ISSN 2224-087X. Електроніка та інформаційні технології. 2020. Випуск 13. С. 3–18
Electronics and information technologies. 2020. Issue 13. P. 3–18

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В НАУКОВИХ ДОСЛІДЖЕННЯХ

УДК 519.765:519.767:004.89 DOI: https://doi.org/10.30970/eli.13.1

ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА СЕМАНТИЧНИХ ОЗНАК В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ТЕКСТОВИХ ДАНИХ

Б. М. Павлишенко

Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна
b.pavlyshenko@gmail.com

У роботі досліджено використання семантичних ознак в інтелектуальному аналізі текстових даних, зокрема у класифікації текстових документів. Як семантичні ознаки, розглянуто семантичні та тематичні поля, складові сингулярного розкладу матриці TF-IDF та складові латентного розміщення Діріхле. Класифікаційний аналіз здійснено за допомогою алгоритму Random Forest та алгоритмів глибинного навчання нейромереж із різною структурою із використанням двонаправлених шарів із довгою короткочасною пам'яттю (LSTM). Використання широкого класу семантичних ознак у задачах інтелектуального аналізу диверсифікує аналітичні підходи і збільшує простір ознак в аналітичних задачах, що є важливим при невеликій кількості даних та при аналізі нестаціонарних процесів.

Ключові слова: аналіз текстів, семантичні ознаки текстів, класифікація текстів, нейронні мережі.

Вступ

Аналіз текстових даних – актуальна проблема сучасних інформаційних технологій. Значних успіхів досягнуто в текстовій аналітиці з використанням сучасних методів машинного навчання, зокрема, глибинного навчання. Актуальною проблемою є пошук нових ознак, на основі яких можна проводити інтелектуальний аналіз даних, зокрема класифікацію текстових даних. Цікавим для дослідження є клас задач, у яких аргументами є як числові, так і текстові дані. Перспективним є використання ознак, які відображають семантику текстів. Складовими аналізу текстів є алгоритми класифікації та кластеризації текстових документів. У цих алгоритмах часто використовують векторну модель текстових документів, яка базується на представленні документів як векторів у фазовому просторі. Базис такого простору часто утворюють за допомогою частотно-дистрибутивних характеристик лексем текстового словника. Одна з основних проблем такого підходу зумовлена великою розмірністю аналізованого векторного простору. Також такий простір не дає можливості виділити задані семантичні складові в інтелектуальному аналізі текстів. У задачах аналізу текстового змісту актуальними є теорії лексичної семантики, зокрема вчення про семантичні поля. Семантичні поля розглядають як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких може бути ефективним у задачах кластеризації та класифікації текстових документів. Семантичні

поля глибоко вивчені у лінгвістичних працях, однак існує необхідність розробки формалізованих математичних моделей для їхнього впровадження в алгоритми інтелектуального аналізу текстових масивів. Поширеною є векторна модель текстових документів. Класичним підходом у формуванні текстових ознак є використання TF-IDF матриць [1]. В основі використання векторної моделі лежить статистична гіпотеза, яка полягає в тому, що статистичні характеристики використання слів відображають ті поняття, які люди мають на увазі в цих текстах [1]. Одним із методів представлення масиву документів у векторному просторі є організація векторів документів у вигляді матриці текстових частот типу лексеми-документи. Рядки таких матриць відповідають лексемам, а стовпці є векторами відповідних документів. Як приклад ієрархічно-організованої семантичної мережі можна розглядати систему WordNet, яка розроблена у Принстонському університеті [2,3]. Лексемний склад у цій системі організований у вигляді синсетів, під якими розуміють набори лексем синонімічного ряду, які є взаємозамінними у заданих контекстах. Бази даних WordNet створені експертами-лексикографами. Іменники, дієслова, прикметники та прислівники організовані у синсети – множини синонімів. Іменники та дієслова згруповані відповідно до семантичних полів. Семантична структурна організація лексемного складу словника може використовуватись у відповідних алгоритмах класифікації та кластеризації текстових об’єктів з точки зору зменшення розмірності задач аналізу та виявлення нових семантичних зв’язків в онтології предметної області, до якої відносять аналізований масив текстів. У [4,5] розглянуто методи класифікаційного аналізу текстових документів. У роботі [6] проаналізовано використання лексемних полів у інтелектуальному аналізі текстових масивів. Семантичні поля розглянуто як групи лексем, об’єднаних спільним поняттям. У [7] запропоновано модель кластеризації текстових документів у семантичному просторі, яка дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності ніж у просторі, утвореному частотними характеристиками лексемного складу текстової вибірки. У роботі [8] проказано, що сингулярний розклад матриці семантичних ознак типу “частоти_семантичних_полів–документи” дає можливість аналізувати текстові документи у новому просторі семантичних концептів. Семантичні групування слів відображають системність лексики. У [9] запропоновано концепцію семантичних доменів, яка розширює поняття семантичних полів. Більшість визначень семантичної класифікації класів лексем є спорідненими, близькими до класичного визначення семантичного поля і базуються на моделі «мішка слів». Відмінності між цими визначеннями зумовлені різним рівнем диференціації семантичних понять, на основі яких утворюють лексемні об’єднання. У цій моделі розглядають сукупність слів текстових документів без врахування їх контекстуальної послідовності. В залежності від вибраної моделі та алгоритму об’єднання лексем можна отримати різні лексемні угрупування. На основі квантитативних характеристик кожного із таких угрупувань можна утворити додатковий вимір у семантичному просторі представлення текстових документів. Уведення цих додаткових вимірів може бути ефективним у задачах інтелектуального аналізу текстів, зокрема у класифікаційних задачах та задачах кластерного аналізу.

