НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Методи машинного навчання в сентимент аналізі текстової інформації

виконала: Рудзевич Анна-Марія Павлівна

студентка групи КА-83МН

науковий керівник: канд. техн. наук Малишевський О.Г.

Об’єкт дослідження

- Сентимент аналіз текстової інформації

Предмет дослідження

- Алгоритми машинного навчання з учителем для задач визначення тональності тексту

Мета дослідження

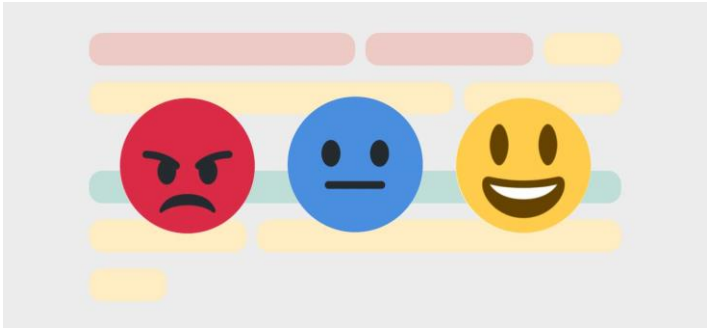
- Метою даної роботи є порівняння алгоритмів машинного навчання для аналізу тональності тексту на українській та російській мовах.

Актуальність

- Відсутність єдиного підходу до вирішення задачі сентимент аналізу
- Широке застосування у різних сферах (маркетинг, соціологія, політика тощо)
- Наявність великої кількості текстових даних на цифрових носіях
- Алгоритми машинного навчання мають більшу точність та узагальнюючу здатність в умовах великих обсягів даних
- Машинне навчання дозволяє знаходити структуру в даних, що дає змогу ефективно вдосконалювати моделі

Загальні поняття

- **Сентимент аналіз** (аналіз тональності тексту) — це розділ глибинного аналізу даних (data mining) і область комп'ютерної лінгвістики, що займається вилученням думок та емоцій з текстових документів.



Підходи до сентимент аналізу

- підхід на основі правил;
- підхід на основі тональних словників;
- підхід з використанням методів машинного навчання (із вчителем і без нього).

Рівні сентимент аналізу

- **рівень документу.** Цей рівень аналізу передбачає, що кожен документ висловлює думку лише про один суб'єкт (наприклад продукт) і не є придатним для документів, які оцінюють декілька суб'єктів.
- **рівень речення.** Цей рівень аналізу тісно пов'язаний з класифікацією тексту на суб'єктивність, який визначає чи є висловлення об'єктивним (тобто таким, що висловлює фактичну інформацію), чи суб'єктивним (думку, погляд)
- **рівень об'єкту та аспекту**. Наприклад «Мені все одно подобається цей фільм, хоча спецефекти погані». Загальний настрій позитивний, хоча все речення явно не є повністю позитивним. Якщо точніше, речення має позитивний сентимент стосовно фільму (об'єкт), але негативний стосовно його спецефектів (аспект фільму).

Алгоритми машинного навчання

Машинне навчання — підрозділ штучного інтелекту, який вивчає методи побудови моделей, які здатні навчатися, та алгоритмів для їх побудови та навчання. Метою машинного навчання є часткова чи повна автоматизація розв'язання складних професійних задач в найрізноманітніших областях.

Наведемо базові позначення:

- X — простір об'єктів;
- Y — простір відповідей;
- $x = (x^1, \ldots, x^d)$ — опис ознак об'єкта;
- $X = (x_i, y_i)_{i=1}^{l}$ — навчальна вибірка;
- $a(x)$ — алгоритм, модель;
- $Q(a, X)$ — функціонал помилки алгоритму a на вибірці X;
- навчання: $a(x) = arg \min_{a \in A} Q(a, X)$.

Виявлення ознак з тексту

Щоб мати можливість працювати з текстом в рамках машинного навчання, його необхідно перевести до векторного вигляду. Розглянемо основні алгоритми для векторизації тексту.

Bag-of-Words	TF-IDF	Word2Vec

Bag-of-Words

Модель **«мішок слів»** є спрощенням подання, що використовується в обробці природних мов і інформаційному пошуку. У цій моделі текст (наприклад, речення або документ) представляється у вигляді мішка (мультимножини) його слів, не беручи до уваги граматику і навіть порядок слів, але зберігаючи множинність.

Document 1

The quick brown fox jumped over the lazy dog's back.

Document 2

Now is the time for all good men to come to the aid of their party.

Term	Document 1	Document 2
aid	0	1
all	0	1
back	1	0
brown	1	0
come	0	1
dog	1	0
fox	1	0
good	0	1
jump	1	0
lazy	1	0
men	0	1
now	0	1
over	1	0
party	0	1
quick	1	0
their	0	1
time	0	1

TF-IDF
(Term frequency – inverse document frequency)

$$TF(t) = \frac{n_i}{\sum_k n_k},$$

де n_i — кількість разів слово t зустрічається у документі, а в знаменнику кількість всіх слів у документі.

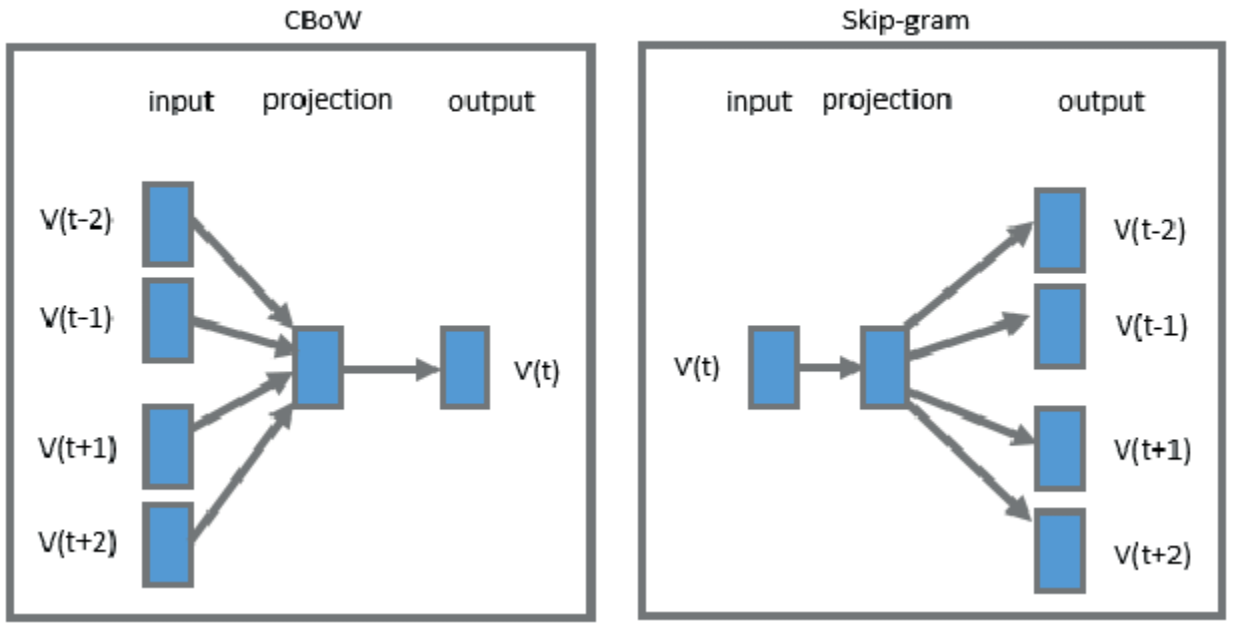
$$IDF(t) = log\frac{|D|}{|d_i \supset t_i|},$$

де $|D|$ — кількість документів в колекції, а $|(d_i \supset t_i)|$ — кількість документів, які включають слово t_i $(n_i \neq 0)$.

$$TF - IDF = TF(t) * IDF(t).$$

Векторизація слів (Word2Vec)

Word2Vec — загальна назва для сукупності моделей на основі штучних нейронних мереж, призначених для отримання векторних представлень слів на природній мові. Використовується для аналізу семантики природних мов.



