Семантичний Аналіз Великих Даних в Задачах Кібербезпеки

Вероніка Островська
кафедра захисту інформації
Вінницький національний технічний університет
Вінниця, Україна
vera_nika.ostrovskaya@mail.ru

Олеся Войтович
кафедра захисту інформації
Вінницький національний технічний університет
Вінниця, Україна
voytovych.op@gmail.com

Big Data Semantic Analysis in the Cybersecurity Tasks

Veronika Ostrovska
dept. of Information Protection
Vinnytsia National Technical University
Vinnytsia, Ukraine
vera_nika.ostrovskaya@mail.ru

Olesia Voitovych
dept. of Information Protection
Vinnytsia National Technical University
Vinnytsia, Ukraine
voytovych.op@gmail.com

Анотація—В роботі надана загальна постанова процедури класифікації текстів, описані основні підходи, що використовуються в задачі класифікації семантичних даних, розглянуті найбільш поширені математичні методи класифікації та регресійний аналіз текстових документів. Розкрито особливості використання, переваги та недоліки зазначених методів. Зроблено висновок щодо необхідності подальшого розроблення алгоритмів класифікації на основі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати при навчанні та високу якість класифікації в реальних завданнях.

Abstract—The general terms in classification of texts, main approaches, which are used in classification of semantic data, are presented at the article. Both mathematical classification and text documents regression analysis methods, which are the most common ones, were researched. The features, advantages and disadvantages of these methods were determined. The deduction of the further development necessarily of classification algorithms based on these methods which would be simple for implementation, efficient, low computational costing in training and high classification quality of real data, was yielded.

Ключові слова—Великі Дані, кібербезпека, аналіз тональності, семантичні дані, машинне навчання, словник тональності, класифікація, опорний вектор, Байєсівський класифікатор, k-найближчий об'єкт, регресія

Keywords—Big Data, cybersecurity, sentiment analysis, semantic data, machine learning, dictionary of tonality, classification, support vector, Bayesian classifier, k-nearest object, regression

I. Вступ

У сфері захисту від сучасних кібератак однією із новітніх тенденцій є використання аналітичних систем кібербезпеки із застосуванням систем машинного навчання і штучного інтелекту на великих обсягах даних (Machine Learning-based Security Analytics using Big Data). Ці системи дають змогу помічати відхилення у поводженні систем чи користувачів від норми і таким чином виявляти більшість небезпечних кібератак [1].

У джерелі [2] наведено спосіб збирання даних із соціальної мережі Facebook з допомогою функцій мови програмування R. За основу дослідником було взято подію – другий тур дебатів кандидатів у президенти США Дональда Трампа та Гілларі Клінтон, що відбулися 10 жовтня 2016 року. Це були гострі політичні протистояння, що спричинили інформаційну війну між таборами прихильників кандидатів. Результати дослідження довели, що реакція людей та підтримка Гілларі Клінтон і Дональда Трампа після дебатів змінилася.

Аналітична обробка Великих Даних дає змогу накопичувати знання, виявляти закономірності і виробляти оптимальні методи. Після терористичного акту на Бостонському марафоні у 2013 році великі набори повідомлень, знімків і відеозаписів із соціальних мереж класифіковано та проаналізовано з допомогою високопродуктивних систем, що в кінцевому підсумку допомогло виявити організаторів теракту [3]. Таким чином, хмари надали обчислювальні потужності для вирішення завдання, а результативність роботи

автоматизованих засобів аналітики було покращено завдяки інформації від учасників соціальних мереж [3].

II. Результати дослідження

Важливою проблемою аналізу текстів у глобальній мережі є забезпечення безпеки і виявлення потенційно небезпечних або небажаних повідомлень (поширення спаму і вірусів, терористичні загрози тощо) [4; 5].

Віртуальні спільноти все активніше і масштабніше використовують у власних інтересах засоби інформаційно-психологічного впливу. Вони надають широкі можливості в плані впливу на формування громадської думки, прийняття політичних, економічних і військових рішень, впливу на інформаційні ресурси противника і поширення спеціально підготовленої інформації (дезінформації) [6]. Процеси в соціальних мережах викликають підвищений інтерес в науці, однак темпи теоретичних досліджень істотно відстають від темпів розвитку соціальних мереж [7].