У даній статті проаналізовано використання семантичних ознак у класифікаційному аналізі різних типів текстових даних. Числове моделювання та візуалізація результатів здійснювались на мові Python у середовищі *Jupyter Notebook* із використанням відповідних пакетів, зокрема *numpy, pandas, scikit-learn, keras, gensim, matplotlib,*

seaborn. У роботі розглянуто формування семантичних ознак текстових даних, класифікаційний аналіз текстових документів та використання рекурентних нейронних мереж в аналітиці текстових даних.

Формування семантичних ознак текстових даних

Розглянемо формування семантичних ознак текстових документів, які можна використати в інтелектуальному аналізі текстових даних. Для аналізу було взято стандартизовану текстову вибірку 20-Newsgroups [10], до якої входить колекція із приблизно 20 тисяч документів близько 20 груп новин. Цю колекцію часто використовують у тестових задачах інтелектуального аналізу текстових масивів, зокрема у задачах класифікації та категоризації текстових масивів. Розглянемо використання різних ознак у методах машинного навчання. Як семантичні ознаки використаємо частоти семантичних та тематичних полів, а також характеристики на основі латентого семантичного аналізу (Latent Semanti Analysis, LSA) [11,12] та латентного розміщення Дирихле (Latent Dirichlet Allocation, LDA) [13]. Для формування семантичних ознак на основі семантичних полів було використано систему WordNet [2,3]. Семантичні поля у мережі WordNet представлено лексикографічними файлами, назви яких відображають основні семантичні значення лексем, які входять у склад цих файлів. У наших дослідженнях ми використали семантичні поля іменників та дієслів. Семантичні поля іменників складаються із 26 лексикографічних файлів: noun.Tops, noun.act, noun.animal, noun.artifact, noun.attribute, noun.body, noun.cognition, noun.communication, noun.event, noun.feeling, noun.food, noun.group, noun.location, noun.motive, noun.object, noun.person, noun.phenomenon, noun.plant, noun.possession, noun.process, noun.quantity, noun.relation, noun.shape, noun.state, noun.substance, noun.time, verb.body. Семантичні поля дієслів містять 15 лексикографічних файлів: verb.change, verb.cognition, verb.communication, verb.competition, verb.consumption, verb.contact, verb.creation, verb.emotion, verb.motion, verb.perception, verb.possession, verb.social, verb.stative, verb.weather. Частотна характеристика семантичного поля розраховувалась як сума текстових частот слів, які входять у це поле, розділена на суму текстових частот лексем всіх семантичних полів, які розглядаються. Розрахунки тематичних полів здійснювалися на тренувальній вибірці даних, яка складала 70% від загального обсягу вибірки. Кількість тематичних полів дорівнює кількості класів у вибірці. У кожне тематичне поле входили слова, які зустрічаються у два рази частіше у документах відповідного класу, ніж у загальній вибірці. Кількісні характеристики текстових документів для кожного тематичного поля розраховувались як суми текстових частот слів, які входять у кожне тематичне поле. Такі розрахунки було здійснено для тренувальної та тестової вибірок на основі множин слів тематичних полів, виявлених на тренувальній вибірці. Семантичні ознаки для латентного семантичного аналізу знайдено за допомогою сингулярного розкладу (Singular Value Decomposition, SVD) матриці TF-IDF. Для реалізації SVD розкладу використано алгоритм пакету scikit-learn, який описаний в [14]. Розрахунок LDA компонент здійснювався за допомогою пакету *gensim* [15]. Розглянемо деякі закономірності в отриманих семантичних ознаках текстових даних. Для аналітики багатомірних даних часто використовують t-SNE перетворення [16] щоб відобразити дані у низькорозмірному просторі. На рис.1 показано відображення документів різних класів (різними відтінками кольорами) у випадку двохвимірного t-SNE перетворення для кількісних ознак тематичних полів.

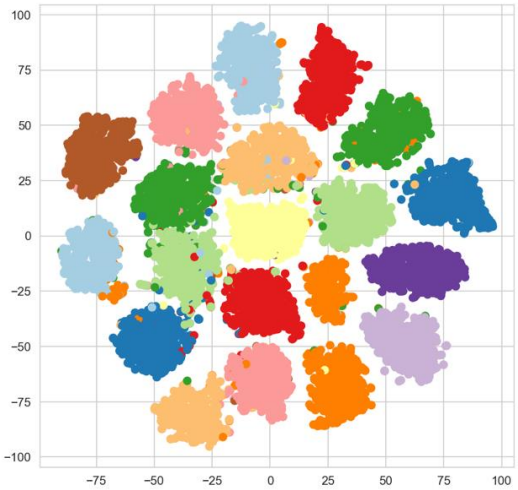

Рис. 1. Відображення документів різних класів у випадку двохвимірного t-SNE перетворення для кількісних ознак тематичних полів.

Як випливає із отриманих результатів, наведених на рис. 1, за допомогою тематичних полів можна чітко диференціювати класи документів. На рис. 2 показано приклад коробкового графіку (boxplot) розподілу по групах новин кількісної характеристики однієї із компонент LDA.

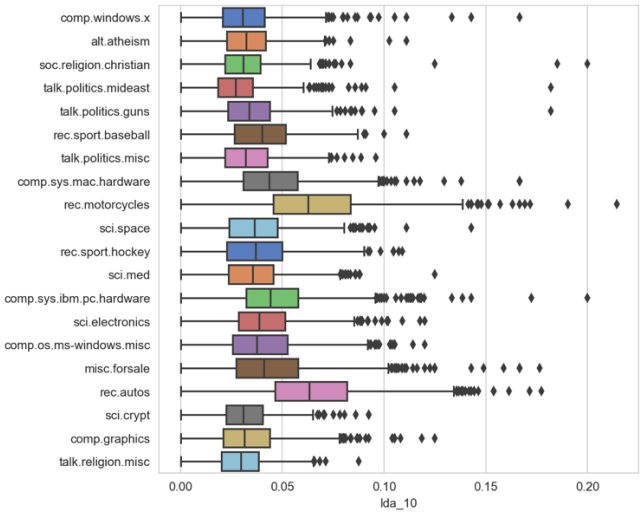

Рис. 2. Розподіл за класами кількісної характеристики однієї із компонент LDA.