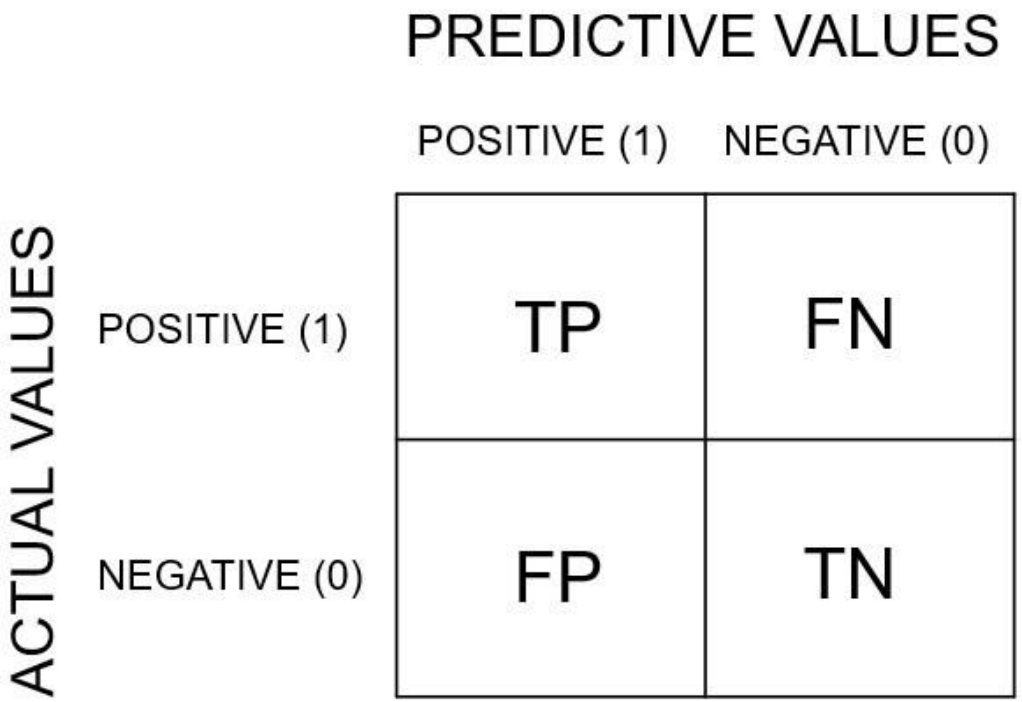
Метрики якості

Якість роботи моделей будемо оцінювати за метрикою F1.

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$



Опис даних

Навчальна вибірка складається з близько 20 тис. записів для обох кандидатів.

Класи є незбалансованими. Для В. Зеленського позитивний клас складає 83%, а для П. Порошенка — 38%.

text	sentiment
Лучший , Вы будущие нашей страны)	1.0
UAUAUAUA🤘🤘🤘UAUAUAUA	1.0
Мы за тебя👍	1.0
Надеяться нужно на себя и на свой голос ! ! ! А,то...	1.0
Тримайтеся і ми всіх переможемо!!!!!	1.0
Слушаешь, закрываешь глаза и слышишь Юлю Тимош...	0.0

Характеристики тексту повідомлень

Обчислимо наступні показники текстової змінної (коментаря) для кожного з кандидатів:

- кількість слів у коментарі;
- посилання на аккаунти інших користувачів (починаються з @);
- кількість тегів (починаються з #);
- кількість слів, написаних заголовними літерами. Інколи використовується, щоб висловити сильні (негативні) емоції;
- кількість знаків питання та знаків оклику;
- кількість посилань у коментарі (починаються з http(s));
- кількість емотиконів.

Попередня обробка тексту

- видалимо згадки, оскільки вони не несуть емоційного навантаження;
- видалимо знак хештега, але не сам хештег, оскільки він може містити інформацію;
- переведемо всі слова до нижнього регістру;
- видалимо всі розділові знаки, включаючи знаки запитання та знаки оклику;
- видалимо URL-адреси, оскільки вони не містять корисної інформації;
- конвертуємо емоджі в одне слово;
- видалимо цифри;
- видалимо стоп-слова;
- застосуємо стемінг, щоб зберегти основу слова без закінчення чи суфіксів.

Приклад даних після обробки

	text	count_words	count_mentions	count_hashtags	count_capital_words	count_excl_quest_marks	count_urls	count_emojis
0	fire	0	0	0	0	0	0	1
2	владимир президент smilingfacewithsmilingey uk...	3	0	0	0	3	0	1
4	посад grimacingfac	6	0	0	0	2	0	1
6	крас redheart fire	1	0	0	0	1	0	2
8	зе став лайк	5	0	0	0	1	0	0

Застосуємо алгоритми МН

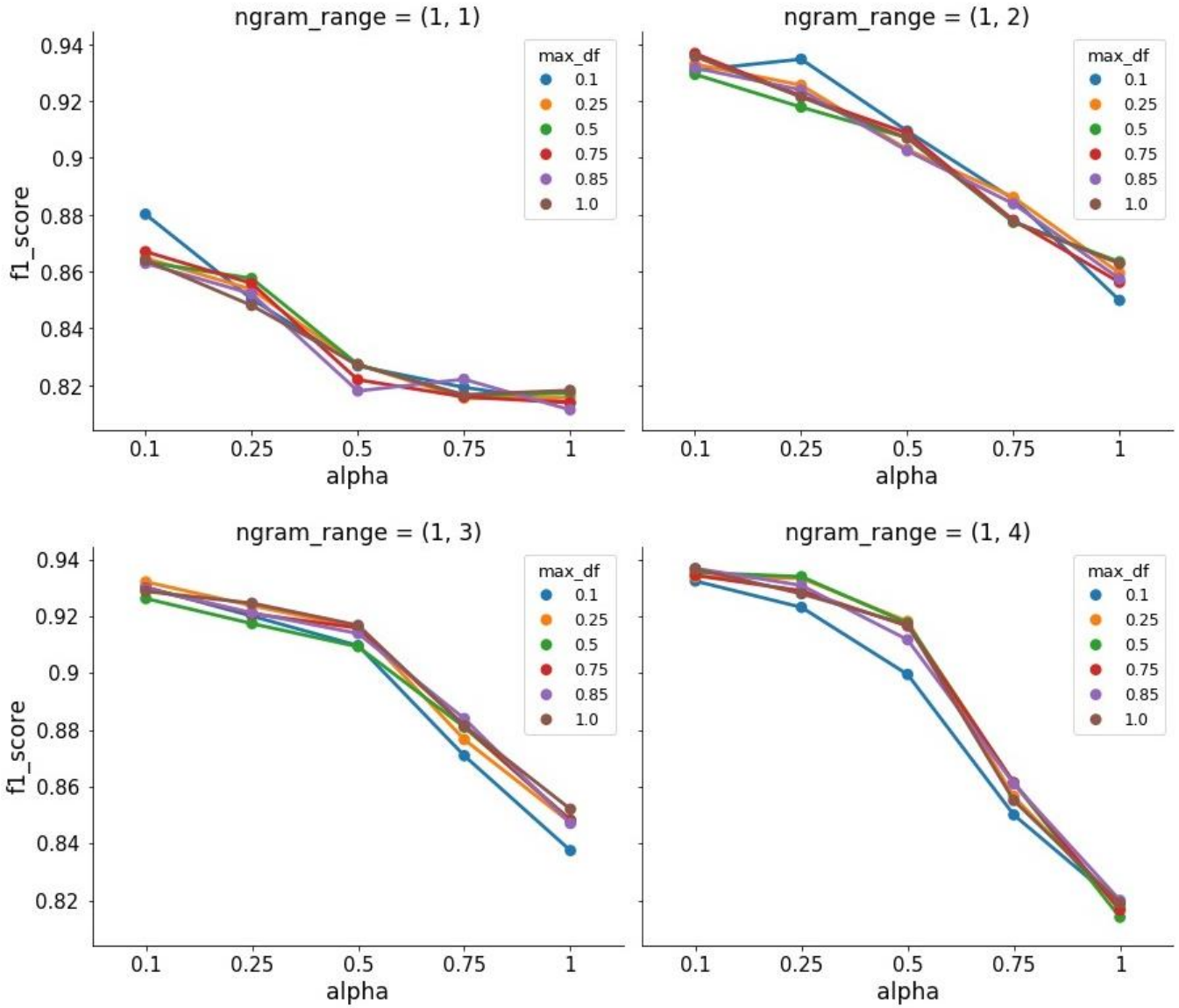
Назва алгоритма	Застосовані векторизатори
Наївний Байєсівський класифікатор	Bag-of-Words, TF-IDF
Метод опорних векторів	Bag-of-Words, TF-IDF
Згорткова нейронна мережа	Word2Vec

Наївний Байєсів класифікатор

Bag-of-Words

Зафіксуємо розбиття кросвалідації рівним **CV = 5**. Будемо шукати:

- коефіцієнт згладжування ***alpha:*** [0.1, 0.25, 0.5, 0.75, 1],
- **ngram_range**: [(1,1), (1,2), (1,3), (1,4)],
- максимальний поріг відсікання термінів **max_df** : [0.1, 0.25, 0.5, 0.75, 0.85, 1].

