

ОЦІНКА КРЕДИТНИХ РИЗИКІВ ЗА ДОПОМОГОЮ IBM WATSON API

Пономаренко Т. В., к.т.н., доцент¹, Мосьпан А. О., магістрант²
^{1,2} Національний університет кораблебудування імені адмірала Макарова
^{1,2} Україна, Миколаїв
¹ ORCID 0000-0003-1631-7574, ² ORCID 0000-0002-5122-7374

Анотація. У статті наведено метод, що базується на машинному навчанні задля багатофакторного аналізу особистості, для визначення її прихованих рис з повсякденної мови користувачів у соціальних мережах. Індивідуальні відмінності часто ґрунтуються на питаннях анкети, які можуть не охопити багатство повсякденної поведінки людини, тому не відображають повною мірою всі фактори що можуть впливати на кредитні ризики.

Ключові слова: багатофакторний аналіз кредитних ризиків, аналіз прихованих рис особистості, персональний аналіз постів у соціальних мережах

Конструкція ознак, заснована на мові, прогнозує результати, що не ґрунтуються на анкетах, але краще, ніж ознаки, що ґрунтуються на відображають реальний стан інтелекту, доходу, відкритості. Цей підхід передбачає цінність у нових конструкціях особистості, похідних від повсякденного використання людської мови (рисунок 1). Згідно з веб-сайтом IBM, інструмент "використовує лінгвістичну аналітику для підрахунку внутрішніх характеристик особистості, включаючи , потреби та цінності, з цифрових комунікацій, таких як електронна пошта, текстові повідомлення, твіти та повідомлення на форумах" [1]. Окрім того, "особисті дані" показують ваші споживчі звички та "часову поведінку" (якщо вхідний текст позначається часом) .

Вхідні дані: після того як користувач дає згоду та вводить особисті аккаунти соціальних мережах, IBM має ваші твіти, електронні листи, текстові повідомлення, повідомлення в блогах та / або все, що написано персоною, особистість якої оцінюється. На даний момент інструмент підтримує англійську, іспанську, японську, корейську та арабську мови, хоча за даними веб-сайту, результати для арабської та корейської мови недостатньо хороші, щоб бути остаточними. Мінімальна вибірка складає 100 слів, щоб отримати результат, однак для найкращої точності потрібен аналіз близько 3000 слів введеного тексту.

Вихідні дані: Після обробки вхідних даних створене програмне забезпечення повертає повний результат (у форматі JSON або CSV), показуючи ваші 52 особистісні характеристики в числових показниках на додаток до вашої споживчої поведінки. Оцінка виражається у відсотках від вибіркової сукупності. Наприклад, якщо мій "пригодний" характерний показник становить 0,25, це означає, що, виходячи з моєї роботи, я більш пригодний, ніж 25% вибіркової сукупності і менш авантюристичний, ніж 75%.

Примітка: еталонна вибірка складається з користувачів Facebook, Twitter, Instagram, Telegram, інформація яких була зібрана з їх згоди та проаналізована за допомогою сервісу IBM Watsons Personal Insights та нашого програмного забезпечення. Результати аналізу одного з них представлено на рисунку 2.