Класифікаційний аналіз текстових даних

Розглянемо класифікаційний аналіз текстових масивів із використанням семантичних ознак текстових документів вибірки 20-Newsgroups [10]. Для проведення

класифікаційного аналізу, масив текстових документів було розділено на тренувальну (70%) та тестову вибірки (30%). Для класифікаційного аналізу було використано алгоритм Random Forest пакету *scikit-learn*. На рис. 3 показано розподіл значущості семантичних полів у класифікаційному алгоритмі Random Forest.

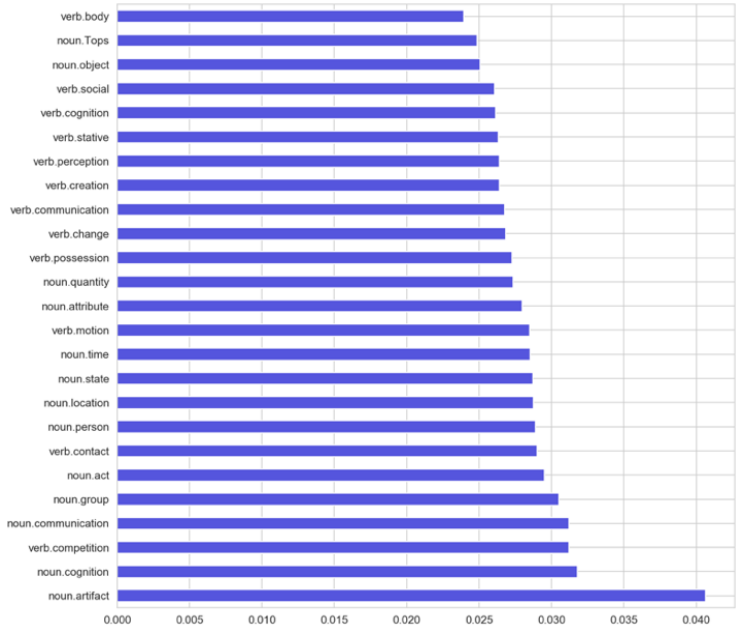

Рис. 3. Розподіл значущості семантичних полів у класифікації текстових документів.

Для оцінки класифікації використовувались точність (precision), повнота (recall) та f1-оцінка. На рис. 4 показано оцінки класифікації текстових документів за класами на основі кількісних характеристик семантичних полів з використанням алгоритму Random Forest. На рис. 5 показано оцінки класифікації текстових документів за класами на основі кількісних характеристик тематичних полів. Оцінки класифікації текстових документів за класами на основі кількісних характеристик компонент LDA показано на рис. 6.

Також розглянуто випадок, у якому взято сукупні семантичні ознаки, які складались із семантичних та тематичних полів, складових компонент SVD розкладу та компонент LDA розміщення. Класифікація здійснювалась на основі нейронної мережі із повністю з'єднаними шарами, структуру якої наведено на рис. 7. Нейронна мережа реалізована за допомогою пакету *keras*. Для зменшення ефекту перенавчання мережі, між шарами були введені *Dropout* шари, які випадковим чином обривають заданий відсоток зв'язків між шарами.

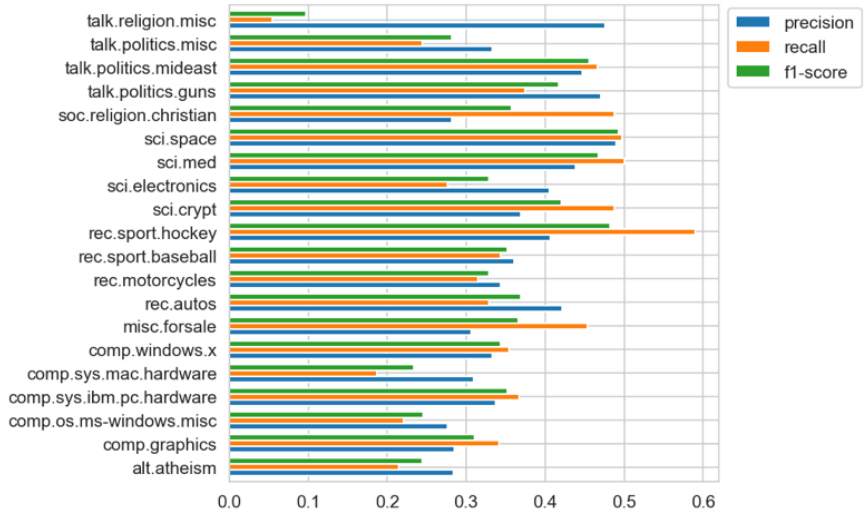

Рис. 4. Оцінки класифікації текстових документів за класами
на основі кількісних характеристик семантичних полів.

