УДК 004.91
МРНТИ 28.23.21

ЛИНГВИСТИЧЕСКИЕ ИНСТРУМЕНТЫ ВЫЯВЛЕНИЯ КРИМИНАЛЬНО ОКРАШЕННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ВЕБ-КОНТЕНТА

О.Ж. МАМЫРБАЕВ[1], К.Ж. МУХСИНА[1, 2], Н.Ф. ХАЙРОВА[1], А.С. КОЛЕСНИК[3]

[1]Институт информационных и вычислительных технологий КН МОН РК
[2]Казахский Национальный университет имени аль-Фараби
[3]Национальный технический университет «Харьковский политехнический институт», Украина

***Аннотация*:** *В работе рассматриваются виды криминально окрашенной текстовой информации Web-контента (киберпреступность, террористический акт или финансовое мошенничество) и анализируются существующие технологии лингвистического анализа, позволяющие выявлять противоправную информацию в текстах. Проводится аналитический обзор использования существующих инструментов обработки языка, позволяющий выявить проблемы использования традиционных подходов NLP для анализа криминально значимой текстовой информации.*
Предлагаемый метод базируется на подходах Information Extraction и фокусируется на методе извлечения фактов из слабоструктурированных текстов. Рассматривается использование технологии, базирующейся на описании семантических функций средствами алгебры конечных предикатов, для извлечения слабоструктурированных фактов из предложений русского и английского языков. Анализируется возможность использования предложенной технологии для текстов казахского языка.

Ключевые слова: *терроризм, мошенничество, киберпреступность, Natural Language Processing, Facts Extraction, алгебра конечных предикатов, тексты русского, английского и казахского языков, семантические функции*

LINGUISTIC INSTRUMENTS OF DETECTING CRIMINALIZED TEXT INFORMATION OF WEB CONTENT

Abstract: *The paper deals with types of criminally colored textual information of Web content (cybercrime, terrorist act or financial fraud) and analyzes existing technologies of linguistic analysis that allow to identify illegal information in texts. An analytical review of the use of existing language processing tools is conducted, which allows to identify problems of using traditional NLP approaches for the analysis of criminal-significant textual information.*
The proposed approach is based on the approaches of Information Extraction and focuses on the method of extracting facts from weakly structured texts. The use of technology based on the description of semantic functions by means of algebra of finite predicates is considered, to extract weakly structured facts from sentences of Russian and English. The possibility of using the proposed technology for the analysis of the Kazakh language texts is analyzed.

Keywords: *terrorism; fraud; cybercrime; Natural Language Processing; Facts Extraction; algebra of finite predicates; texts of Russian, English and Kazakh languages; semantic functions*

WEB-КОНТЕНТТІҢ КРИМИНАЛДЫҚ МӘНДІ МӘТІНДІК АҚПАРАТЫН АНЫҚТАУДЫҢ ЛИНГВИСТИКАЛЫҚ ҚҰРАЛДАРЫ

***Аңдатпа*:** *Жұмыста Web-контенттің (киберқылмыстық, террористік акт немесе қаржылық алаяқтық) криминалдық мәнді мәтіндік ақпарат түрлері қарастырылып, мәтіндердегі заңға қайшы ақпаратты анықтауға септігін тигізетін қолданыстағы лингвистикалық талдау технологияларына талдау жасалған. Криминалдық мәнді мәтіндік ақпаратты талдауға арналған NLP дәстүрлі әдістерін*

пайдалану проблемаларын анықтауға мүмкіндік беретін қолданыстағы тілді өңдеу құралдарын қолдануға аналитикалық шолу жүргізіледі.
Ұсынылып отырған тәсіл Information Extraction тәсілдеріне сүйене отырып, жартылай құрылымдалған мәтіндерден фактілер алу әдісіне назар аударады. Орыс және ағылшын тілдеріндегі сөйлемдерден жартылай құрылымдалған фактілерді алу үшін соңғы предикаттар алгебрасы құралдарының көмегімен семантикалық функциялардың сипаттамасына негізделген технологияларды пайдалану көзделіп отыр. Ұсынылған технологияны қазақ тіліндегі мәтіндерді талдау үшін қолдану мүмкіндігіне талдау жасалады.

