

Метод розпізнавання терористичних загроз у текстових повідомленнях

Бісікало О.В.¹, Богач І.В.²

¹ Проф., д.т.н., декан факультету комп'ютерних систем та автоматики, Вінницький національний технічний університет, Хмельницьке шосе 95, м. Вінниця, Україна, obisikalo@gmail.com
к.т.н., доцент кафедри АІВТ, Вінницький національний технічний університет, Хмельницьке шосе 95, м. Вінниця, Україна, ilona.bogach@gmail.com

Анотація — Запропоновано метод розпізнавання терористичних загроз у текстових повідомленнях на основі моделі образного мислення людини. Задачу класифікації текстів загрозливого характеру представлено як ідентифікацію параметрів класифікаційної функції шляхом машинного навчання. Обґрунтовано можливість застосування чисельних параметрів складних залежностей між реченнями текстового документу для проведення індексації текстової інформації та зменшення розмірності векторів лінійної on-line класифікації. Отримано програмний прототип системи моніторингу загроз на основі лінгвістичного пакету DKPro Core.

Ключові слова: *текстові повідомлення, терористичні загрози, складні залежності, лінійний класифікатор, машинне навчання.*

Terrorist threats detection method in the text messages

Bisikalo O.V.¹, Bogach I.V.²

¹ Prof., Head of Department of Computer Systems and Automation, Vinnytsia National Technical University 95 Khmelnytske shose, Vinnytsia, Ukraine, obisikalo@gmail.com

² Prof., Department of Computer Systems and Automation, Vinnytsia National Technical University 95 Khmelnytske shose, Vinnytsia, Ukraine, ilona.bogach@gmail.com

Abstract — In this work, a method for recognition of terrorist threats in the text messages is suggested, basing on the human creative thinking model. The text classification task is represented as identification of the classification function's parameters by means of machine learning. Also this article substantiates the possibility to use the numerical parameters of complex relationships between the sentences of the document text for indexing textual information and reducing the dimension of the linear vectors on-line classification. The software prototype of the threats monitoring system that is based on DKPro Core linguistic package was retrieved.

Keywords: *text messages, terrorist threats, complex relationships, linear classifier, machine learning.*

ВСТУП

Останніми роками обсяг електронних документів зростає настільки швидко, що в багатьох випадках людина фізично неспроможна осягнути закладену в них інформацію. Практично єдиним способом розв'язання цієї проблеми можна вважати застосування формальних методів опрацювання текстів, здатних автоматизувати частину необхідних виробничих операцій, наприклад пошуку та фільтрації певних повідомлень. Проте машинна обробка текстової інформації супроводжується низкою специфічних труднощів, пов'язаних, у першу чергу, з вирішенням семантико-залежних задач аналізу і синтезу природно-мовного контенту. Недостатній рівень розуміння природної мови в сучасних засобах комп'ютерної лінгвістики складає основну перепону для розпізнавання таких типів повідомлень, де сенс зазвичай передається інваріантними мовними структурами або навмисно завуальовується. Для людини виявлення прихованого змісту повідомлення не є надскладною задачею, проте модерація відповідальними особами

потужного мовного потоку обмежена порівняно невеликою швидкістю сприйняття тексту.

Розбудова інформаційного суспільства, що не визнає кордонів та штучних перепон, супроводжується новими ризиками, зокрема локальними терористичними загрозами. Швидкість та легкість спілкування через соціальні мережі або інші інформаційно-комунікаційні технології дозволяє значно прискорити створення небезпечних засобів та пристроїв, а також організувати їх злочинне застосування. Проте природно-мовний характер інформації, що передається в таких випадках, має важливу перевагу – автоматизація розпізнавання терористичних загроз може спиратися на відомі формальні методи класифікації текстів (*Text Categorization*, TC). Під класифікацією розуміють сортування текстових документів за заздалегідь визначеними категоріями, а методи TC належать до перетину двох областей – інформаційного пошуку (*information retrieval*) та машинного навчання (*machine learning*). Спільним у цих підходах є способи уніфікованого представлення документів та оцінки якості TC, а

відрізняються, власне, тільки методи класифікації [1].

ОСНОВИ МЕТОДУ РОЗПІЗНАВАННЯ ТЕРОРИСТИЧНИХ ЗАГРОЗ У ТЕКСТОВИХ ПОВІДОМЛЕННЯХ

У відомих методах оперативного моніторингу загрозові текстові повідомлення відбирають шляхом аналізу їх словарного складу та пошуку в них слів з «терористичного» словника, наприклад – [2]. Проте точність такого відбору є надзвичайно низькою внаслідок лексичної омонімії та контекстного характеру вживання будь-якого слова. Це призводить до численних помилок 1-го роду – нейтральне повідомлення отримує статус загрозового (*false positive, FP*), а черга текстів, що потребують модерації, починає перевищувати можливості відповідної системи моніторингу загроз. З іншого боку, запобігання цьому шляхом збільшення нижньої межі для кількості різних слів з «терористичного» словника або визначення окремих з них як ключових слів усього тексту може мати гірший наслідок у вигляді помилок 2-го роду – загрозове повідомлення не розпізнається (*false negative, FN*). Зрозуміло, що алгоритм розпізнавання має максимально «розуміти» зміст повідомлення, у т.ч. прихований, інакше застосування зловмисниками синонімічних та метафоричних конструкцій унеможливить визначення загроз через порівняння з «терористичним» словником.