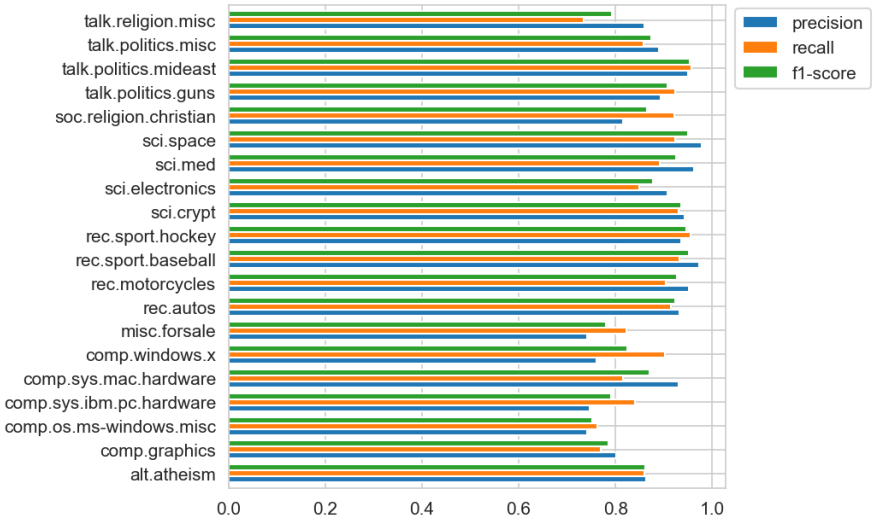

Рис. 5. Оцінки класифікації текстових документів за класами
на основі кількісних характеристик тематичних полів.

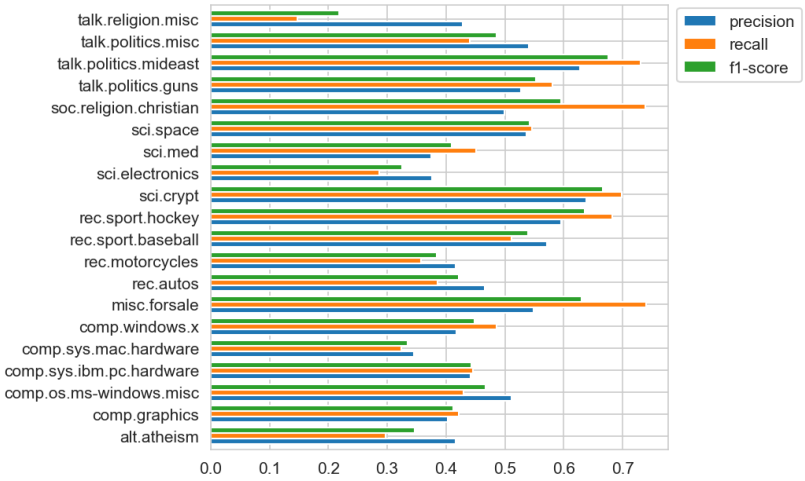

Рис. 6. Оцінки класифікації текстових документів за класами на основі кількісних характеристик компонент LDA.

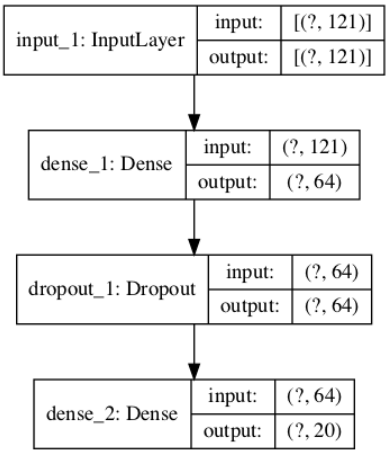

Рис. 7. Структура нейронної мережі із повністю з‘єднаними шарами.

На рис. 8 наведено динаміку кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання нейронної мережі. На рис. 9 наведено величини точності (precision), повноти (recall) та *f1*-оцінки для класифікації текстових документів нейронною мережею із повністю з‘єднаними шарами у випадку сукупності семантичних характеристик різних типів.

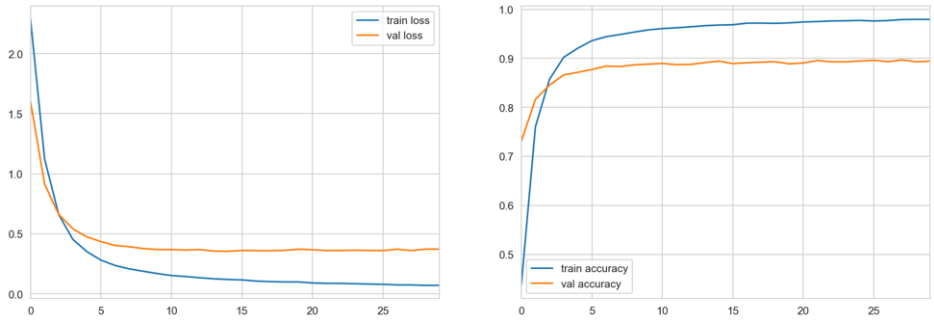

Рис.8. Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання нейронної мережі.

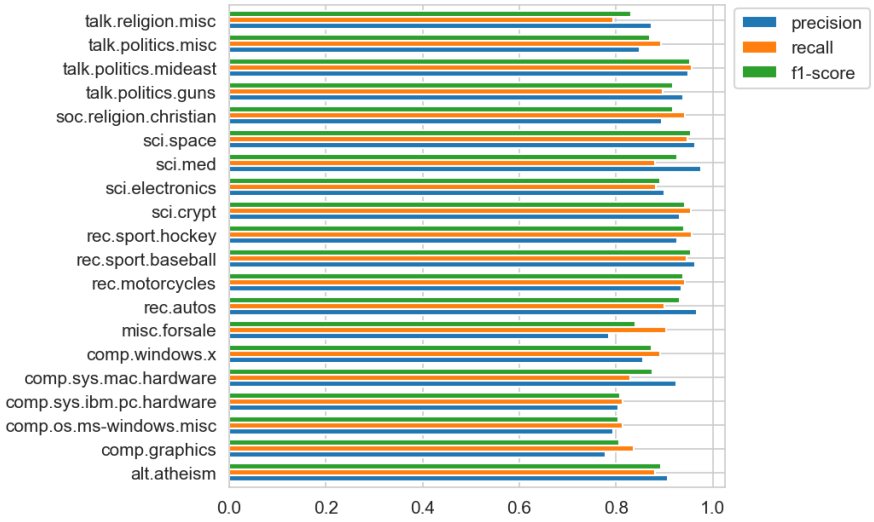

Рис. 9. Оцінки класифікації текстових документів нейронною мережею із повністю з'єднаними шарами.