Є два основних підходи до проблеми аналізу тональності семантичних даних: підхід, що ґрунтується на методах машинного навчання, та підхід, що ґрунтується на використанні словників тональної лексики. У першому підході завдання аналізу тональності зводиться до класифікації текстів, яка може бути вирішена шляхом навчання класифікатора на заздалегідь розміченій колекції текстів. Під класифікацією текстів (Text Categorization) розуміється розподіл текстових документів по заздалегідь визначених категоріях.

Загальна методика класифікації полягає в тому, щоб за допомогою набору прикладів із кожного класу відшукати правила, які можуть бути застосовані до нових прикладів. Це один з найважливіших режимів машинного навчання [8].

Основою другого підходу є аналіз тональності окремих слів (термів) у тексті і подальше визначення тональності тексту згідно з оцінками окремих слів, що входять до даного тексту. Для цього в основному використовуються словники тональності, в яких кожному слову відповідає величина, яка відображає «вагу» слова в тональності всього тексту. У подальшому згідно із запропонованим методом будується функція, яка на вхід приймає кількість входжень у текст кожного слова й обчислює агреговану величину тональності всього тексту.

Однак проблемою такого аналізу є те, що не завжди можна просто визначити точне емоційне забарвлення тексту, опираючись тільки на окреме слово. Поширення набули слова, які в сукупності можуть мати зовсім інший емоційний зміст, ніж поодинці. Або ж текст може містити велику кількість негативних або позитивних слів і все одно виражати протилежну думку. Тому одним із напрямів аналізу тональності тексту є вибір методів таким чином, щоб здійснювати класифікацію максимально точно, враховуючи різні можливі комбінації.

Завдання методів аналізу тональності текстів полягає в тому, щоб якнайкраще обрати такі ознаки і сформулювати правила, відповідно до яких прийматиметься рішення про віднесення документа до певного класу. До найвідоміших методів аналізу тональності текстів на основі машинного навчання відносяться метод Байєсової (наївної) класифікації, метод опорних векторів, метод k-найближчого сусіда та регресія.

Метод Байєсової (наївної) класифікації використовує ймовірнісну модель, в якій класифікація та включення у відповідну категорію документів реалізується шляхом оцінювання ймовірності появи слів у документі. Ймовірності можуть бути використані для оцінювання найбільш близьких категорій тестового документа. Основні переваги байєсівського класифікатора – простота реалізації і низькі обчислювальні витрати при навчанні та класифікації [9]. Особливо добре цей класифікатор справляється з задачею узагальнення, коли навчальний набір містить недостатньо даних, щоб покрити весь простір ймовірностей. У тих рідкісних випадках, коли ознаки дійсно незалежні (або майже незалежні), байєсівський класифікатор (майже) оптимальний [8]. Основним недоліком методу є відносно невисока якість класифікації в більшості реальних завдань. Зазначений метод часто використовується як базовий при порівнянні різних методів машинного навчання.

Метод опорних векторів (Support Vector Machine, SVM) використовує процес пошуку площини вирішення, яка може розділити позитивні і негативні приклади в багатовимірному просторі функції, в якому навчальні документи представлені як вектори. Метод набув величезної популярності завдяки своїй високій ефективності. Результати класифікації текстів з допомогою методу опорних векторів є одними з найкращих у порівнянні з іншими методами машинного навчання [9]. Точність даного методу залежить від добору оптимальної площини вирішення; області розв’язання є ефективними за умови збільшення вхідних змінних [10]. Однак, швидкість навчання даного алгоритму є однією з найнижчих. Метод опорних векторів вимагає великого обсягу пам’яті і значних витрат машинного часу на навчання, що знижує його масштабованість. Проте даний алгоритм можна використовувати як еталон з точки зору якості класифікації. Так, метод працюватиме ефективно, якщо опорних векторів буде порівняно небагато, якщо ж їхня кількість зростатиме, то метод стає малоефективним через значно збільшену складність.

