УДК 519.21 https://doi.org/10.17721/1812-5409.2020/1-2.4

В.П. Зубченко[1], *к.ф.-м.н.*
Є.О. Костюк[2],
М.О. Лукащук[3],
А.М. Ярошевський[4]

Класифікація зміни рейтингу фінансових показників страхових компаній

[1–4]Київський національний університет імені Тараса Шевченка, 01033, Київ, вул. Володимирська, 64.
[3]Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico City, Av. Juan de Dios Batiz S/N, Nueva Industrial Vallejo, Gustavo A. Madero, 07738 Ciudad de Mexico, Mexico
e-mail: [1]v_zubchenko@ukr.net,
[2]kosteugeneo@gmail.com,
[3]lukashuk.nikola@gmail.com,
[4]yaro.andriy@gmail.com

V.P. Zubchenko[1], *Ph.D.*
Ye.O.Kostiuk[2],
M.O. Lukashchuk[3],
A.M. Yaroshevskyi[4]

Rating change classification of insurance companies indicators

[1–4]Taras Shevchenko National University of Kyiv, 01033, Kyiv, 64 Volodymyrska st.
[3]Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico City, Av. Juan de Dios Batiz S/N, Nueva Industrial Vallejo, Gustavo A. Madero, 07738 Ciudad de Mexico, Mexico
e-mail: [1]v_zubchenko@ukr.net,
[2]kosteugeneo@gmail.com,
[3]lukashuk.nikola@gmail.com,
[4]yaro.andriy@gmail.com

У роботі досліджуються залежності між фінансовими показниками страхових компаній та простором новин, який розглядається у вигляді сукупності тематик. Метою роботи є побудова моделі для передбачення напрямку зміни рейтингу страхової компанії за заданим показниками (вгору або вниз від поточної позиції в рейтингу) на основі множини новинних статей за відповідний проміжок часу. Було створено підхід, який включає пошук найбільш впливових тем для заданого показника. Використано метод тематичного моделювання Latent Dirichlet Allocation (LDA) та наївний баєсів класифікатор. Для оцінки було використано Leave-One-Out валідацію часових рядів із метрикою точності.

Ключові слова: LDA, статистичний аналіз новин, тематичне моделювання, наївний баєсів класифікатор, фінансові показники страхової компанії.

In this paper we investigate the relationship between financial indicators of insurance companies and news space. The news space is considered as a set of topics. The goal of the paper is to fit the model in order to forecast company's rating change for given indicators – whether rating will go up or down regarding the current value. As the data set we use news articles of the relevant insurance topics for the specified time period. The approach we use includes search for the most influential topics for the given indicator. To retrieve topics, we used Latent Dirichlet Allocation (LDA) algorithm and Naive Bayes model. For the validation the Leave-One-Out approach was used with accuracy metric.

Key Words: LDA, news space analysis, topic modelling, Naive Bayes, financial indicators of insurance company.

1 Вступ

У роботі досліджується питання прогнозування зміни рейтингів фінансових показників страхових компаній. Використано відкриті дані страхового ринку України та, для прикладу, показник сумарних активів страхових компаній. Відкриті дані щодо історичних показників фінансового рейтингу страхових компаній України доступні за посиланням https://forinsurer.com/ratings/nonlife. Дана задача є важливою, оскільки в багатьох галузях, зокрема в страхуванні, недостатньо даних для побудови прогнозного механізму в режимі реального часу. Обмеження виникають з різних причин, таких як законодавче регулювання, занадто велика ціна підготовки фінансових звітів, недостатня автоматизація й цифровізація

бізнес-процесів компаній тощо. Наприклад, різниця між наявними даними щодо фінансової динаміки страхових компаній та ринком акцій полягає в тому, що інформація щодо ціни акції компанії на біржі доступна майже у довільний час. Натомість щодо страхового ринку оперативної інформації дуже мало. Єдине, що відомо – це динаміка ключових фінансових показників страхових компаній на основі щоквартальних фінансових звітів. Також, страховики зобов'язані на щорічній основі публікувати баланс, звіт про прибутки та збитки, та дані щодо ключових показників страхової діяльності.

2 Попередні дослідження

Дослідження ринку показало, що вплив засобів масової інформації у фінансовому секторі за останні роки значно збільшився, особливо у сфері фондового ринку. Так, опубліковано велику кількість результатів щодо залежності між інформацією, поширеною в ЗМІ, та ціною акцій, волатильністю та трендом їх динаміки [7, 8, 9, 10]. Наприклад, у [11] було показано, що медіапростір пов'язаний із фондовими ринками. Такі зв'зки дуже складні і включають вплив на поведінку інвесторів.