Використання рекурентних нейронних мереж в аналітиці текстових даних

Рекурентні нейронні мережі із шарами з довгою короткочасною пам'яттю (Long short-term memory, LSTM) [17,18] часто використовують в аналітиці текстових даних. Розглянемо аналітику повідомлень соціальної мережі Твіттер за допомогою рекурентних нейронних мереж. Повідомлення соціальної мережі Твіттер можна розглядати як короткі текстові документи. Для аналізу твітів ми вибрали базу даних із повідомленнями Твітів, які стосуються сервісу декількох авіакомпаній [19]. Ця вибірка містить близько 14640 твітів. Твіти мають мітки 'negative', 'neutral', 'positive'. Розподіл класів за авіакомпаніями показано на рис. 10.

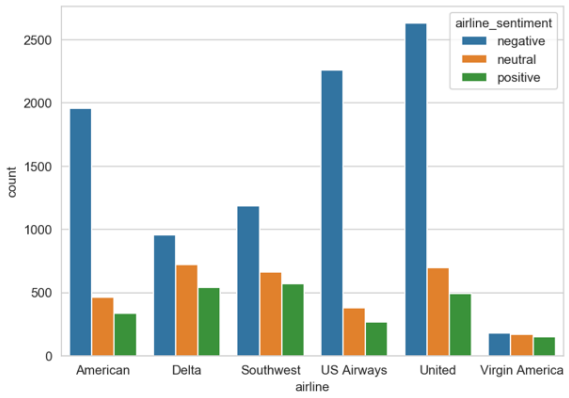

Рис. 10. Розподіл кількості твітів за класами цільової змінної для кожної авікомпанії.

Для класифікаційного аналізу було вибрано рекурентну нейронну мережу із двонаправленим шаром LSTM. Структуру такої мережі наведено на рис. 11. Для чисельних розрахунків були вибрані такі параметри: максимальна довжина текстових стрічок – 50 слів, максимальна кількість ознак вбудованого шару – 300, розмірність вбудованого шару – 5, параметр швидкості навчання – 0.0003, розмір пакету даних на ітерації навчання – 64.

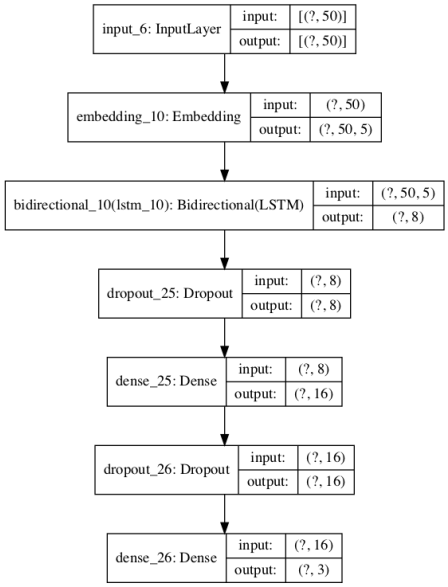

Рис. 11. Структура рекурентної нейронної мережі із двонаправленим LSTM шаром.

Набір твітів було розділено на тренувальну та валідаційну вибірки, розмір валідаційної вибірки 30 %. На рис. 12 наведено динаміку кількісних характеристик

втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентної нейронної мережі.

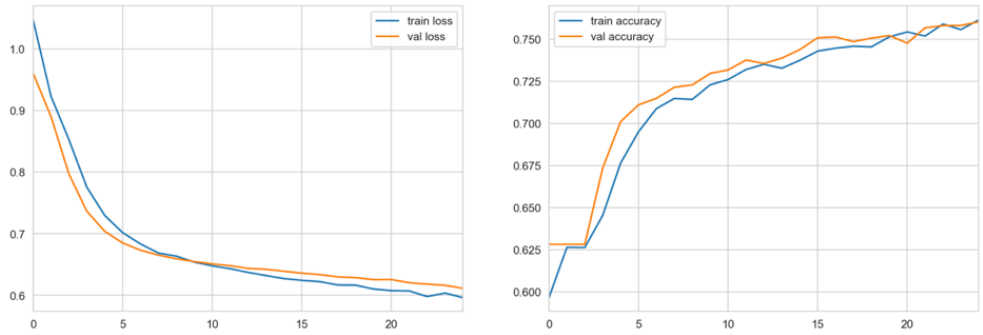

Рис.12. Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентної нейронної мережі.

На рис. 13 зображено кількісні величини точності, повноти та f1-оцінки класифікації твітів рекурентною мережею.

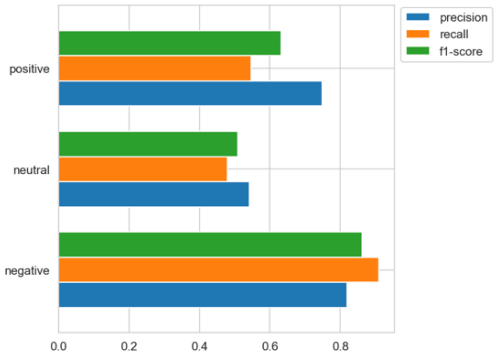

Рис.13. Оцінки класифікації твітів рекурентною мережею.