Метод k-найближчого сусіда (k-nearest neighbor) [8] як навчальну вибірку використовує набір об'єктів, кожен з яких належить до одного з двох або більше класів. Невідомий об'єкт відноситься до одного з класів за таким принципом: знаходяться k-найближчі об'єкти з навчальної вибірки в просторі образів (зазвичай використовується міра відстані Евкліда). Потім визначається, до якого класу належить більшість найближчих об'єктів навчальної вибірки – до цього класу належить і невідомий об'єкт.

Даний метод є найпоширенішим алгоритмом плоскої кластеризації, однак його недоліком є необхідність зберігання в оперативній пам’яті комп’ютера всіх об’єктів для порівняння кожного із них з невідомим об’єктом [8]. У дослідженнях, присвячених аналізу роботи різних

алгоритмів машинного навчання для задачі класифікації текстів, цей метод демонструє одні з найкращих результатів [11].

Головною особливістю, що виділяє зазначений метод серед інших, є відсутність у нього стадії навчання. Належність документа відповідному класу визначається без побудови функції класифікації. Основною перевагою такого підходу є можливість оновлювати навчальну вибірку без перенавчання класифікатора [12].

У великих даних задачу класифікації розглядають як визначення значення одного з параметрів об'єкту на основі значення інших параметрів [13]. Задача регресії подібна до задачі класифікації і дозволяє визначити за відомими характеристиками об'єкту значення деякого його параметру. Тут значенням параметру є не кінцева множина класів, а множина дійсних чисел. Перевагою цього алгоритму є те, що на виході для кожного об'єкта отримується ймовірність приналежності до класу. До найбільш дієвих видів регресії в задачах, що стосуються Big Data, відносяться логістична, лінійна та гребнева регресії.

До переваг лінійної регресії можна віднести швидкість і простоту отримання моделі, а також модель можна інтерпретувати. Лінійна модель є прозорою і зрозумілою для аналітика. За отриманими коефіцієнтами регресії можна судити про те, як той чи інший фактор впливає на результат, зробити на цій основі додаткові корисні висновки [14].

Логістична регресія має деякі недоліки, що також властиві лінійній регресії – низька стійкість до помилок, залежність від набору даних, але в загальному працює краще, і може бути приведена до вигляду лінійної регресії для спрощення обчислень [15].

Гребнева регресія – удосконалення лінійної регресії з підвищеною стійкістю до помилок, що накладає обмеження на коефіцієнти регресії для отримання більш наближеного до реальності результату [15]. До того ж, цей результат набагато простіше інтерпретувати.

Класифікація та регресія передбачають здійснення двох обов'язкових етапів. Перший етап – виділення набору об'єктів, для яких відомі значення залежних і незалежних змінних. На основі отриманого набору будується модель визначення значення залежної змінної (функція класифікації або регресії). На другому етапі побудовану модель застосовують до об'єктів, які аналізуються. Недоліком класифікації та регресії є те, що розробник системи повинен фіксувати кількість класів та характеристик, за якими буде проводитись дослідження. Це означає, що якщо система не виявить ознаки або класу, він не буде коректно оброблений.

Кожен із проаналізованих методів має свої переваги і недоліки залежно від конкретного вирішуваного завдання. Ефективність розв'язання завдання значною мірою залежить саме від особливостей засобу. У табл. 1 наведена порівняльна характеристика основних методів аналізу тональності семантичних даних, що застосовуються при обробці природної мови. Тут знак «+» означає високий, знак «–» – низький та «+/–» – нейтральний показник характеристики методу.

ТАБЛИЦЯ І. ПОРІВНЯЛЬНА ХАРАКТЕРИСТИКА МЕТОДІВ АНАЛІЗУ ТОНАЛЬНОСТІ СЕМАНТИЧНИХ ДАНИХ

Метод	**Характеристика**						
	Трудомісткість	***Швидкість***	***Масштабованість***	***Сфера використання***	***Алгоритми***	***Відстань між словами***	***Поєднання з іншими методами***
Метод Байєсової (наївної) класифікації	+\-	+\-	-	Аналіз переходів по web-сторінках	Класичні методи статисти–ки	+	-
Метод опорних векторів	+	+	-	Класифіка ція та регресій–ний аналіз даних	Методи аналізу	+	+
Метод k-середніх	+\-	+	-	Машинне навчання	Методи ієрархіч–ного аналізу	+	-
Регресія	+\-	+	+	Оцінка допусти–мості прийняття рішення	Класичні методи статисти–ки	-	-

Для підвищення оцінок ефективності алгоритмів, ймовірно, слід доповнити їх елементами лінгвістичного аналізу. Таким чином, необхідною умовою для аналізу тональності є складання словарного списку тональної лексики. Для цього в тексті виділяються оціночні слова, для них обчислюється емоційна вага, потім ці ваги об'єднуються за допомогою деякої функції (наприклад, середнє арифметичне або сума). Існує кілька підходів до вилучення оцінних слів і обчисленню їх емоційної ваги.