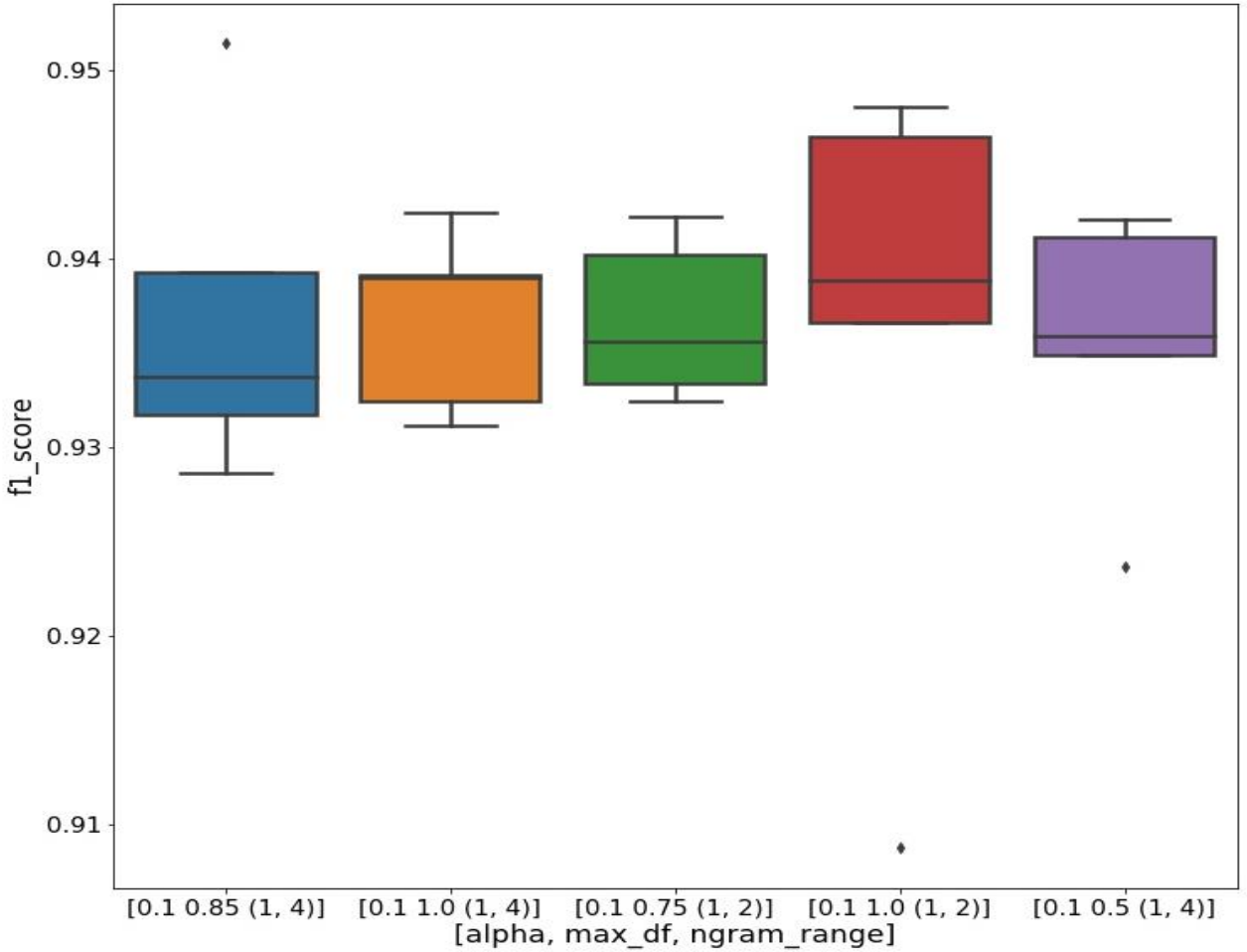


Наївний Байєсівський класифікатор з векторизатором Bag-of-Words

Найкраща модель:

- *alpha = 0.1*,
- *max_df = 0.85*,
- *ngram_range = (1, 4).*

Точність F1-score ≈ *0.94*.



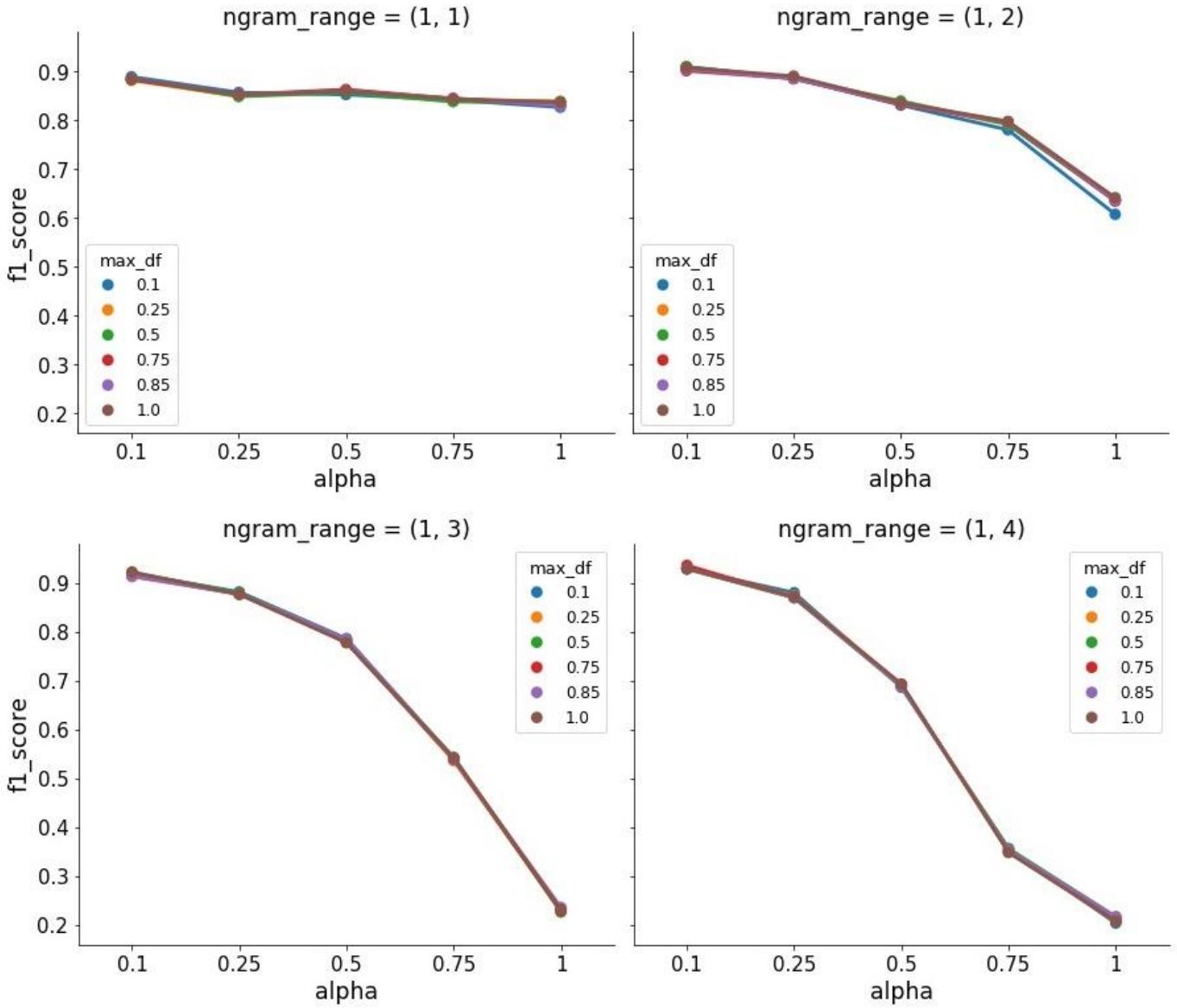
Кращі 5 моделей зі 120 побудованих з точки зору середньої точності за кросвалідацією.

