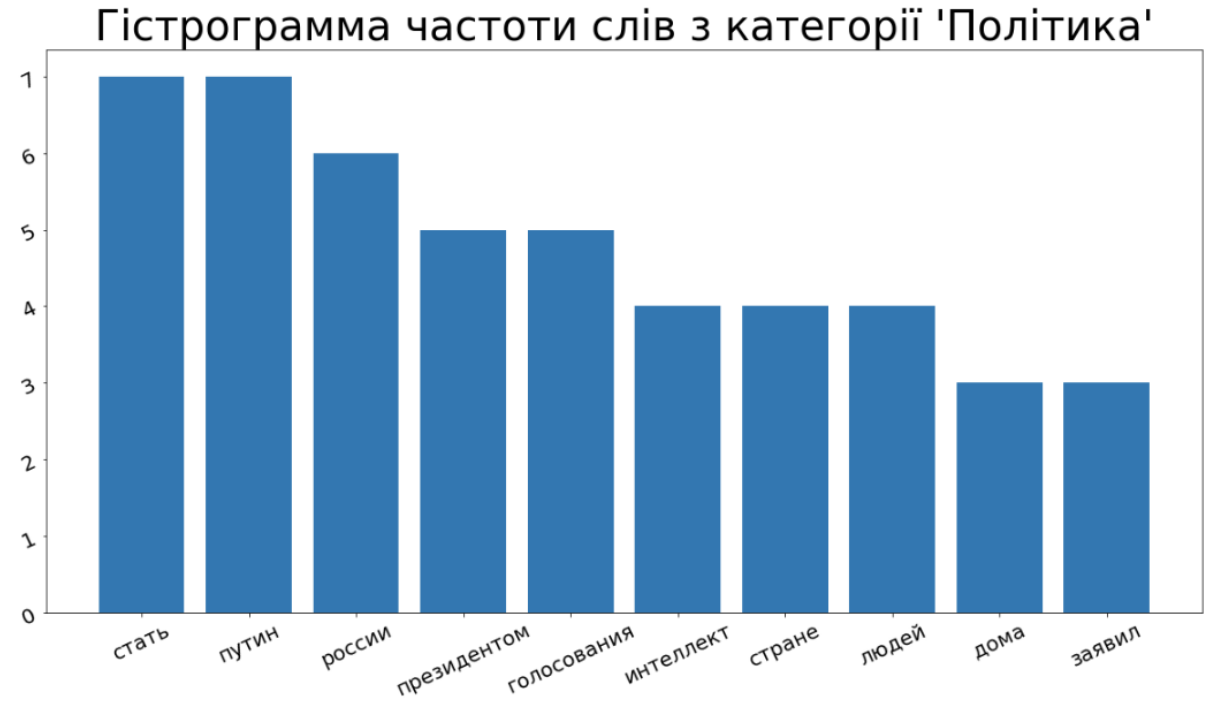

Рис. 2.4 - гістограма кількості вживання слів у категорії “Політика”

3. Дослідження впливу розміру навчальної вибірки до ймовірності потрапляння документу до классу

У таблиці 3.1 наведено значення ймовірностей відношення документу до классу відносно розміру початкової вибірки.

Таблиці 3.1 - ймовірності

	Навчальні вибірки	Ймовірність
0	Розмір: 10 x 10	0.600
1	Розмір: 20 x 20	0.625
2	Розмір: 30 x 30	0.725

На рисунку 3.1 продемонстровано графік оснований на таблиці 3.1.

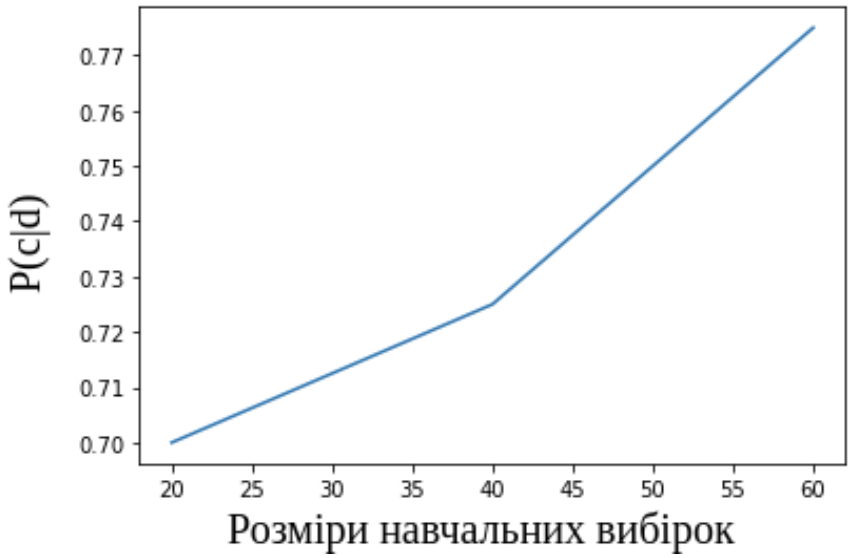

Рис. 3.1 - графік зростання ймовірності відносно збільшення розміру навчальної вибірки

Висновок:

Була досліджена задача визначення тональності тексту за допомогою використання наївного Баєсіва класифікатора.

При дослідженні впливу розміру навчальної вибірки на визначення ймовірності потрапляння документу до певного класу спостерігається кореляція - при збільшенні розміру навчальної вибірки збільшується і якість визначення ймовірності.