В джерелі [16] спочатку вибираються дві еталонних множини оціночних слів: позитивна і негативна. Далі з відгуків вибираються набори, що складаються з прикметників в поєднанні з іменниками і прислівники в поєднанні з дієсловами. Автор Turney використовує набори, вважаючи, що, хоча ізольоване слово може вказувати на суб'єктивність, його може виявитися недостатньо для визначення контексту емоційної оцінки. Тональність відгуку розраховується як середнє арифметичне емоційних оцінок наборів, взятих з цього відгуку. Для розрахунку емоційної оцінки для набору Turney використовував пошукову систему Altavista.

Множини оціночних слів також створюються вручну експертами. Для поповнення даних множин можуть використовуватися словники. В джерелі [17] запропонований метод, який використовує тезаурус для поповнення заданої вручну множини оціночних слів. Ідея методу полягає в наступному: якщо слово оціночне, то його синоніми також будуть оціночними і відносяться до однієї тональності, а антоніми – до протилежної

тональності. Ще один підхід представлений в роботі [18], де за допомогою тлумачень слів в словнику визначається їх орієнтація. Даний метод ґрунтується на ідеї, що слова з однаковою емоційною оцінкою мають схожі тлумачення.

В джерелі [19] описується підхід до автоматичного створення словника оціночної лексики в області товарів і послуг для російської мови ProductSentiRus. Словник ProductSentiRus був отриманий застосуванням навченої моделі до наборів відгуків в декількох предметних областях. Словник представлений як список 5 тисяч слів, упорядкованих у міру зниження обчисленої ймовірності їх оціночності без вказівки позитивної або негативної тональності. Для покращення якості списку оціночних слів використовується тезаурус російської мови РуТез [20].

На вхід алгоритму надходить список слів, що впорядковані за ймовірністю їх оціночності. Основна ідея уточнення оціночного лексикону – це автоматичний розподіл понять тезауруса на оціночні та нейтральні, і подальше використання отриманої розмітки для переранжування списку оціночних слів. Цей процес регулюється за допомогою раніше отриманих ваг оціночності слів [19].

Чим більше ресурсів, таких як семантично розмічені корпуси та тезауруси, розроблено для природної мови, тим вище вдається дослідникам підняти якість вирішення цього завдання. На даний момент найбільший синтаксично розмічений корпус російською мовою – SynTagRus [21], який входить до складу Національного корпусу російської мови. На сайті Національного корпусу російської мови (www.ruscorpora.ru) також містяться корпуси для інших мов, в тому числі корпус текстів української мови, однак він доступний лише для онлайн-пошуку.

Переважна більшість кращих підходів у задачах класифікації семантичних даних ґрунтуються на застосуванні методу опорних векторів SVM, який може комбінуватися з додатковими ресурсами на зразок словників або правил. Даний результат узгоджується з результатами, отриманими для завдання аналізу тональності англомовних текстів, де також було доведено, що метод опорних векторів зазвичай породжує кращі за якістю результати в цьому завданні.

III. Висновки

В процесі дослідження методів аналізу тональності текстових даних надано загальну процедуру класифікації текстів, наведено огляд підходів до вирішення задачі класифікації, описані основні підходи, що використовуються в задачі класифікації текстів, визначено етапи процесу класифікації та розглянуті найбільш поширені математичні методи класифікації текстових документів. Розкриті особливості використання, переваги та недоліки зазначених методів дають змогу зробити висновок про необхідність подальшого вдосконалення алгоритмів класифікації на основі зазначених методів, що були б простими в реалізації, ефективними, мали низькі обчислювальні витрати при навчанні та високу якість класифікації в реальних завданнях.