Наївний Байєсів класифікатор

TF-IDF

Будемо шукати:

- коефіцієнт згладжування ***αlpha:*** [0.1, 0.25, 0.5, 0.75, 1],
- **ngram_range**: [(1,1), (1,2), (1,3), (1,4)],
- максимальний поріг відсікання термінів **max_df** : [0.1, 0.25, 0.5, 0.75, 0.85, 1].

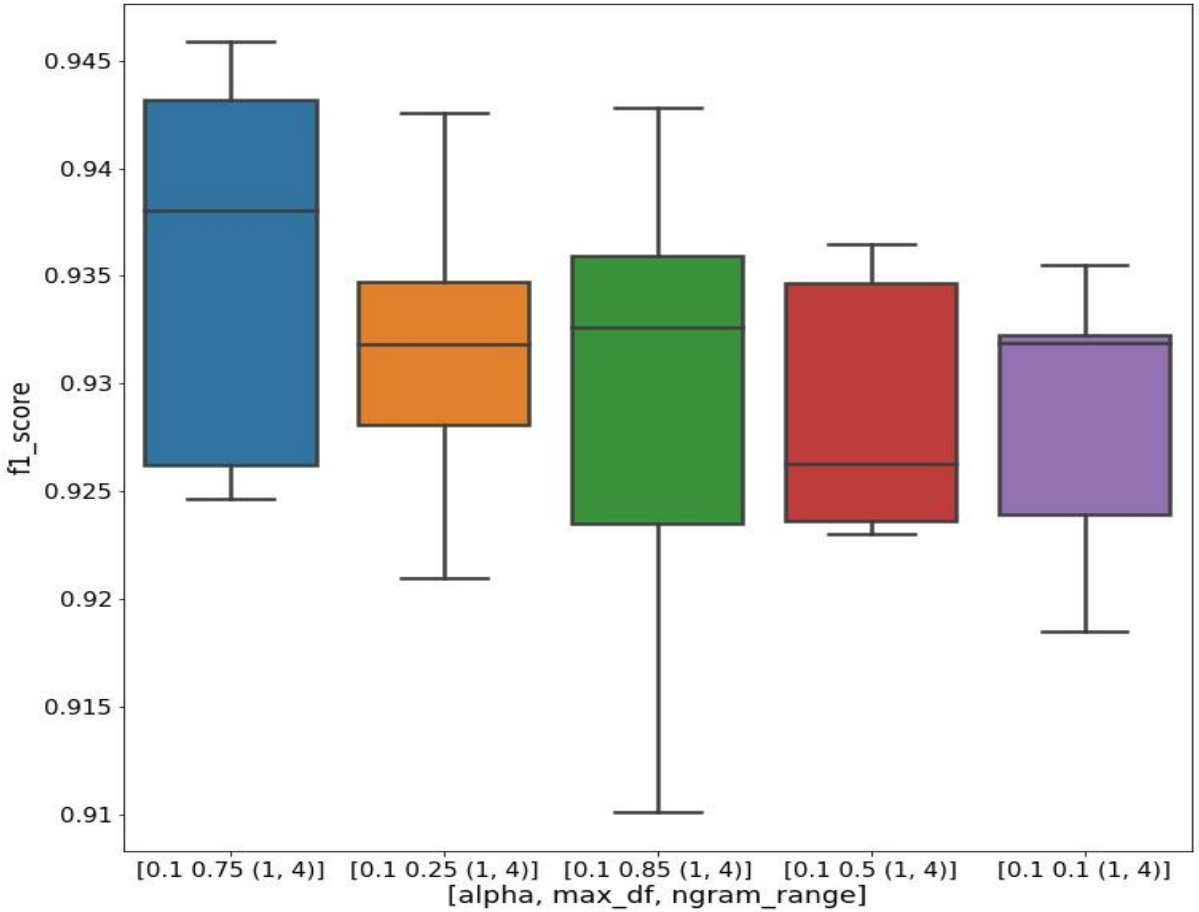


Наївний Байєсівський класифікатор з векторизатором TF-IDF

Найкраща модель:

- *alpha = 0.1*,
- *max_df = 0.75*,
- *ngram_range = (1, 4).*

Точність F1-score ≈ *0.932*.



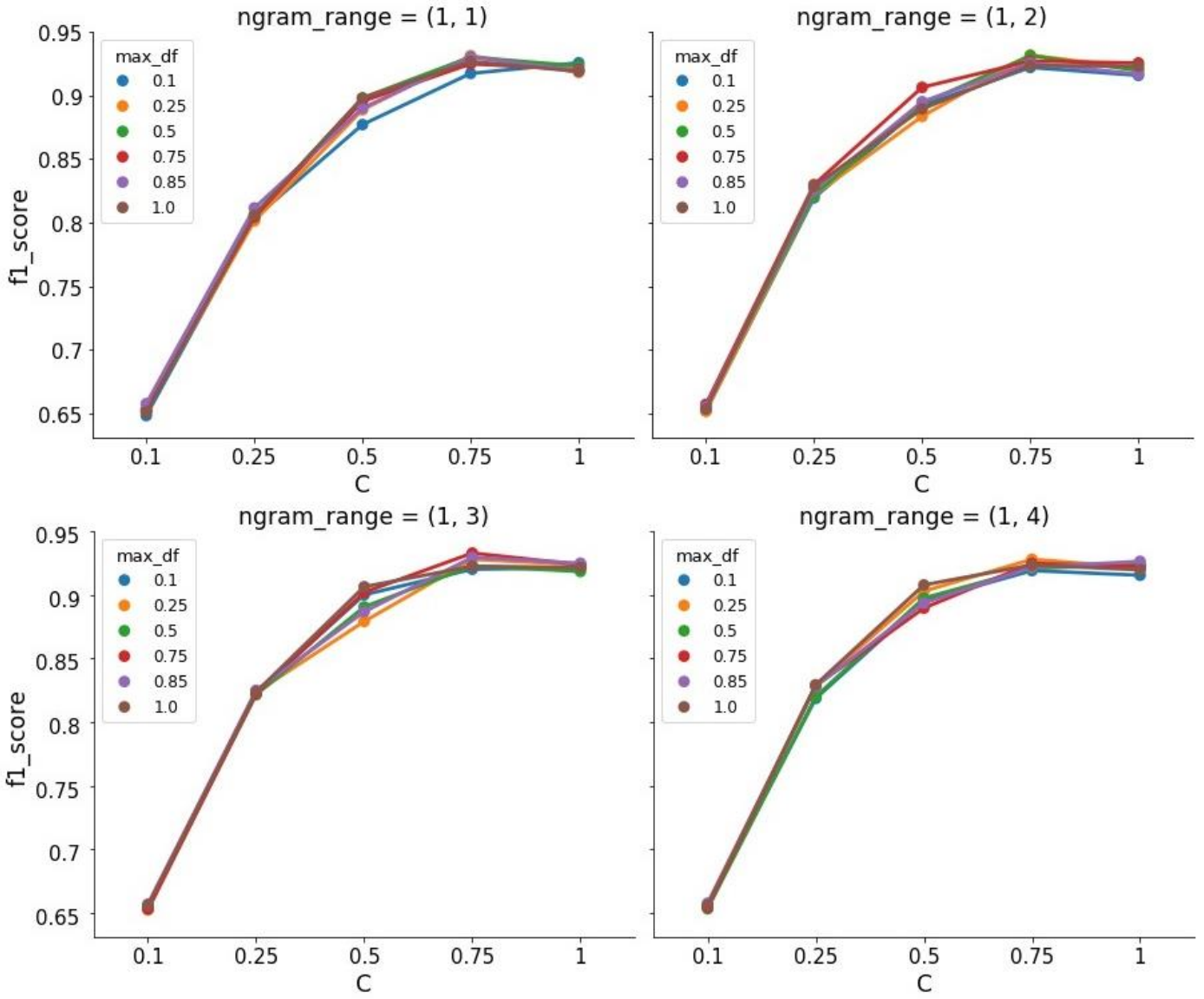
Кращі 5 моделей зі 120 побудованих з точки зору середньої точності за кросвалідацією.