Розглянемо випадок класифікаційного аналізу за допомогою мережі, яка складається із двох підмереж. Одна підмережа містить двонаправлений LSTM шар, а інша складається із повністю з'єднаних шарів. На вхід першої мережі подаються текстові дані, а на вхід другої – числові дані, які характеризують текстові документи. Такими числовими даними можуть бути кількісні характеристики семантичних ознак. Для простоти розгляду вибрано один тип семантичних характеристик, який формується на основі сингулярного розкладу TF-IDF матриці. Для розрахунків взято 30 перших компонент такого розкладу. Структуру нейронної мережі із підмережами для текстових та числових даних зображено на рис. 14.

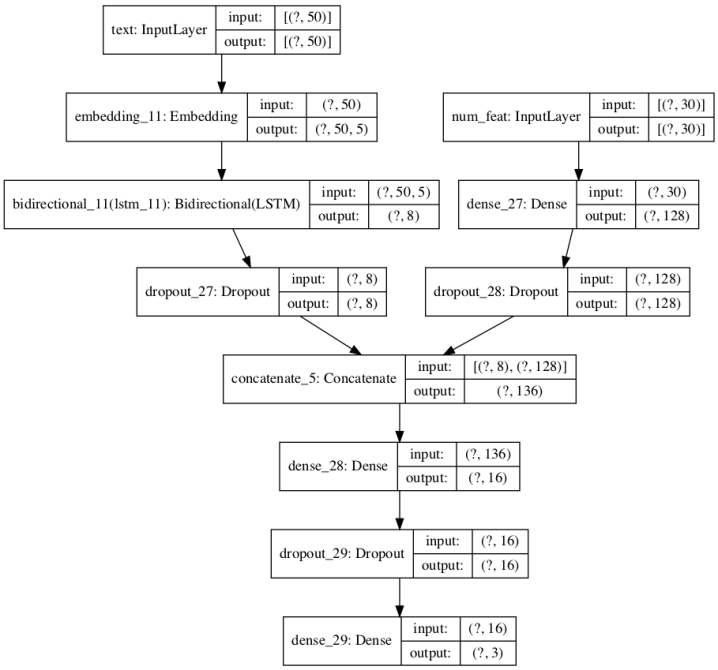

Рис.14. Структура нейронної мережі із підмережами для текстових та числових даних.

На рис. 15 наведено оцінки класифікації твітів такою комбінованою нейронною мережею, а на рис.16 - динаміку кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання такої нейронної мережі.

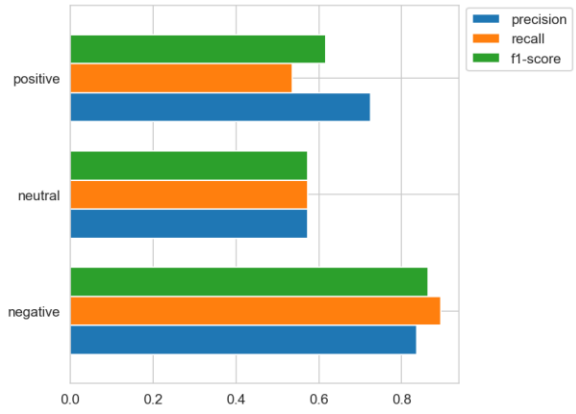

Рис. 15. Оцінки класифікації твітів комбінованою нейронною мережею.

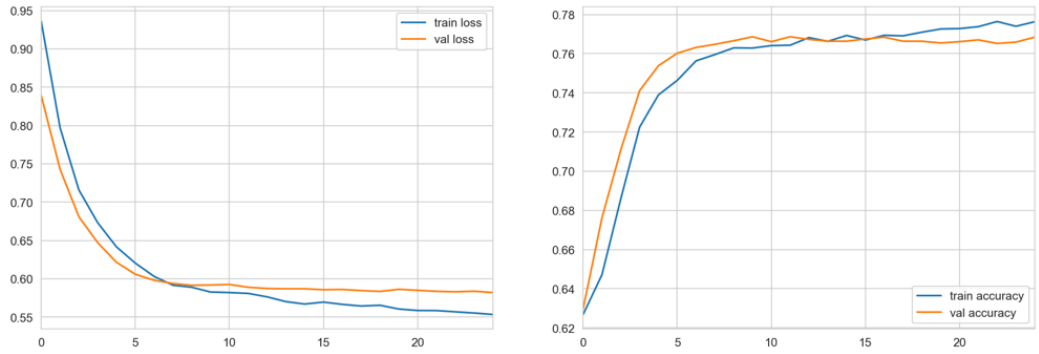

Рис.16. Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання комбінованої нейронної мережі.

Як випливає із отриманих результатів, використання нейронної мережі, яка складається із об'єднаних підмереж дає дещо кращі результати на валідаційному сеті, криві втрат та точності прогресують швидше, ніж у випадку рекурентної мережі, яка складається лише із шарів для обробки текстової інформації.