Для підвищення точності класифікації та аналізу прихованого змісту терористичних загроз пропонується метод на основі аналізу синтаксичних зв'язків між словоформами в реченнях текстового повідомлення. Метод базується на застосуванні моделі образного мислення людини та нечіткого відношення сенсу [3] і забезпечує самовдосконалення бази знань системи моніторингу загроз. Згідно з парадигмою машинного навчання класифікаційне правило будується поступово за допомогою тренувальної колекції. Такий підхід забезпечує якість класифікації, яка може порівнюватися з якістю класифікації, що здійснюється людиною.

Розглянемо множину документів $D = \{d_1, \dots, d_{|D|}\}$, інформацію про кожний з яких можна отримати тільки з самого документа без залучення зовнішніх джерел. Відомою також є $C = \{c_1, \dots, c_{|C|}\}$ – множина категорій або абстрактних міток, кожна з яких може помічати один з документів множини D . Нехай $\Phi: D \times C \rightarrow \{0,1\}$ – невідома цільова функція, що за парою $\langle d_i, c_j \rangle$ визначає, чи належить документ d_i категорії c_j (1 або Т) або ні (0 або F). Задача полягає у побудові (класифікаційної) функції Φ^k , максимально близької до Φ .

ОТРИМАННЯ ФОРМАЛЬНИХ ПАРАМЕТРІВ ТЕКСТУ

Для отримання формальних параметрів тексту та побудови відповідного класифікатора текстові документи індексуються, як правило, на основі частотних характеристик термів (слів) – такий підхід вважають базовим. Головною проблемою при цьому є зменшення розмірності векторів, якими представлені тексти та які пропорційні кількості різних слів, що зустрілися в колекції документів [4]. Для зменшення розмірності векторів використовують як евристичні (фільтрація рідких та найчастотніших слів (прийменники, артиклі), визначення коефіцієнтів корисності окремих термів, їх групування тощо), так і математичні методи (сингулярний розклад матриць). Проте маємо додаткову проблему – уникнення авторами повідомлень слів загрозового характеру заздалегідь нівелює ефективність зазначених методів та знижує якість класифікатора.

У роботі [5] було отримано формальні переваги складних залежностей на противагу традиційному частотному словнику окремого тексту з огляду на визначення семантично-залежних параметрів цього тексту. З метою розв'язання задачі побудови функції класифікатора загроз Φ^k пропонується застосувати можливості сучасних лінгвістичних пакетів для визначення складних залежностей між реченнями текстового документу та покласти отримані чисельні параметри в основу процедури індексації текстової інформації. У першому наближенні спосіб визначення порога для лінійного класифікатора та, власне, алгоритм лінійної on-line класифікації можна обрати стандартними.

У базовому підході текст розглядається як мультимножина термів (слів). Кожному слову ставиться у відповідність певне число (вага) – узагальнена характеристика частоти знаходження цього слова у тексті. Порядок слів, як правило, не враховується, а окрім частоти знаходження (головна ознака) враховується, зазвичай, додаткові ознаки, такі як «слово зустрілося в заголовку», «слово позначено іншим кольором», «слово позначено напівжирним або курсивом або іншими відмінними параметрами шрифту» тощо. На основі всіх цих ознак для кожного слова в тексті розраховується його вага.

Одним з найбільш популярних способів представлення ваги слова вважається $TF \cdot IDF$ ($TF = \text{term frequency}$, $IDF = \text{inversed document frequency}$). TF показує наскільки часто слово зустрічається в конкретному документі, а IDF навпаки – наскільки рідко слово зустрічається в усіх документах деякої колекції. Іноді проводиться нормалізація по окремому документу для того, щоб сума квадратів усіх ваг для нього дорівнювала 1. Власне кожний текстовий документ за таким підходом можна вважати вектором у багатовимірному просторі, координати якого – це номери слів, а значення координат – значення відповідних ваг. Тоді розмірність вектора дорівнює кількості слів, що зустрічаються в колекції документів. Але з-за того,

що враховуються всі слова, які коли-небудь зустрічалися в усій колекції документів, отримані вектори є розрідженими, тобто такими, що мають надвелику кількість координат, причому значення більшості з них – нулі.