У роботі [12] використовується метод LDA з класифікатором, який забезпечує прогнозування волатильності для часових діапазонів з 20-хвилинним зміщенням. Взявши за основу даний підхід, ми адаптуємо його для побудови моделі прогнозування руху рейтингів фінансових показників страхових компаній відносно конкурентів (на відміну від прогнозування абсолютних значень). У роботі [13] автор проводить аналіз впливу на фінансову динаміку акцій інформації із соціальних мереж та ЗМІ. У задачі прогнозування волатильності автору вдалося досягнути точності 83,2%.

Результатів щодо аналізу динаміки страхових компаній за допомогою методі машинного навчання та data-science значно менше, оскільки звітні дані доступні здебільшого лише на щоквартальні основі. Залежно від законодавчого регулювання конкретної країни звітні дані можуть подаватись із різною частотою, втім у відкритому доступі найчастіше доступна саме щоквратальна динаміка, на відміну від значно вищої частоти даних із фондових ринків, де ціни можна отримати практично в будь-який момент часу. Отже, страховий ринок має принципово іншу структуру даних, ніж фондовий ринок, оскільки ми не можемо щодня здійснювати моніторинг інформації щодо динаміки зміни фінансових показників страхових компаній. Також одним із факторів виступає доступ до інформаційних медіа ресурсів. Коли новини про фондові ринки досить популярні у суспільстві та мають величезну аудиторію, то зі страхуванням дещо важче. В Україні, де страхування досі є темою більш вузькоспеціалізованою, окремих медіа ресурсів, які спрямовані саме на цю тему не так багато. Зазвичай новини розріджені по всіх можливих журналах, сайтах тощо, що ускладнює задачу збору та аналіху даних.

3 Використані алгоритми

3.1 Latent Dirichlet allocation (LDA)

У роботі Blei et al. [1] була описана генеративна статистична модель, яка кожному документу ставить у відповідність ймовірності наявності у ньому наперед заданої кількості тематик. LDA спирається на деякі припущення щодо структури документів, такі як:

- порядок слів не має значення;
- кількість тем заздалегідь відома (або оцінюється заздалегідь);
- розподіл тематик моделюється розподілом Діріхле.

У LDA кожен документ складається з тематик, які розглядаються як латентні (приховані) змінні. Кожна тема розглядається як набір слів із ймовірностями кожного слова потрапити у тему. Відповідно, після процедури підгонки є можливим оцінити за цими ймовірностями розподіл тем на документ. Ці ймовірності і є "основними" параметрами моделі.

Більш детально, LDA – це ієрархічна баєсівська модель, що складається з двох рівнів: перший рівень складається з компонентів, які відповідають "тематикам" (або його ще називають латентним рівнем); другий рівень – мультиноміальна змінна з апріорним розподілом Діріхле, яка відповідає за класифікацію "тематик" у документі. Потім для кожного слова будується його розподіл належності до кожної з тем.

Однією з переваг підходу є робота на нерозмічених текстах. Досить часто розмітка є проблемою за рахунок вартості та часу. Як вже зазначалося, в умовах страхового медіа простору така розмітка та збір даних інколи є досить суттєвим обмеженням.

3.2 Тестування

Leave-One-Out валідація [4] – це метод оцінки моделі. Для проведення оцінки наявні дані діляться на n частини, де n – загальна кількість зразків у наборі даних. Потім $n-1$ частина використовуються в процесі навчання, а для решти частини ми проводимо тести. Процес повторюється n разів. Таким чином для тестування використовується весь наявний набір даних, що покращує якість перевірки моделі. З іншого боку, такий підхід вимагає більшого часу роботи та ресурсів для навчання n моделей у порівнянні з класичним розбиттям на тренувальну та тестувальну вибірки. Однак з іншого боку, незважаючи на те, що такий підхід дає більш точне уявлення про якість моделі, він витрачає значно більше часу на впровадження - замість однієї моделі треба навчити n. а опісля - перевчити на всіх даних.

3.3 Коефіцієнт узгодженості

Коефіцієнт узгодженості [5] – коефіцієнт, який обчислюється для кожної теми. Його можна трактувати як оцінку смислової схожості чи спорідненості слів у зазначеній темі. Коефіцієнт узгодженості допомагає зрозуміти, як узгоджуються між собою слова, які визначають тематику. Існує багато різних підходів до обчислення цієї оцінки. У нашому випадку він обчисляється наступним чином. Покладемо $t_i = \{w_k^i, k = 1, ..., K\}$ – тема, яка представлена за допомогою K найбільш імовірних слів. Нехай сумарна кількість тематик рівна T. Для векторизації слів використовувалися попередньо обчислені вектори fastText [3] для української мови[1]. Вибір векторизації досить суттєвий. Було обрано саме цей згідно наступних причин:

- fastText досить непогано уміє працювати з одруковками, яких у наших текстах очікується немало;
- модель був анавчена на великих об'ємах тексту, що дозволило їй отримати хороші властивості у отриманого векторного простору;

Коефіцієнт обчислюється згідно наступної формули:

$$\frac{1}{T}\frac{1}{K}\sum_{k=1}^{T}\sum_{\substack{i\neq j\\ i,j=1}} cos(w_k^i, w_k^j)$$

Чим більший показник узгодженості – тим краще, оскільки ми очікуємо, що слова, які визначають тематику, мають багато спільного.

3.4 Наївний баєсів классифікатор

Ймовірнісна модель наївного баєсового касифікатору [2] – це умовна модель $p(C \mid F_1, \ldots, F_n)$ над залежною змінною C, яка залежить від декількох змінних $F_1, \ldots, F_n$. Умовний розподіл змінної C за $F_1, \ldots, F_n$ можна зобразити у вигляді: $p(C \mid F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C)$, де Z – нормуючий множник, який залежить лише від $F_1, \ldots, F_n$. Значення прогнозу C визначається згідно $\text{classify}(f_1, \ldots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i \mid C = c)$

4 Вибірка

Набір даних був отриманий з різних українських новинних веб-сайтів, які містять теги для статті, наприклад, "страхування" та мітку конкретно обраної компанії. Ми розглядали часовий діапазон з початку 2014 року до кінця 2019 року, розділений на квартали (загалом 24 різних періоди часу). Що стосується рейтингу показників для конкретної компанії – дані з відкритих джерел збиралися за кожен квартал. Рейтинги показників були отримані як позиції в рейтингу для досліджуваної страхової компанії відносно динаміки решти компаній у поточному кварталі. Ми зібрали 12829 новинних статей за вказаний часовий проміжок та дослідили рейтинги 22 фінансових показників.

4.1 Обробка ознак та відгуків

Як ознаки, простір новин розділили на відповідну кількість тем (більш докладний опис наведено в розділі "Експерименти та результати". Для кожної теми $topic_t$ обчислюється сума її

[1] https://fasttext.cc/docs/en/crawl-vectors.html

ймовірностей за визначений проміжок часу t. Після цього всі теми ранжуються за цим значенням із зміщенням на один часовий проміжок назад. Позначимо отриманий результат як $topic_t^r$. Після цього до кожної теми застосовується функція:

$$\begin{cases} 0 & \text{topic}_t^r \leqslant \text{topic}_{t-1}^r \\ 1 & \text{topic}_t^r > \text{topic}_{t-1}^r \end{cases}$$

Таким чином, простір новин перетворюємо на бінарний вхід, який показує зміни популярності тем з часом. Аналогічна функція була застосована до відгуку.

5 Експерименти та результати

Для того, щоб встановити характер залежності між поведінкою коефіцієнтів та поточним станом простору новин, на отриманих даних була побудована модель LDA. Кількість тематик була підібрана за допомогою коефіцієнту узгодженості та рівна 60. Як вхідні ознаки, були використані попередньо оброблені бінарні ознаки та відгуки. На цих даних була підлагоджена модель наївного баєсового класифікатора, яка досягла точності 0.77 тобто 17/22, із довірчим інтервалом рівня 0.95: $[0.5463, 0.9218]$. Для кожної тематики було проведено тест Фішера на перевірку значимості залежностей для кожної з тематик. Нижче наводимо основні результати, що свідчить про існування залежності між певними новинними тематиками та обраним фінансовим показником сумарних активів страхової компанії. Опис тематик з найбільш ймовірних 5 слів для кожної новинної тематки виглядає наступним чином:

- Тематика 18, p-value: 0,039, Кореляція: 0.441;
- Тематика 19, p-value: 0,039, Кореляція: 0.442;
- Тематика 26, p-value: 0,009, Кореляція: 0.541.
- Тематика 18; Слова: грудень, субота, понеділка, перенести, карта;
- Тематика 19; Слова: медичний, державний, суб'єкт, ступінь, ризик;
- Тематика 26; Слова: ліцензія, страхування, страховий, фінансовий, здійснювати.

Можна помітити, що слова, що визначають тематики вдалося отримати досить непогано - загалом, вони мають зміст. Можна виділити тему медичного страхування, яка була досить популярною в українському суспільстві. Також виділється тема деяких часових новин - має місце сезонність новин – наприклад, зміна пор року та відповідні проблеми, що виникають під час таких процесів. Згідно отриманих результатів статистичних тестів можна стверджувати існування статистичної залежності між відгуком та ознаками.