Розглянемо задачу числової регресії за наявності комбінованих даних текстового та числового типу. Для цього випадку було вибрано текстові дані опису товарів та їхні ціни. Ці дані були завантажені із платформи Kaggle [20]. Для простоти аналізу вибрано лише одну категорію даних, для аналізу було взято 15000 зразків даних. Крім текстового опису дані містять також категоріальні змінні, які було об'єднано як стрічкові дані із текстовим описом. Як числові характеристики вибрано 30 перших компонент SVD розкладу TF-IDF матриці. Структуру нейронної мережі, яка була використана у цьому аналізі зображено на рис.17. Для аналізу було вибрано такі основні параметри: максимальна довжина текстових стрічок – 250 слів, максимальна кількість ознак вбудованого шару – 3000, розмірність вбудованого шару – 10, параметр швидкості навчання – 0.0003, розмір пакету даних на ітерації навчання – 32. Як функція втрат розглядались характеристики RMSE, MAE. Цільову змінну, яка описує ціну товару розглянуто в алгоритмі навчання у логарифмічному масштабі на основі перетворення $y=\lg(x+1)$. Після отримання прогнозованих значень було проведено зворотнє перетворення масштабу даних за фукцією $y=10^x-1$. Оцінка похибки MAE, отримана на валідаційному наборі даних дорівнює 4.1, відносна оцінка MAE, дорівнює відношенню абсолютної оцінки MAE до середнього значення ціни, рівна 0.3. На рис. 18 наведено динаміку функцій RMSE та MAE на ітераціях навчання нейронної мережі для тренувального та тестового наборів даних. Як випливає із отриманих даних, алгоритм глибинного машинного навчання покращує точність прогнозування після реалізації деякої кількості ітерацій процесу навчання.

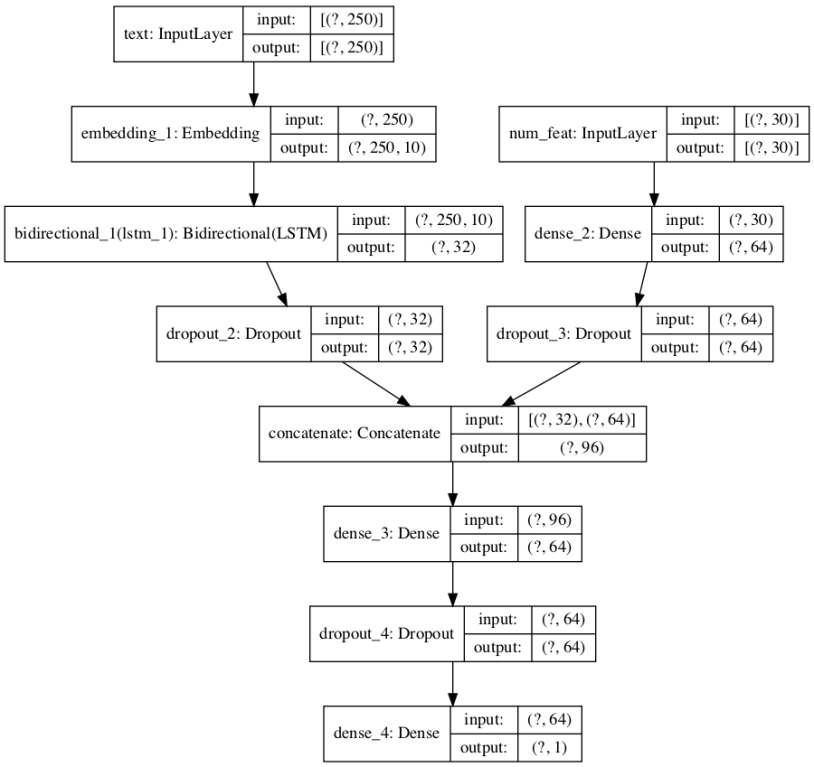

Рис.17. Структура нейронної мережі з підмережами для текстових та числових даних.

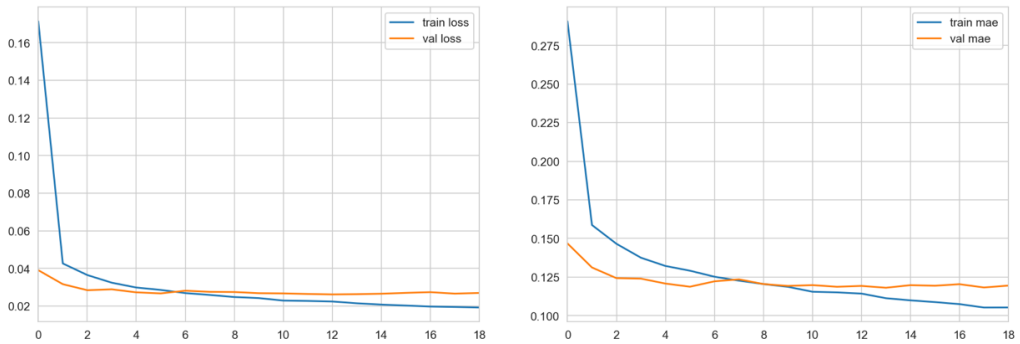

Рис.18. Динаміка кількісних характеристик втрат та точності для тренувальної та валідаційної вибірок на ітераціях навчання рекурентої нейронної мережі.