Түйінді сөздер: *терроризм, алаяқтық, киберқылмыстық, Natural Language Processing, Facts Extraction, соңғы предикаттар алгебрасы, орыс, ағылшын және қазақ тіліндегі мәтіндер, семантикалық функциялар*

Введение

В последние десятилетия, в связи с распространением сетевых компьютерных технологий, мобильной связи и Интернет, информационные ресурсы современного общества подвергаются растущему числу угроз, чреватых экономическим ущербом и ставящих под угрозу безопасность национальной информационной инфраструктуры. Подобным атакам подвергаются как государственные, так и коммерческие системы, в то время как рост криминальной активности в глобальных сетях (в таких формах, как финансовые мошенничества, нарушения авторского права, распространение детской порнографии и т.д.) создает угрозы безопасности личности и общества в целом [14]. Благодаря компьютерным сетям насильственный экстремизм может глобально распространяться, сохраняя низкую стоимость и высокую скорость.

В то же время, существующие в настоящее время технологии обработки текстов позволяют специалистам по анализу разведывательных данных и полиции осуществлять превентивную обработку текстовых данных компьютерной сети, собирая, соединяя и анализируя ‘слабые сигналы’ или ‘цифровые следы’ огромных текстовых массивов, которые присутствуют в Интернете. В некоторых случаях такой анализ может помочь обнаружить потенциал противоправного действия прежде, чем оно будет осуществлено.

В то же время, одной из главных проблем такой превентивной обработки текстов, наряду с громадным объемом информации Интернет, доступной до такого анализа [6], является проблема слабой «окрашенности» криминальных текстов для использования традиционно принятых подходов классификации, кластеризации и выделения шаблонов Natural Language Processing (NLP).

Обзор литературы

В настоящее время направление поиска противоправной информации в текстовых данных, обнаружение шаблонов преступления, и оценка риска киберпреступлений становятся одними из самых популярных исследований Natural Language Processing (NLP). Всё больше исследователей сфокусировались на способах и формах применения технологий обработки естественного языка в рамках широкого спектра видов деятельности, имеющих отношение к предотвращению террористической активности.

Одно из направлений исследований, связанных с предотвращением террористических атак, направлено на анализ использования террористами и террористическими организациями Интернета и социальных сетей [2], [3], [4], [11]. Например, одно из исследований, посвященных обнаружению “лингвистических маркеров насильственного экстремизма в онлайн среде” [6] фокусируется именно на идентификации цифровых следов, которые имеют отношение к потенциальному «террористу-одиночке» и к другим потенциальным видам насилия [5]. Для обнаружения лингвистических маркеров, которые свидетельствуют о потенциальном «предупреждающем» поведении, в работе предложено использовать списки слов насильственных действий,

подготовка и поиск которых базируется на стандартных подходах обработки текстов, таких как лемматизация и POS-тегирование, а также на использовании лексических баз данных, подобных WordNet [2]. Однако, такие лингвистические маркеры, использующиеся в качестве дополнения к стандартным алгоритмам обработки текстов, могут распознать потенциальные признаки оговоренного, заранее предполагаемого радикального насилия. Они не могут принять автоматизированные решения по любым видам преступлений. Кроме того, если отдельные этапы обработки естественного языка будут неточными, точность выделения лингвистических маркеров значительно уменьшится и количество ошибок увеличится.

Еще одним направлением NLP, используемым в рамках решения задачи выделения криминально значимой информации и потенциально связанных с терроризмом текстов, является анализ стиля текста и выявление его эмоциональной составляющей, связанной с неявным выражением намерения. Такая эмоциональная составляющая может включать хвастовство, идеологические заявления или восхищение террористическими лидерами [1]. Текстовый анализ стиля, в этом случае, позволяет обнаружить шаблоны фраз, связанных с такими эмоциональными мотивациями как гнев, унижение или позор. В данном контексте следует подчеркнуть, что стиль общения не зависит от определенной темы или от содержания. Добавление для его анализа более глубокой психологической обработки, использование лингвопсихологии речевой деятельности и социолингвистики позволяет не только идентифицировать «предупреждающее» преступление поведение, но и раскрыть в некоторых случаях корпоративное мошенничество [7].