Література References

[1] Експерт: кіберзахист – це не параноя [Електронний ресурс]. – Режим доступу : http://www.bbc.com/ukrainian/features-39364360. – Назва з екрану.

[2] Voitovych O. Badania sieci społecznych jako źródła informacji w czasie wojny / Voitovych O., Holovenko V. // Inżynier XXI wieku projectujemy przyszlosc : monografia / [pod red. : Jacek Rysiński]. – Bielsko-Biała : Wydawnictwo Naukowe Akademii Techniczno-Humanistycznej w Bielsku-Białej, 2016. – С. 111–119.

[3] Analytics of Big Data and social networks [Електронний ресурс]. – Режим доступу : http://www.osp.ru/os/2013/08/13037856. – Назва з екрану.

[4] Kontostathis A. Text mining and cybercrime / Kontostathis A., Edwards L., Leatherman A. // Text Mining. Applications and Theory / ed. by Berry M. W., Kogan J. – Chichester : Wiley. – 2010. – P. 149–164.

[5] Dua S. Data Mining and Machine Learning in Cybersecurity / Dua S., Du X. – New York. – 2011. – 224 p.

[6] Гриненко І. Вплив віртуальних спільнот на інформаційну безпеку: сучасний стан та тенденції розвитку / І. Гриненко, Д. Прокофьєва-Янчиленко // Правове, нормативне та метрологічне забезпечення систем захисту інформації в Україні. – 2012. – № 1 (23). – С. 18–23.

[7] Ефимов Е. Г. Социальные интернет-сети (методология и практика исследования : монография / Е. Г. Ефимов / Волгоградский гос. тех. ун-т. – Волгоград. – 2015. – 169 с.

[8] Коэльо Л. П., Ричарт В. Построение систем машинного обучения на языке Python. 2-е издание / пер. с англ. Слинкин А. А. – М.: ДМК Пресс. – 2016. – 48 – 167 с.

[9] Text Mining: Aplications and theory / ed. by Berry M. W., Kogan J. [Електронний ресурс]. – Режим доступу : http://www.mohamedrabeea.com/books/book1_1165.pdf. – Назва з екрану.

[10] Thorsten J. Text Categorization with Support Vector Machines: Learning with Many Relevant Features [Електронний ресурс]. – Режим доступу : https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf. – Назва з екрану.

[11] Sebastiani F. Machine learning in automated text categorization / F. Sebastiani // ACM Comput. Surv. – 2010. – P. 1-47.

[12] Yang Y. A re-examination of text categorization methods / Y. Yang, X. Liu // Proc. of Int. ACM Conference on Research and Development in Information Retrieval (SIGIR-99). – 2007. – P. 42-49.

[13] Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Viual Mining, Text Mining, OLAP. – Спб.: БХВ-Петербург, 2007. – 384 с.

[14] Линейная регрессия [Електронний ресурс]. – Режим доступу : https://basegroup.ru/deductor/function/algorithm/linear-regression. – Назва з екрану.

[15] 10 типов регрессии – какой выбрать? [Електронний ресурс]. – Режим доступу : http://datareview.info/article/10-tipov-regressii-kakoy-vyibrat/. – Назва з екрану.

[16] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.

[17] Hu M., Liu B. Mining and Summarizing Customer Reviews // KDD, Seattle, 2004, pp. 168–177.

[18] Esuli A., Sebastiani F. Determining the Semantic Orientation of Terms through Gloss Classification // Conference of Information and Knowledge Management (Bremen). ACM, New York, NY, 2005, pp. 617–624.

[19] Лукашевич Н. В., Четверкин И. И. Комбинирование тезаурусных и корпусных знаний для извлечения оценочных слов, Системы и средства информ., 2015, том 25, выпуск 1, С. 20–33.

[20] О лингвистической онтологии «Тезаурус РуТез» [Електронний ресурс]. Режим доступу: http://www. labinform. ru/pub/ruthes/index.htm. – Назва з екрану.

[21] Синтаксически размеченный корпус русского языка: информация для пользователей [Електронний ресурс]. Режим доступу: http://www.ruscorpora.ru/instruction-syntax.html. – Назва з екрану.