Висновки

У роботі досліджено використання семантичних ознак в інтелектуальному аналізі текстових даних, зокрема у класифікації текстових документів. Як семантичні ознаки розглянуто семантичні та тематичні поля, складові сингулярного розкладу матриці TF-IDF та складові латентного розміщення Дирихле. Класифікаційний аналіз здійснено за допомогою алгоритму Random Forest та алгоритмів глибинного навчання нейромереж із різною структурою з використання двонаправлених шарів із довгою короткостроковою пам'яттю (LSTM). LSTM шари нейронної мережі дають можливість враховувати порядок та комбінації лексем. Розглянуто випадок використання комбінованої нейронної мережі, яка складається із рекурентної нейронної підмережі для аналізу текстових даних та підмережі для числових семантичних ознак текстових документів. Також розглянуто числову регресію, у якій як вхідні розглядались текстові дані для випадку аналізу цін за текстовим описом товарів. Для аналізу була вибрано аналогічну комбіновану нейронну мережу із LSTM підмережею для текстових даних і підмережею із повністю з'єднаними шарами для числових компонент SVD розкладу TF-IDF матриці. Результати показують, що у текстових даних опису товарів можна знайти відповідні патерни і як наслідок, точність прогнозування ціни товару за текстовим описом покращується на ітераціях тренування такої нейронної мережі. Комбінації різних семантичних ознак дають можливість отримати вищу точність у задачах класифікації текстових документів. Використання широкого класу семантичних ознак у задачах інтелектуального аналізу диверсифікує аналітичні підходи і збільшує простір аналітичних ознак, що є важливим при невеликій кількості даних та при аналізі нестаціонарних процесів, коли прогнозний потенціал різних ознак може змінюватись із часом.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Pantel P., Turney P. D.* From Frequency to Meaning: Vector Space Models of Semantics // J. of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141–188.
2. *Fellbaum C.* WordNet. An Electronic Lexical Database // Cambridge, MA : MIT Press, 1998. – 432 p.
3. WordNet [Електронний ресурс].–Режим доступу: http://wordnet.princeton.edu
4. *Sebastiani F.* Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34. – N 1. – P. 1–47.
5. *Pavlyshenko B.* Classification Analysis of Authorship Fiction Texts in The Space of Semantic Fields // Journal of Quantitative Linguistics. – 2013. – Vol. 20, №3. – P. 218–226. http://dx.doi.org/10.1080/09296174.2013.799914
6. *Павлишенко Б. М.* Використання лексемних полів у інтелектуальному аналізі текстових масивів. Штучний інтелект. – 2013. – № 1. С. – 98-109
7. *Bohdan Pavlyshenko*. Clustering of authors' texts of english fiction in the vector space of semantic fields. Cybernetics and Information Technologies, 14(3):25–36, 2014
8. *Павлишенко Б. М.* Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів // Математичні машини і системи. – 2012. – №1. – С. 69–76.
9. *Gliozzo A.* Semantic Domains in Computational Linguistics / Alfio Gliozzo, Carlo Strapparava. – Springer, 2009. – 132 p.
10. 20 Newsgroups. [Електронний ресурс]. – Режим доступу: http://qwone.com/~jason/20Newsgroups/

11. *Deerwester S., Dumais S., Furnas G*. et al. Indexing by Latent Semantic Analysis // J. of the American Society for Information Science. – 1990. – Vol. 41. – Is. 6. – P. 391–407.
12. *Landauer T.K., Foltz P.W. and Laham D.* 1998. An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), pp.259-284.
13. *Blei D.M., Ng A.Y. and Jordan M.I.* 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.
14. *Halko N., Martinsson P.G. and Tropp J.A.*, 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, *53*(2), pp.217-288.
15. *Rehurek R. and Sojka P.*, 2010. Software framework for topic modelling with large corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.
16. *Maaten L.V.D. and Hinton G.* 2008. Visualizing data using t-SNE. Journal of machine learning research, 9 (Nov), pp.2579-2605.
17. *Gers F. A., Schraudolph N. N., Schmidhuber J.* "Learning Precise Timing with LSTM Recurrent Networks", Journal of Machine Learning Research 3, 2002, pp. 115–143
18. *Sundermeyer M., Schlüter R. and Ney H.* LSTM neural networks for language modeling. In Thirteenth annual conference of the international speech communication association, 2012.
19. Airline Twitter sentiment. in Data For Everyone WordNet. [Електронний ресурс].– Режим доступу: https://www.figure-eight.com/data-for-everyone/
20. Mercari Price Suggestion Challenge, Kaggle.com. [Електронний ресурс].–Режим доступу: https://www.kaggle.com/c/mercari-price-suggestion-challenge

USING MACHINE LEARNING AND SEMANTIC FEATURES IN INTELLECTUAL ANALYSIS OF TEXT DATA

Bohdan M. Pavlyshenko

Ivan Franko National University of Lviv,
50 Drahomanov St., 79005, Lviv, Ukraine
b.pavlyshenko@gmail.com

In the paper, the study of using supervised machine learning and the semantic features was carried out. As semantic features, semantic and thematic fields, the components of singular value decomposition and the components of latent Dirichlet allocation were considered. As semantic fields, groups of lexemes united by some specified concept were considered. Groups of lexemes with text frequencies twice as high in specified classes of text documents as in a mutual set of text documents were considered as thematic fields. Using thematic fields features, text documents classes can be differentiated accurately. The classification analysis was conducted using Random Forest algorithm and deep learning algorithms for neural networks. Neural networks with fully connected layers and semantic quantitative features were analyzed. Neural networks with embedded layers for text representation and with bidirectional LSTM layer were considered. LSTM layer makes it possible to take into account the order and combinations of words. The approach with the use of neural networks which consists of the recurrent neural subnetwork for text data processing and the subnetwork for numerical semantic features is considered. Precision, recall and f1-scores were used for classification scoring. The cases with the combination of

semantic and thematic field features, singular value decomposition components for TF-IDF matrix and latent Derichlet allocation components were considered. The numeric regression using text data as input features for the case of product analytics using product text description was considered. For this regression analysis, combined neural network with LSTM layer for text analytics and fully connected layer for numerical text semantic features were considered. The results show that the patterns in the product text descriptions can be found by this kind of neural network, the accuracy for price prediction improves on the training iteration of such combined neural network. The use of the wide class of semantic features in intellectual analysis of text data diversifies analytic approaches and increases semantic feature space in analytical problems when the prediction potential of the features can change with time.

Key words: text analytics, text semantic features, text classification, neural networks.

Стаття: надійшла до редакції 20.02.2020,
доопрацьована 27.04.2020,
прийнята до друку 29.04.2020