Для анализа больших объемов разнородных текстов, тематика которых не известна заранее, используются методы классификации и кластеризации NLP. Например, объединение в кластеры может выделить такие темы как оружие, тактика или цели [1]. В этом случае, технологии распознавания речи и машинного перевода могут значительно увеличить объем текста, доступного для анализа.

Одной из разновидностей классификации является сентимент анализу. Различные формы онлайнового выражения авторского мнения (например, обзоры, личные мнения, рейтинги и рекомендации) стали основными источниками информации как для компаний, надеющихся продавать свои продукты и управлять своей репутацией [9], так и для СМИ, определяющих отношение общества к реальным событиям. Например, в работе [11] сентимент анализ используется при анализе твитов для определения мнений авторов по отношению к определенным зонам преступлений в режиме реального времени.

Многие исследования, которые сосредоточились на обнаружении шаблона преступления, используют методы сбора данных в их временном изменении. Такие исследования, кроме твитов, блогов и социальных сетей, используют информацию СМИ для обнаружения преступления в каждой определенной области [15].

Методы классификации NLP довольно хорошо разработаны и отлажены. В то же время их использование при анализе эмоциональной составляющей текста или выявления намерения не всегда дает хорошие результаты. Основным недостатком подобных подходов является неспецифичность выделения закономерностей, когда выявленные закономерности (даже если они явно угрожающие), могут быть не связаны с угрозами и их интерпретация часто зависит от культурных и отдельных особенностей человека. Анализ и классификация текстов исключает включение в классы и выбор текстов, содержащих не прямое использование терминологии, строго относящееся к криминалу. Такой терминологией может быть прямое использование названий оружия, насильственных действий, угрожающей лексики и т.д.

В рассмотренных выше статьях и подходах, при семантическом анализе эмоциональной составляющей текста, параграфы, которые представляют факты, как правило, удаляются и исследователи сосредотачиваются на параграфах, в которых автор выражает свое мнение, используя распространенные классификаторы – наивный Байесовский метод,

максимальную энтропию или support vector machine.

В данном исследовании предлагается фокусироваться именно на фактическом материале текстов, и использовать технологии, базирующиеся на подходах Information Extraction, в частности на методах извлечений фактов из слабоструктурированных текстов [17].

Предлагаемая информационно-лингвистическая технология

В текстовых документах информация о компонентных элементах состава преступления представляется в виде слабоструктурированных фактов, которые семантически сочетают партиципанты предметной области и их отношения в триаду субъект - атрибут - значение (или субъект - отношение - объект).

Поскольку такой слабоструктурированный факт обычно выражается различными нерегламентированными конструкциями естественного языка, то для его идентификации необходимо извлечь из текстовой информации некоторый предикат, выражаемый определенными глаголами, и определить партиципантов отношения, выражаемых данным предикатом (рис. 1).

Для задания таких смысловых связей предлагается использовать семантические функции, выражаемые отношениями морфологических и семантических категорий партиципантив предложения, которые могут быть описаны средствами алгебры конечных предикатов [16].

Используемая модель вводит конечное множество грамматических и семантических характеристик партиципантив предложения $M = \{m_1, \dots, m_n\}$, где n – количество указанных характеристик. Отношение между характеристиками может быть представлено в виде $m_i \bullet m_j \bullet \dots \bullet m_k$, где $m_i, m_j, \dots, m_k \in M$, а знак • – обозначает, что данные характеристики соответствуют существительному, который выполняет определенную семантическую функцию.

На множестве M вводится система предикатов S так, чтобы любой предикат $P(q_m) \in S$, равнялся 1 на множестве существительных с грамматико-семантической информацией, соответствующей определенной семантической функции, и был равным 0 в противном случае. Таким образом, множество предикатов S можно сопоставить со множеством семантико-грамматических характеристик, которые приписаны определенному партиципанту предложения.

В связи с тем, что в разных естественных языках глубинные семантические отношения выражаются различными поверхностными характеристиками, модель необходимо отдельно реализовывать для разных естественных языков. Например, для русского языка семантические функции определяются в своем большинстве падежами, в то время как для английского языка большее значение имеют предлоги.