Метод опорних векторів

Bag-of-Words

Будемо шукати:

- коефіцієнт регуляризації ***C:*** [0.1, 0.25, 0.5, 0.75, 1],
- **n-gram_range:** [(1,1), (1,2), (1,3), (1,4)],
- максимальний поріг відсікання термінів **max_df:** [0.1, 0.25, 0.5, 0.75, 0.85, 1].

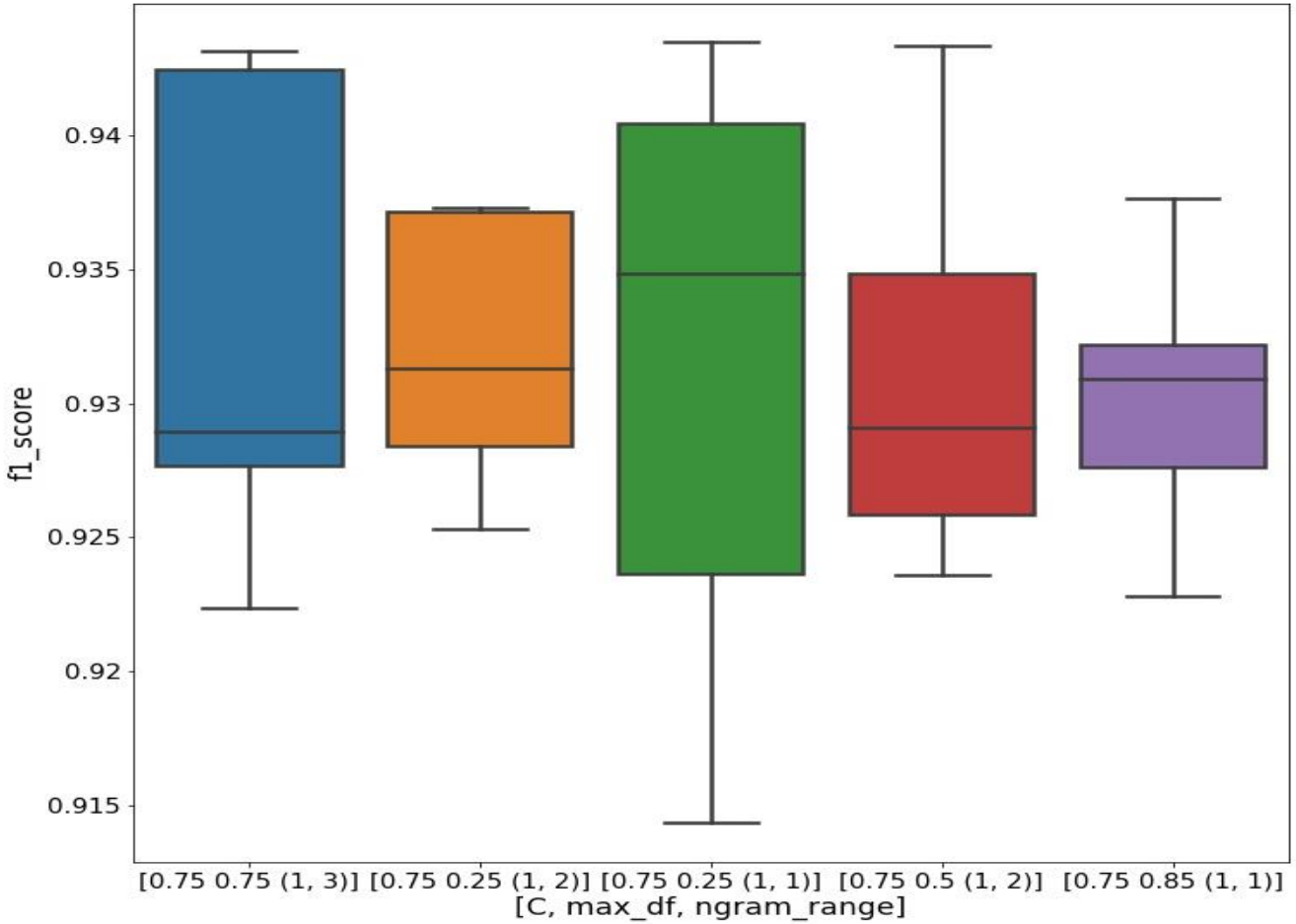


SVM з векторизатором Bag-of-Words

Найкраща модель:

- *C = 0.75*,
- *max_df = 0.75*,
- *ngram_range = (1, 3).*

Точність F1-score ≈ *0.945*.



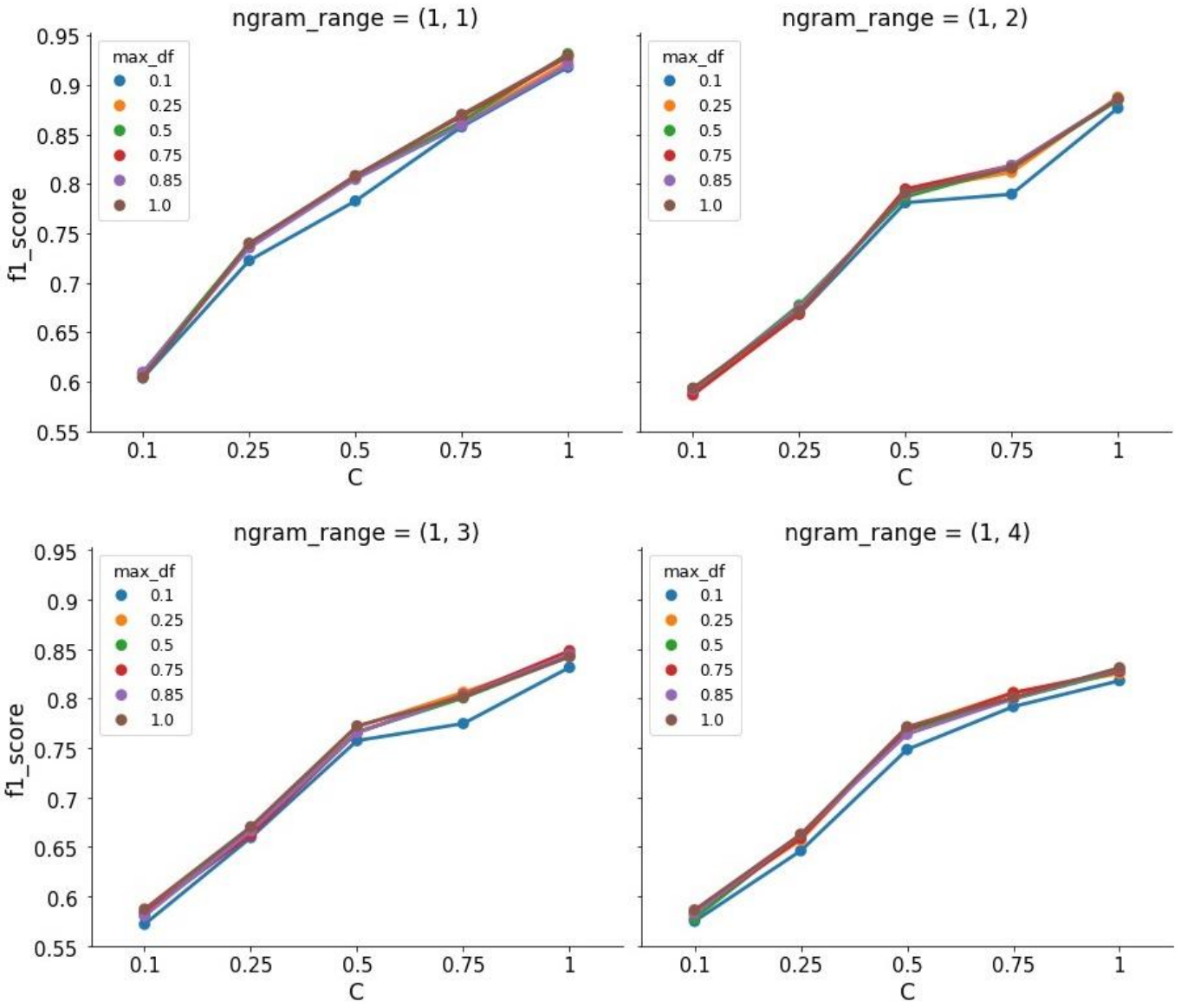
Кращі 5 моделей зі 120 побудованих з точки зору середньої точності за кросвалідацією.