На відміну від базового у підході, що пропонується, замість слів (термів) координатами вектора вважаються зв'язки між словами. Таке, на перший погляд, парадоксальне рішення у квадратичній залежності лише посилює проблему розмірності векторів. Але, відповідно до моделі образного мислення людини [3], найбільш інформативними змістовними ознаками тексту можна вважати асоціативні зв'язки між мовними образами, які об'єднують лема та словоформи споріднених за ознакою спільного кореня, але, можливо, навіть відмінних за частинами мови слів. Тому розмірність більш інформативного вектору текстового документу на основі складних залежностей між словоформами [5] не перевищує розмірність відповідного вектору за базовим підходом. Технологічно визначення складних залежностей забезпечують сучасні лінгвістичні пакети, зокрема DKPro Core (на основі фреймворку Apache UIMA та з програмною підтримкою Java / Maven / Eclipse) для англomовного тексту – цю платформу було обрано для створення програмного прототипу системи моніторингу загроз.

Формально ранжувальним класифікатором будемо вважати визначення функції $CSV_i : D \rightarrow [0,1]$, що для кожного документу d_j повертає значення приналежності (*categorization status value*) d_j до c_i . Побудувати точний класифікатор можна таким чином: або відразу будувати функцію $CSV_i : D \rightarrow \{T, F\}$ або обчислити аналогічну ранжувальну функцію $CSV_i : D \rightarrow [0,1]$, а потім визначити поріг (*threshold*) τ_i такий, що $CSV_i \geq \tau_i$ інтерпретується як T , а $CSV_i < \tau_i$ інтерпретується як F .

Зазвичай класифікатори за методами машинного навчання будуються на основі колекцій $\Omega = \{d_1, \dots, d_{|\Omega|}\}$ заздалегідь класифікованих експертами документів, тобто таких, для яких значення цільової функції точно відомо. Отже, для визначення якості, наприклад, бінарної класифікації загрозливих текстових повідомлень потрібно колекцію $\Omega = \{d_1, \dots, d_{|\Omega|}\}$ відповідних документів розбити на множини, що не перетинаються – T_r навчальної (*training*), V_a перевіркової (*validation*) та T_e тестової (*test*) колекцій. Якщо перші дві з них використовуються для побудови (T_r) та оптимізації параметрів (V_a) класифікатора, то множину незалежних з точки зору створення класифікатора документів T_e потрібно застосувати для

визначення метрик повноти π (доля вірно знайдених загрозливих документів серед усіх загрозливих документів) та точності ρ (доля справді загрозливих документів серед усіх загрозливих документів, що визначені так класифікатором). Формально

$$\pi = \frac{TP}{TP + FN}, \quad (1)$$

$$\rho = \frac{TP}{TP + FP}, \quad (2)$$

де TP – число вірно знайдених загрозливих документів (*true positive*).

ВИСНОВКИ

Запропоновано метод розпізнавання терористичних загроз у текстових повідомленнях на основі моделі образного мислення людини. Метод реалізовано на платформі фреймворку Apache UIMA та лінгвістичного пакету DKPro Core з програмною підтримкою Java / Maven / Eclipse. Отриманий програмний прототип системи моніторингу загроз буде мережу мовних образів (значимих словоформ) повідомлення шляхом накопичення складних синтаксичних зв'язків між лемами. Формальний аналіз отриманого графу лексичної онтології дозволяє визначити ступінь його близькості до класифікованого простору аналогічних графів, що відповідають відомим типам загроз. Запропонований метод забезпечує отримання чисельних показників повноти (1) та точності (2) визначення загрозливих документів.

ЛІТЕРАТУРА REFERENCES

- [1] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization [E-resource] / Available: <http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf>.
- [2] Облік та контроль ядерного матеріалу, фізичний захист ядерного матеріалу і ядерних установок. Тлумачний словник українських термінів. Словники термінів.: українсько-англо-російський, русско-украино-английський, English-Russian-Ukrainian / Затверджено наказом Державного комітету ядерного регулювання України від 08.06.2004 р. № 101. – НП 306.7.086-2004. – 165 с
- [3] Bisikalo O. Solving problems on base of concepts formalization of language image and figurative meaning of the natural-language constructs [E-resource] / Oleg Bisikalo, Slawomir Cieszczyk, Gulbahar Yussupova // 16th Conference on Optical Fibers and Their Applications. International Society for Optics and Photonics. – December 18, 2015. – Pp. 98161U-98161U-14. – Available:
- [4] <http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleId=2478661>.
- [5] Юрий Лифшиц. Автоматическая классификация текстов [E-resource] / Available:
- [6] <http://logic.pdmi.ras.ru/~yura/internet/06ianote.pdf>.
- [7] Бісікало О.В. Статистичний аналіз складних залежностей у тексті / О.В. Бісікало // Вісник Нац. ун-ту "Львівська політехніка": Інформаційні системи та мережі. – 2015. – № 814. – С. 228–236.