В работе [17] для формализации семантических функций партиципантив предложе-

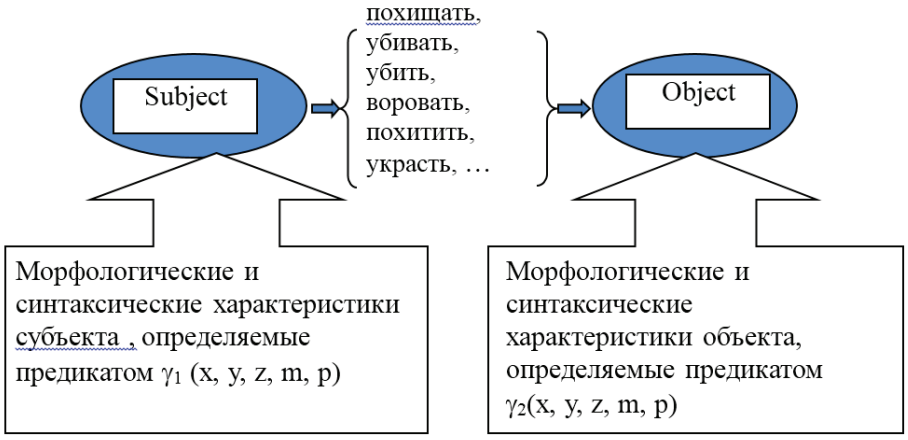

Рис. 1 – Схема идентификации криминально-значимого факта

ния русского языка и их явного представления средствами поверхностной структуры были выделены и описаны предметными переменными морфологические (грамматический падеж) (1) и семантические категории существительных. В работе рассмотрены семантические категории: живой / неживой (2), инструмент, часть тела, объемное пространство, пункт назначения, место отправления, плоскость / точка, механизм, определенное (3).

$$z^{и} \vee z^{р} \vee z^{д} \vee z^{в} \vee z^{т} \vee z^{п} = 1 \quad (1)$$

$$x^{о} \vee x^{н} = 1 \quad (2)$$

$$y^{с} \vee y^{м} \vee y^{и} \vee y^{ч} \vee y^{т} \vee y^{о} \vee y^{в} \vee y^{п} \vee y^{ц} = 1 \quad (3)$$

Тогда предикат, выражающий формальное представление семантико-грамматических признаков, выражающих семантические падежи существительных русского предложения будет выглядеть следующим образом:

$$P(x_n, y_n, z_n) = \gamma_k (x_n, y_n, z_n) \bullet P(x_n) \bullet P(y_n) \bullet P(z_n), \quad (4)$$

где • – операция конъюнкции, а $\gamma_k (x_n, y_n, z_n)$ – предикат, задающий определенный семантический падеж. Например, представление семантического падежа **агенс** через семантико-грамматические признаки

$$\gamma_1(x_n, y_n, z_n) = x_n^{\,o} z_n^{\,u} \vee z_n^{\,u} x_n^{\,n} y_n^{\,м} \vee z_n^{\,u} x_n^{\,o} y_n^{\,c}, \quad (5)$$

В работе [18] для формализации семантических падежей предложений английского языка выделены предметные переменные, описывающие использование предлогов в английских предложениях (6), использование притяжательного падежа (7), позицию существительного в предложении (8), наличие глагола «*to be*» (9) и форму основного глагола:

$$z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{out} = 1 \quad (6)$$

$$y^{ap} \vee y^{aps} \vee y^{out} = 1 \quad (7)$$

$$x^{f} \vee x^{l} \vee x^{kos} = 1 \quad (8)$$

$$m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out} = 1 \quad (9)$$

$$p^{III} \vee p^{ed} \vee p^{I} \vee p^{ing} \vee p^{II} = 1 \quad (10)$$

Тогда, предикат, описывающий семантико-грамматические признаки, партиципатнов английского предложения будет выглядеть следующим образом:

$$P(x, y, z, m, p) = \gamma_k (x, y, z, m, p) \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p), \quad (11)$$

Используя предикаты (10) и (11) можно получить субъект, объект факта, а также его атрибуты: время события, место события, инструмент, цель и т.д.