Метод опорних векторів

TF-IDF

Будемо шукати:

- коефіцієнт регуляризації ***C:*** [0.1, 0.25, 0.5, 0.75, 1],
- **n-gram_range:** [(1,1), (1,2), (1,3), (1,4)],
- максимальний поріг відсікання термінів **max_df:** [0.1, 0.25, 0.5, 0.75, 0.85, 1].

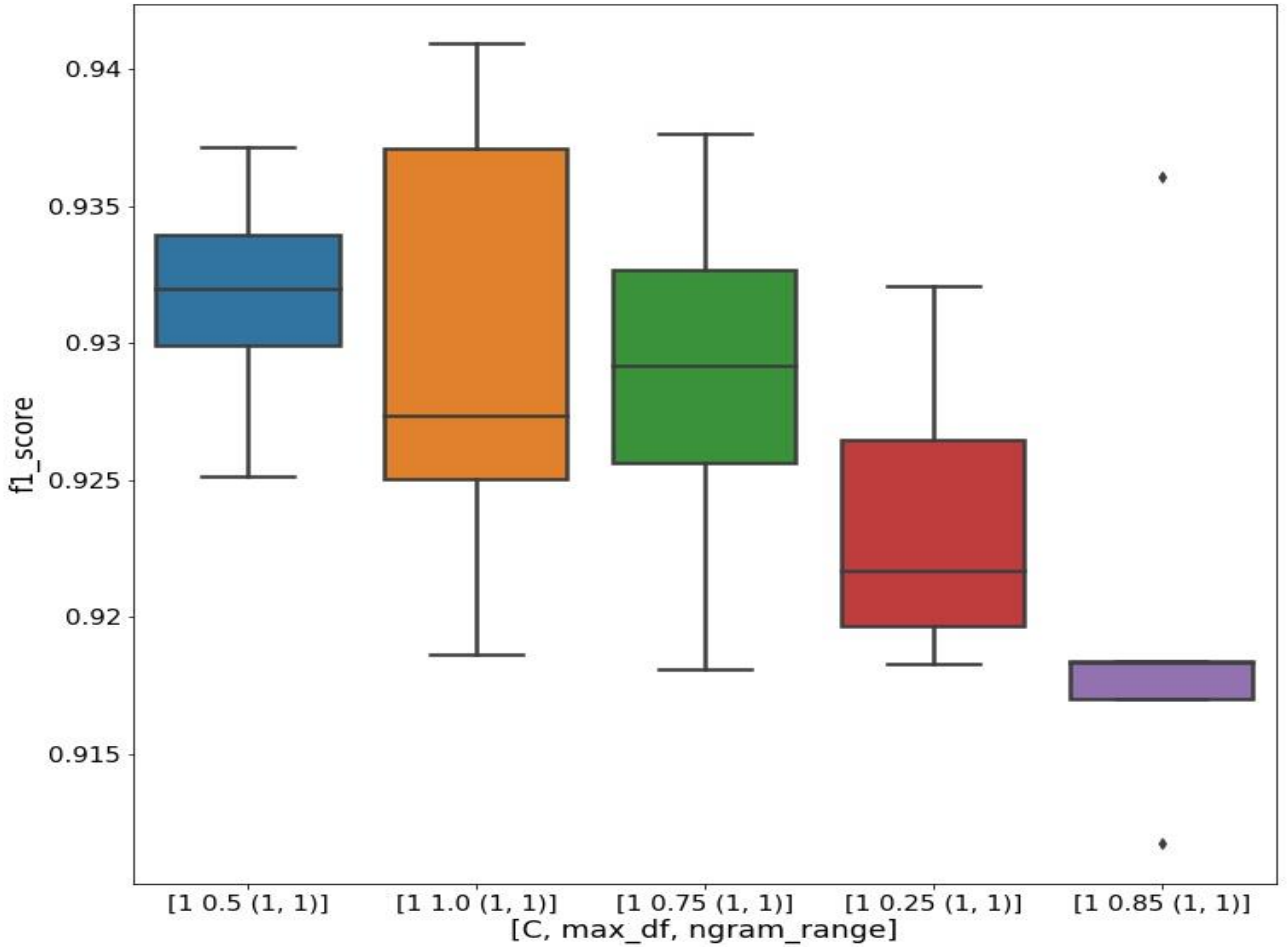


SVM з векторизатором TF-IDF

Найкраща модель:

- *C = 1*,
- *max_df = 0.5*,
- *ngram_range = (1, 1).*

Точність F1-score ≈ *0.934*.



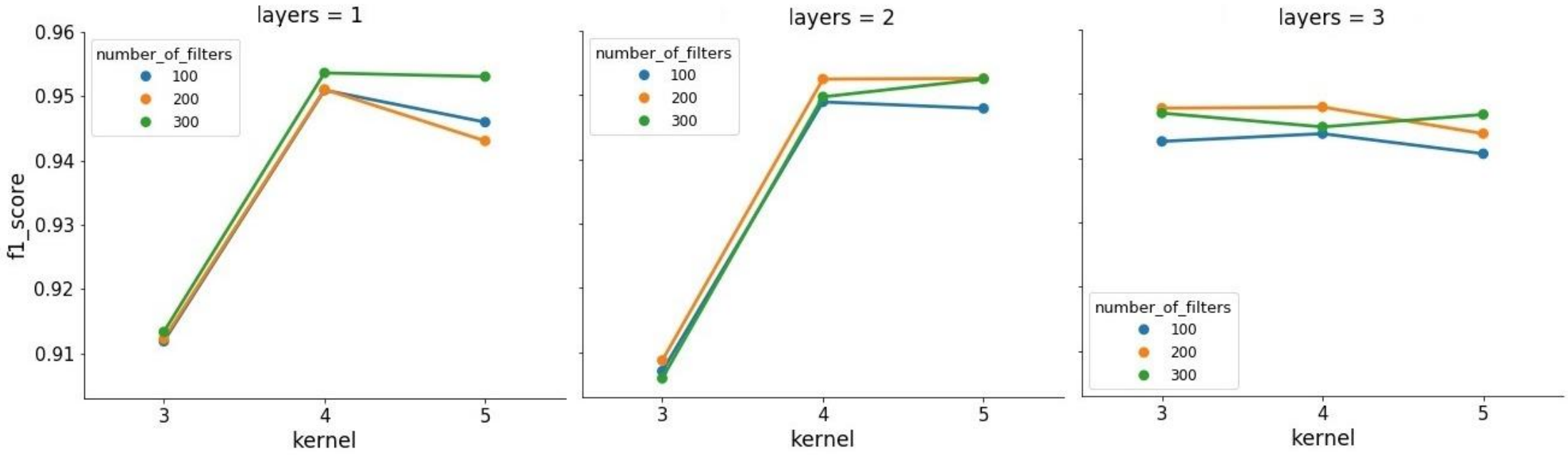
Кращі 5 моделей зі 120 побудованих з точки зору середньої точності за кросвалідацією.