6 Висновок

У роботі показано існування статистично значущої залежності між змінами простору новин у засобах масової інформації та ключовими показниками фінансової динаміки страхової компанії. Результат показує, що за допомогою використаних в роботі методів машинного навчання можна краще зрозуміти поточний фінансовий стан страхової компанії та спрогнозувати тренд її фінансової динаміки на конкурентному ринку. Використання більшої кількості даних та часових фрагментів може допомогти краще зрозуміти вплив ЗМІ на компанії, про які ми маємо обмежену кількість інформації. Дослідження можуть бути корисними для оцінювання ефективності потенційних інвестицій та трендів розвитку ринку.

Далі від дослідження локального ринку страхових послуг України ми плануємо перейти до дослідження ринків страхових послуг інших країн, зокрема Великобританії, США, Гонконгу, Білорусі та інші. Прогнозуємо, що таке дослідження може виявити схожість динаміки страхового ринку в країнах із однаковими політичними режимами, структурою економіки тощо. Подальший аналіз може показати й принципові відмінності між динамікою страхового ринку різних країн. Загалом використані методи покликані забезпечити можливість прогнозування ключових фінансових показників на основі інформації із ЗМІ та соціальних мереж навіть в умовах відсутності відкритого доступу до фінансової звітності страхових компаній. З іншого боку, такий аналіз дає краще розуміння залежностей новинного простору та його впливом на цілком реальні фінансові показники різних компаній. У епоху такого різнома-

ніття думок та суспільних ресурсів для компаній досить важливо мати надійний інструмент для підрахунку у відносних чи абсолютних значеннях змін у суспільній думці щодо компанії. Більш того, часто необхідно зважувати та порівнювати ступінь впливу того чи іншого видання на кінцевий результат компанії.

Список використаних джерел

1. Blei, David M. and Ng, Andrew Y. and Jordan, Michael I. (2016), "Latent Dirichlet Allocation", *JMLR.org*, v. 3, pp. 993–1022.
2. Rish, Irina (2001), An Empirical Study of the Naive Bayes Classifier", *Empir. methods Artif. Intell. Work. IJCAI 2001*, v. 22230, pp. 41-46.
3. Bojanowski, Piotr and Grave, Edouard and Joulin, Armand and Mikolov, Tomas (2017), Enriching Word Vectors with Subword Information", *Transactions of the Association for Computational Linguistics*, v. 5, pp. 135-146.
4. Stone, M. (1974), Cross-Validatory Choice and Assessment of Statistical Predictions", *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 36 , pp. 111-147.
5. Aletras, Nikolaos and Stevenson, Mark (2013), Evaluating Topic Coherence Using Distributional Semantics", *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pp. 13-22.
6. Gagniuc, Paul A. (2017), *From Theory to Implementation and Experimentation*, John Wiley & Sons, pp. 1–235.
7. Shunrong Shen and Haomiao Jiang and Tongda Zhang (2012), *Stock Market Forecasting Using Machine Learning Algorithms*, Department of Electrical Engineering, Stanford University, Stanford, CA.
8. Hegazy, Osman and Soliman, Omar S. and Abdul Salam, Mustafa (2013), A Machine Learning Model for Stock Market Prediction", *International Journal of Computer Science and Telecommunications*, v. 4, pp. 17-23.
9. Coupelon, Olivier (2007), Neural network modeling for stock movement prediction: A state of the art", *https://cutt.ly/Js9lhiM*
10. Tetlock, Paul (2007), Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *Journal of Finance*, v. 62, pp. 1139-1168.
11. Engelberg, Joseph and Parsons, Christopher (2011), The Causal Impact of Media in Financial Markets", *The Journal of Finance*, v. 66, pp. 67-97.
12. Atkins, Adam and Niranjan, Mahesan and Gerding, Enrico (2018), Financial News Predicts Stock Market Volatility Better Than Close Price", *The Journal of Finance and Data Science*, v. 4.
13. Khan, Wasiat and Ghazanfar, Mustansar ali and Azam, Muhammad Awais and Karami, Amin and Alyoubi, Khaled and Alfakeeh, Ahmed (2020), Stock market prediction using machine learning classifiers and social media, news", *Journal of Ambient Intelligence and Humanized Computing*.
14. Floreddu, P. and Cabiddu, Francesca (2014), Managing Online Reputation: The Role of Social Media in Insurance Industry", *Academy of Management Proceedings*, v. 1.
15. Cerqueira, Vitor and Torgo, Luis and Mozetic, Igor (2019), *Evaluating time series forecasting models: An empirical study on performance estimation methods*.

Received: 17.01.2020