Выводы

Анализ существующих подходов лингвистического анализа текстов противоправного содержания показывает, что используемые сегодня методы классификации, кластеризации и выделения «лингвистических маркеров» не позволяют эффективно анализировать Веб-контент, содержащий неявно выраженную криминальную информацию. Для анализа такой информации предлагается использовать модель извлечения фактов из слабоструктурированной текстовой информации. Для извлечения факта предложения, представляющего собой триплет субъект-предикат-объект, выделяется глагол, выражающий противоправные действия, и определяются партиципанты – существительные участники данного действия. Для задания смысловых связей между глаголом и партиципантами используются семантические функции глагола, явно выраженные грамматическими и семантическими категориями языка. Модель описывает данные лингвистических категорий для русского и английского языков средствами алгебры конечных предикатов. Полученные предикатные уравнения позволяют определить субъект, объект факта, а так же его атрибуты (время события, место события, инструмент, цель и т.д.) в предложении текста. Следующим этапом исследования является построение данной модели для казахского языка, сложность формализации которого заключается в необходимости описывать семантические функции как грамматическими падежами, так и синтаксическими связями.

ЛИТЕРАТУРА

1. Using Behavioral Indicators to Help Detect Potential Violent Acts. Paul K. Davis, Walter L. Perry, Ryan Andrew Brown, Douglas Yeung, Parisa Roshan, Phoenix Voorhies. – 2013, 258p.
2. Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. doi:10.1080/09546553.2014.849948
3. Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. doi:10.1111/j.0956-7976.2004.00741.x PMID:15447640
4. Europol. (2012). TE-SAT 2012: European Union terrorism situation and trend report. European Law Enforcement Agency.
5. Meloy, J. R. (2011). Approaching and attacking public figures: A contemporary analysis of communications and behaviour. In C. Chauvin (Ed.), *Threatening communications and behaviour: Perspectives on the pursuit of public figures* (pp. 75–101). Washington, DC: The National Academies Press.
6. Meloy, J. R., Hoffmann, J., Guldimann, A., & James, D. (2012). The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, *30*(3), 256–279. doi:10.1002/bsl.999 PMID:22556034
7. Meloy, J. R., Hoffmann, J., Roshdi, K., & Guldimann, A. (2014). Some warning behaviors discriminate between school shooters and other students of concern. *Journal of Threat Assessment and Management*, *1*(3), 203–211. doi:10.1037/tam0000020
8. Meloy, J. R., Mohandie, K., Knoll, J. L., & Hoffmann, J. (2015). The concept of identification in threat assessment. *Behavioral Sciences & the Law*, *33*(2-3), 213–237. doi:10.1002/bsl.2166 PMID:25728417
9. Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: An Overview and Case Studies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003,4p.
10. J. Bollen, A.pepe, and H.Mao. Modelling public mood and emotion: Twitter sentiment and socio-economic phenomena. arxiv:0911.1583v0911 [cs.CY], pages 09-19, 2010.
11. Crime Pattern Detection Using Data Mining. Shyam Varan Nath. Oracle Corporation.- 4p.
12. Bjelopera, J. P. (2013). American jihadist terrorism: Combating a complex threat. CRS Report for Congress. Washington, DC: Congressional Research Service.
13. C McCue, "Using Data Mining to Predict and Prevent Violent Crimes", availableat:http://www.spss.com/dirvideo/richmond.htm?source=dmpage&zone=rtsidebar
14. Учебно-методический комплекс «Современный терроризм: сущность, причины, модели и механизмы противодействия». – Часть 2. – Москва, 2008 г.
15. Bolla, Raja Ashok, "Crime pattern detection using online social media" (2014).*Masters Theses*. 7321.
16. Бондаренко М.Ф. Теория интеллекта : учебник / М.Ф. Бондаренко, Ю.П. Шабанов-Кушнаренко. – Харьков : ООО «СМИТ», 2007. – 576 с.
17. Хайрова Н., Шаронова Н. Логико-лингвистическая модель извлечения фактов из слабоструктурированной текстовой информации // International Journal "Information Models and Analyses" – Varna, Bulgaria. Vol.2, Number 2, 2013. – С. 167-175.
18. Khairova, N.F., Petrasova, S., Gautam, A.P. The logical-linguistic model of fact extraction from English texts . Information and Software Technologies. Volume 639 of the series Communications in Computer and Information Science, Springer, ISBN: 978-3-319-46253-0, 2016,pp. 625-635. doi> 10.1007/978-3-319-46254-7_51