Згорткова нейронна мережа

Word2Vec

Задамо наступні параметри для ЗНМ: **функція активації** ReLU, **дропаут** = 0.4, **розмір батча**: 100.

Будемо шукати кількість шарів згортки **layers:** [1, 2, 3], **number_of_filters**: [100, 200, 300], розмір ядра **kernel:** [3, 4, 5].

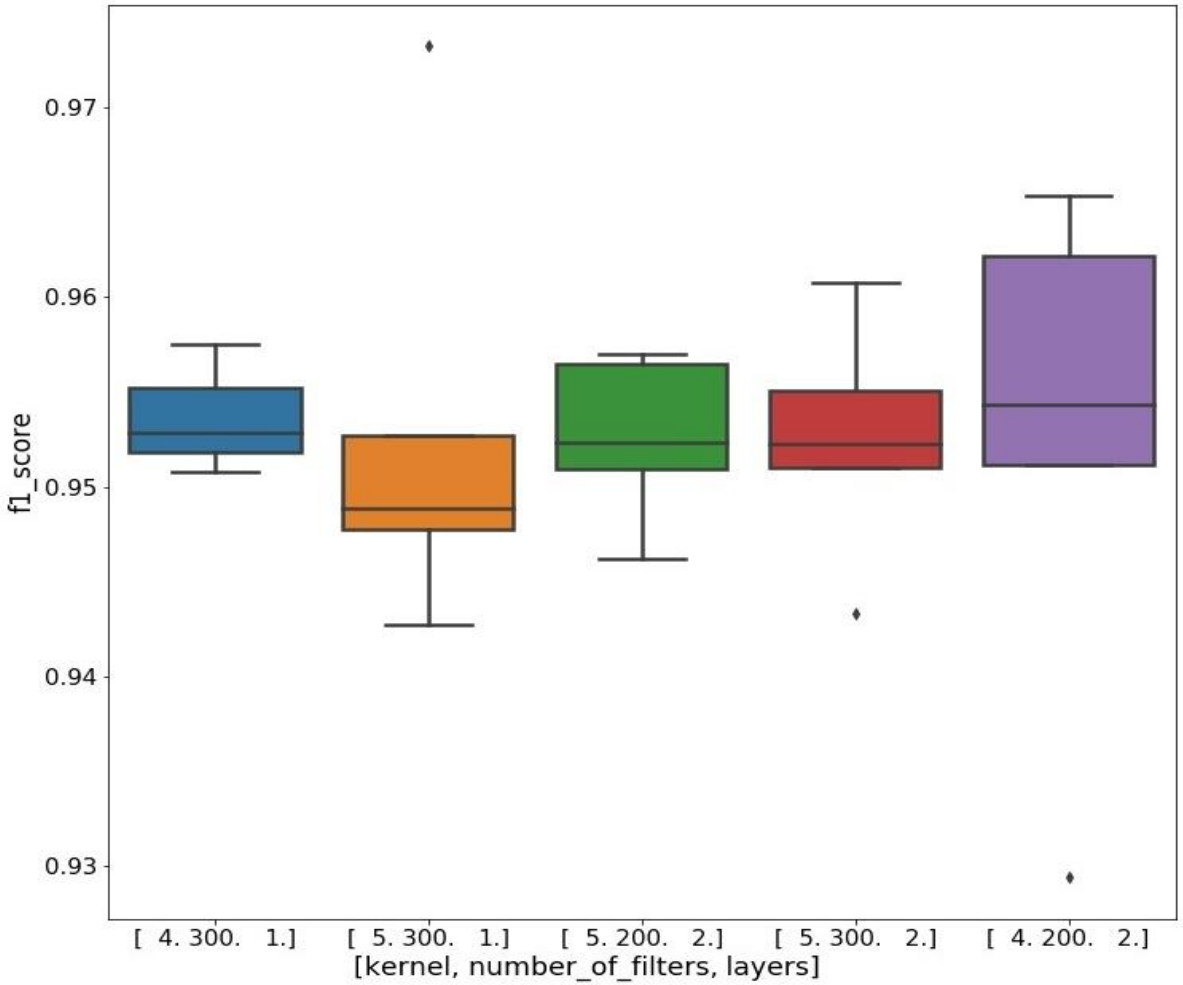


ЗНМ з векторизатором Word2Vec

Найкраща модель:

- *kernel = 4*,
- *number_of_filters = 300*,
- *layers=1.*

Точність F1-score ≈ *0.956*.



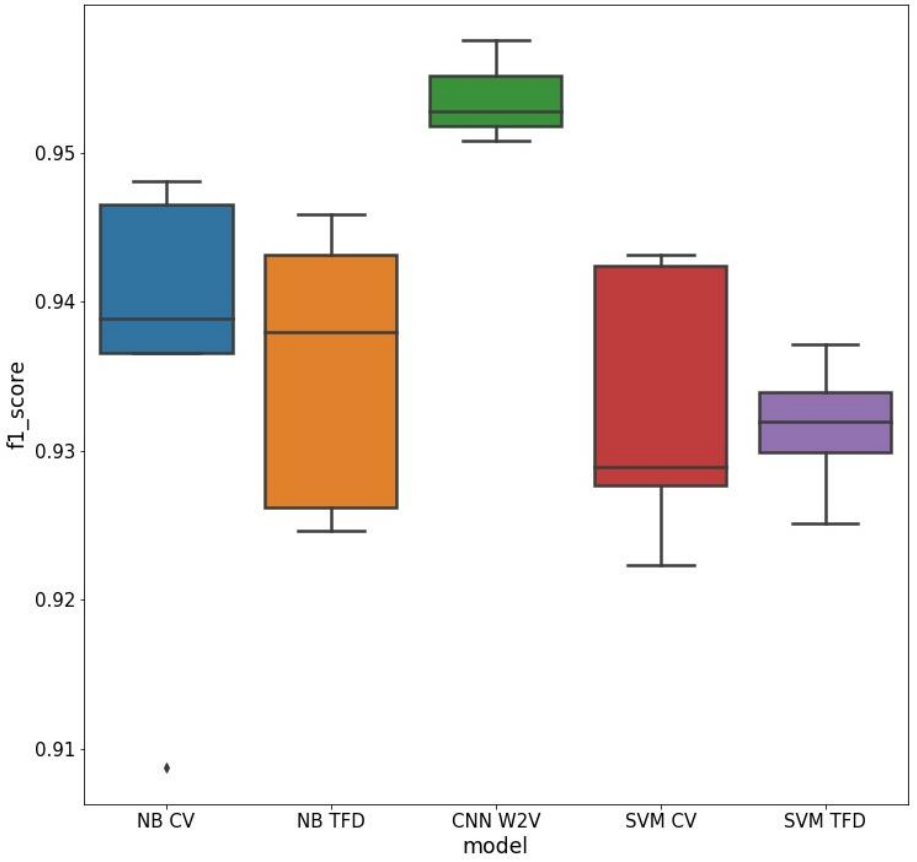
Кращі 5 моделей зі 120 побудованих з точки зору середньої точності за кросвалідацією.

