УДК 004.021

М. О. Гранік
В. І. Месюра

РОЗРОБКА ТЕОРЕТИКО-МЕТОДИЧНИХ ЗАСАД ОЦІНЮВАННЯ ПРАВДИВОСТІ НОВИННОЇ ІНФОРМАЦІЇ. ВИЗНАЧЕННЯ НЕПРАВДИВИХ НОВИН ЗА ДОПОМОГОЮ НАЇВНОГО БАЄСІВОГО КЛАСИФІКАТОРА

Вінницький національний технічний університет;

Анотація
У цій роботі показується простий спосіб застосування наївного баєсівого класифікатора для визначення неправдивих новин. Результати роботи підтримують ідею про визначення неправдивих новин за допомогою методів штучного інтелеку
Ключові слова: неправдиві новини, наївний баєсів класифікатор, штучний інтелект

Abstract
This paper shows a simple approach for fake news detection using naive Bayes classifier. Received results suggest, that fake news detection problem can be addressed with artificial intelligence methods.
Keywords: fake news; naive Bayes classifier; artificial intelligence

Вступ

Інтернет та соціальні медіа зробили доступ до новин легшим ніж будь-коли. Часто користувачі мережі Інтернет можуть слідкувати за новинами онлайн, і розповсюдження мобільних пристроїв тільки спростило цей процес.

Але разом з великими можливостями приходять і великі виклики. Засоби масової інформації мають великий вплив на суспільство, і, як це часто буває, знаходяться люди, які хочуть цим скористатись. Це призводить до виникнення новин, які повністю або частково не є правдивими. На сьогоднішній день, проблема неправдивих новин є глобальною, адже вона впливає на політичну ситуацію у багатьох країнах (Україна, США, Німеччина, Китай тощо).

Багато вчених вірять, що проблема визначення неправдивих новин може бути вирішена за допомогою методів штучного інтелекту. Причиною для цього є те, що алгоритми штучного інтелекту дуже добре проявили себе у задачах класифікації у декілька попередніх років.

Ця стаття показує один із алгоритмів визначення неправдивих новин за допомогою наївного баєсового класифікатора — одного з найпростіших методів класифікації за допомогою штучного інтелекту [1].

Схожість між спам-повідомленнями та неправдивими новинами

Спам-повідомлення та неправдиві новини мають багато спільних рис. Приведемо деякі з них:

- Велика кількість граматичних помилок
- Часто — емоційно забарвлені
- Часто носять маніпуляційний характер
- Використовують певний обмеженеий набір слів та словосполучень

Отже, можна зробити висновок, що для визначення спам-повідомлень та неправдивих новин можна використовувати аналогічні методи [2].

Наївний баєсів класифікатор

У машинному навчанні, наївний баєсів класифікатор — простий ймовірнісний алгоритм, що базується на застосуванні теореми Баєса [3].

Наївніий баєсів класифікатор є доволі популярним методом для визначення спам-повідомлень.

Основна ідея наївного баєсівого класифікатора — розглядати кожне слово в текстів повідомлення незалежно. Як вже зазначалось раніше, неправдиві новини часто використовують слова-індикатори, що вказують на можливу неправдивість тексту. Звичайно, не можна казати, що якщо у деякому тексті зустрілись деякі слова, то цей текст однозначно є неправдивим, проте такі слова впливають на ймовірність такого факту. Формула для обрахунку умовної ймовірності того, що текст є неправдивим при умові, що він містить певні слова, виглядає наступним чином:

$$Pr(F|W) = Pr(W|F)\cdot Pr(F)/(Pr(W|F)\cdot Pr(F)+Pr(W|T)\cdot Pr(T)), \quad (1)$$

де:

Pr(F|W) – умовна ймовірність того, що текст є неправдивим;

Pr(W|F) – умовна ймовірність знаходження слова W у неправдивих текстах;

Pr(F) – апріорна ймовірність того, що будь-який новинний текст є неправдивим;

Pr(W|T) – умовна ймовірність знаходження слова W у правдивих текстах;

Pr(T) – апріорна ймовірність того, що будь-який новинний текст є правдивим.

Припустимо, що ми знаємо Pr(F|W) для кожного слова у тексті. Наступний крок — комбінування цієї інформації для визначення ймовірності того, що текст є неправдивим.

Формули для цього виглядають так:

$$p1 = Pr(F|W1)\cdot ...\cdot Pr(F|Wn), \quad (2)$$

$$p2 = (1 – Pr(F|W1))\cdot ...\cdot (1 – Pr(F|Wn)), \quad (3)$$

$$p = p1/(p1 + p2), \quad (4)$$

Результат обрахунку формули 3 і є шуканою ймовірністю того, що текст є неправдивим.

Останнє запитання — як обраховувати ймовіросні Pr(W|F) та Pr(W|T) для кожного слова. Припустимо, що існує тренувальний набір із великою кількістю новин, помічених як правдиві чи неправдиві. Тоді можна визначити ймовірність знаходження слова у правдивому тексті як відношення кількості правдивих текстів із цим словом до загальної кількості правдивих текстів. Аналогічним чином обраховуються ймовірності для неправдивих текстів.

Загальний огляд тренувального набору даних

Набір даних, що використовувався для тренування і тестування, було зібрано командою BuzzFeed News. Набір даних представляє собою інформацію про Facebook публікації, кожна з яких містить інформацію про новини. Працівники BuzzFeed перевірили правдивість кожної з цих новин вручну. Вони позначили кожну з публікацій як “правдива”, “неправдива” і “суміш правди і неправди”. Вони також зібрали додаткову інформацію: кількість лайків, поширень тощо. Тренувальний набір містить 2282 приклади [4].

Отримані результати

Тестування реалізованого алгоритму показало наступні результати:

- Правдиві новини: точність класифікації 75.59%
- Неправдиві новини: точність класифікації 71.73%
- Загальна точність класифікації: 75.40

Висновки

Результати дослідження показали, що навіть доволі простий алгоритм штучного інтелекту (такий, як наївний баєсів класифікатор) може показати хороші результати у задачі визначення неправдивих новин. Отже, результати цього дослідження ще більше підтримують гіпотезу про те, що алгоритми штучного інтелекту можуть бути успішними для розв’язання задачі про визначення неправдивих новин.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Granik M.O., Mesyura V.I. Fake News Detection Using Naive Bayes Classifier / Conference proceedings 2017 IEEE First Ukraine Conference on ELECTRICAL AND COMPUTER ENGINEERING (UKRCON), May 29 – June 2, 2017 Kyiv, Ukraine. – 2017. – 900-904.

2. Spamming. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Spamming. Accessed Feb. 6, 2017.

3. Naive Bayes classifier. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier. Accessed Feb. 6, 2017.
4. Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, Jeremy Singer-Vine. (2016, Oct. 20). Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. [Online]. Available: https://www.buzzfeed.com/craigsilverman/partisan-fb-pages-analysis?utm_term=.twM44ywz1B#.cxEnnGWD6g

Гранік Михайло Олександрович — аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, e-mail: Fcdkbear@gmail.com.
Володимир Іванович Месюра — кандидат технічних наук, доцент, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Granik Mykhailo O. — Postgraduate student of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia, e-mail: Fcdkbear@gmail.com.
Mesyura Volodymyr I. — Cand. Sc., Assistant professor, professor of the Computer Science Chair, Vinnytsia National Technical University, Vinnytsia