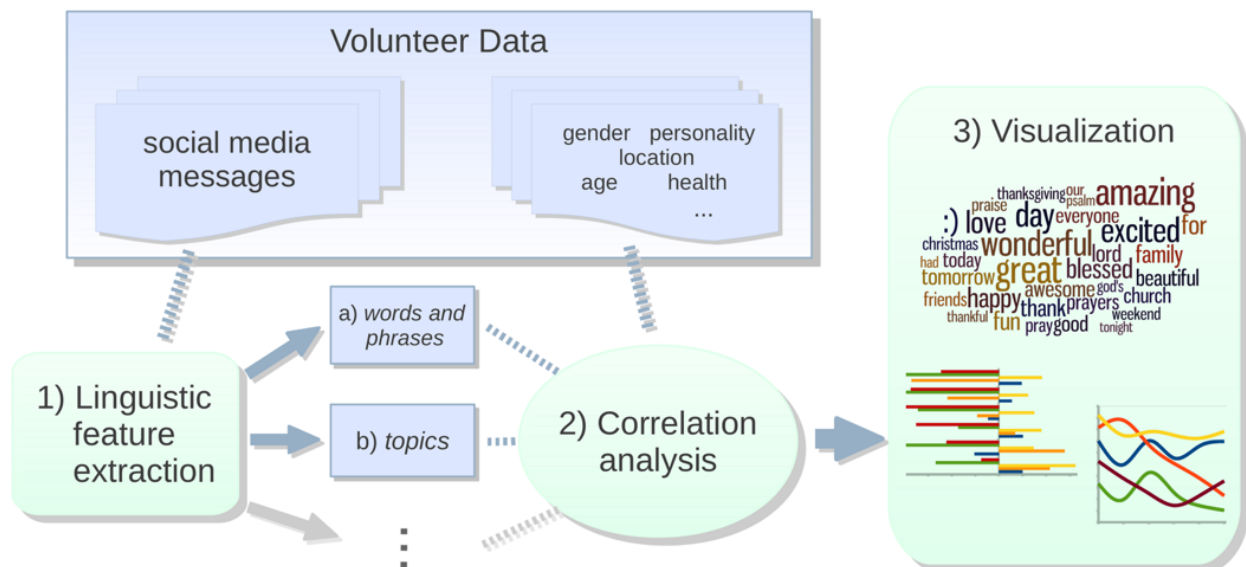


Рисунок 1 – Алгоритм багатofакторного аналізу з використанням соціальних мереж

Модель: Основний метод заснований на підході до відкритого словника. Цей метод був розроблений дослідниками Пенсільванського університету, які проаналізували статуси у Facebook 75 000 добровольців. На основі цього аналізу та супровідних особистих опитувальників вони побудували моделі для прогнозування віку, статі та особистості людини [1].

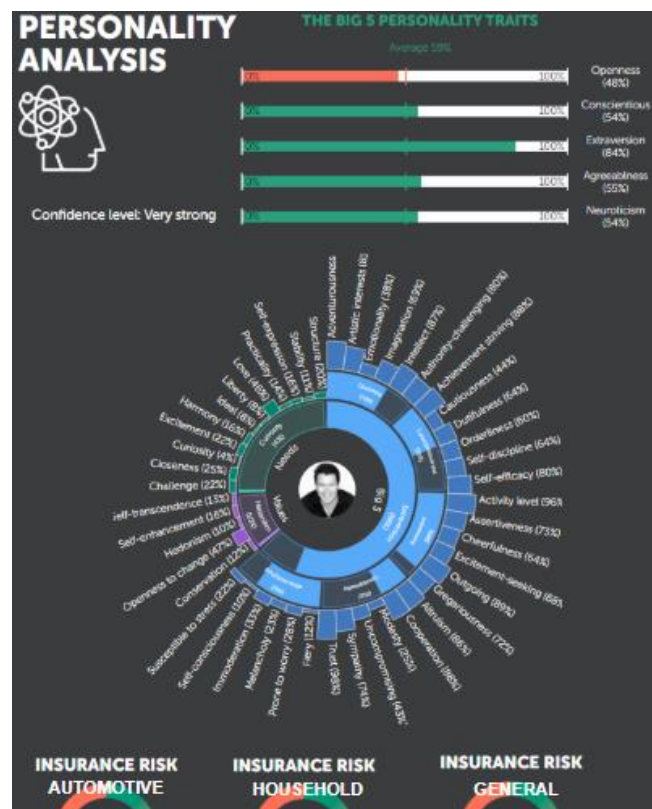


Рисунок 2. Результати аналізу кредитних ризиків одного з волонтерів

Запропоновано метод, що базується на багатофакторному аналізі, для визначення прихованих рис особистості з повсякденної мови користувачів у соціальних мережах. Людина - це "організована, динамічна, агентська система, що функціонує в соціальному світі", з характеристиками, поведінкою, почуттями та думками, які є одночасно послідовними та змінними у часі та ситуаціях [3]. Індивідуальні відмінності часто базуються на питаннях анкети, які можуть не зафіксувати багатства повсякденної поведінки людини.

Висновки. Представлений метод визначає приховані фактори з мовної поведінки в соціальних медіа - середовищі, що дозволяє отримати доступ до великих розмірів вибірки; непередбачуваний доступ до думок, емоцій та мови користувача; та керовані даними підходи. Отримано п'ять додаткових факторів з мови та продемонстровано, що ці фактори є узагальнюючими з високим ступенем прогнозування та є стабільними у часі. Ми розглядаємо це як достовірний інструмент доповнення картини кредитних ризиків особистості, що може широко використовуватися кредитними установами, виходячи з того що він заснований на масштабних поведінкових даних, а не на власних звітах опитувальника.

REFERENCES

- [1] Allport GW, Odbert HS. Trait-names: A psycho-lexical study. Psychological monographs. 1936;47(1):i
- [2] Golbeck J, Robles C, Edmondson M, Turner K. Predicting personality from twitter. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE; 2011. p. 149–156.
- [3] Liu F, Perez J, Nowson S. A Language-independent and Compositional Model for Personality Trait Recognition from Short Texts. arXiv preprint arXiv:161004345. 2016.

Ponomarenko Tetyana Viktorivna, Mospan Adam Olexandrovich

CREDIT RISK ASSESSMENT USING API IBM WATSON

Abstract. The article presents a multivariate analysis method for using hidden personality traits from everyday users' language on social networks. Individual differences are often based on survey questions that may not capture the wealth of everyday human behavior, so they do not fully reflect all the factors that may affect credit risk.

Keywords: multivariate analysis of the credit risks, analysis of the hidden personality traits, personal analysis of the posts on social networks.