Порівняння алгоритмів

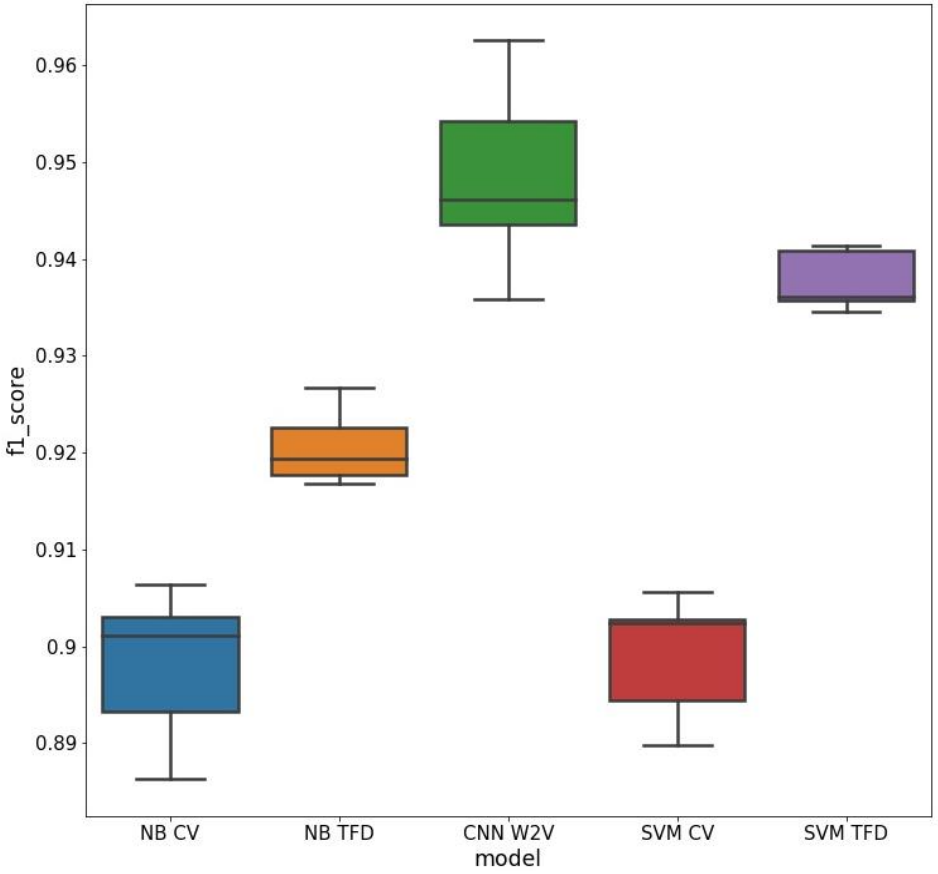
	В. Зеленський	П. Порошенко
Naïve Bayes (Bag-of-Words)	93.9%	90%
Naïve Bayes (TF-IDF)	93.2%	92%
SVM (Bag-of-Words)	93.5%	90%
SVM (TF-IDF)	93.2%	94%
Convolutional neural network	95.6%	95.5%

- Знайдемо найкращу з кращих моделей для кожного з кандидатів

Краща модель

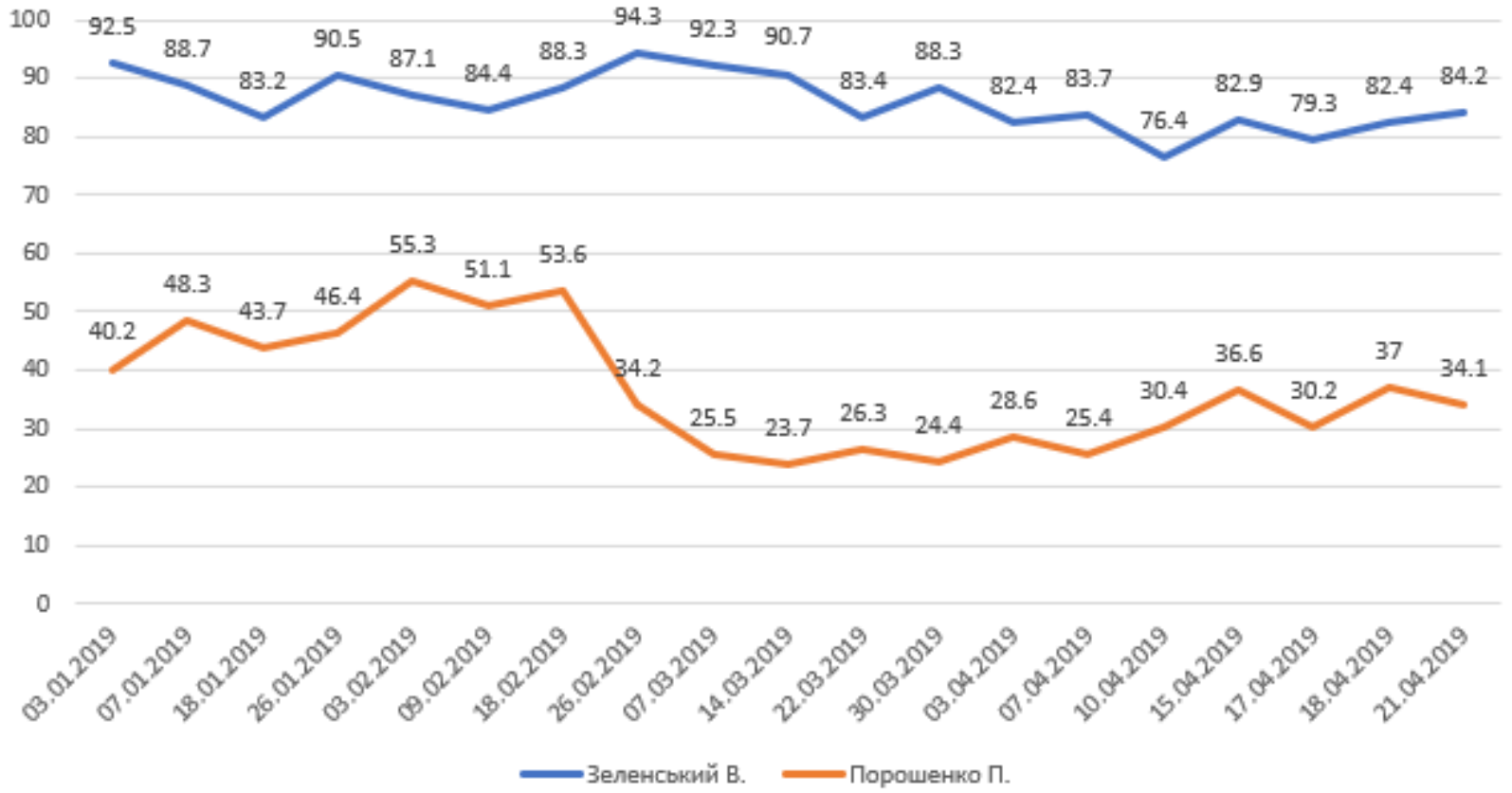


CNN з параметрами kernel = 4, layers=1 number_of_filters = 300, є кращою моделлю для класифікації коментарів адресованих В. Зеленському



CNN з параметрами kernel = 3, layers=1 number_of_filters = 300 є кращою моделлю для класифікації коментарів адресованих П. Порошенко

Застосування моделей



Застосуємо моделі для визначення зміни прихильності громадськості до кандидатів. Для цього визначимо сентимент коментарів під публікаціями кандидатів в період з 03.01.2019 по 21.04.2019. Відобразимо на графіку відсоток позитивних коментарів.

Висновки

- Досліджено підходи до вирішення задачі сентимент аналізу
- Досліджено методи машинного навчання, які застосовують для сентимент аналізу
- Досліджено етапи попередньої обробки тексту
- Наведено дослідні дані, описано і програмно реалізовано алгоритми машинного навчання
- Порівняно ефективність моделей за метрикою F1
- Проведено дослідження зміни громадської думки під час виборчої кампанії 2019 року

Апробація результатів дослідження

- За результатами проведеного дослідження було укладено статтю “Методи машинного навчання в сентимент аналізі текстової інформації на прикладі настроїв користувачів стосовно кандидатів у Президенти України 2019” і опубліковано в міжнародному науково-технічному журналі “Системні дослідження та інформаційні технології”.

Перспективи подальших досліджень

Як перспективу до подальших досліджень ми вбачаємо:

- проведення сентимент аналізу на іншому рівні: на рівні речення та/або на рівні об’єкта-аспекта.
- дослідження ефективності застосування методів машинного навчання без учителя.

Дякую за увагу!