УДК 004.8

МЕТОДЫ АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ НА ОСНОВЕ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Дюсупова Нургуль Болатхановна
Nurgulya_1991@bk.ru
Докторант кафедры Вычислительной Техники Евразийского национального университета им. Л.Н.Гумилева
Научный руководитель – Д.Ж. Сатыбалдина

Социальные сети (такие как Twitter, Facebook, VK, Instagram и т.д.), пользуются все большей популярностью [1]. Одна из самых известных и популярных социальных сетей Facebook - это более 2 миллиардов активных пользователей в 2018 году [2]. Аспект, который следует учитывать в этом контексте, заключается в том, что использование социальных сетей может привести к негативным последствиям и угрозам безопасности пользователей, например, к киберзапугиванию / преследованию в Интернете [1]. Для обнаружения вредоносных программ и пользователей, противодействия распространению нежелательной информации необходим автоматический мониторинг и анализ открытых информационных ресурсов, в том числе семантический анализ комментариев, постов, сайтов, призывающих к межнациональным конфликтам, терроризму, оппозиции. В связи с этим целью настоящей работы является аналитический обзор исследовательских работ, посвященных разработке и использованию программного обеспечения для обнаружения угроз безопасности в текстах на естественном языке на основе методов обработки естественнного языка (Natural Language Processing, NLP).

В [3] представлены результаты иссследований казахстанских ученых, которые впервые создали базу данных экстремистских ключевых слов для казахского языка. Для обнаружения экстремистской направленности в тексте была использована модель, которая состоит из пяти уровней: идентификация сайтов экстремистов, подготовка данных к выгрузке, выгрузка данных, анализ данных и их классификация. Авторы обработали 150 текстов, 80 из которых были с экстремистской направленностью. Для определения часто употребляемых слов был использован метод TF-IDF (term frequency – inverse document frequency). Была создана программа на Visual C#, которая на первом этапе проводит морфологичекий анализ входящего текста, далее ведется поиск базовых слов в подготовленной базе данных и результаты поиска выводятся на интерфейс пользователя. Для продолжения исследований авторы планируют классифицировать слова с помощью таких методов как: Naïve Bayes, Random forest, Logistic regression и Support Vector Machine.

Авторы работ [4], [5] исследуют твиты про ИГИЛ, которые были опубликованы в социальной сети Twitter. Исследователи в начале собирают базу данных с твиттера, связанные с ИГИЛ, затем очищают от шума, таких как, слова с ошибками, странные символы и т.п. и выполняют предварительную обработку данных. Во время семантического анализа данных автор столкнулся с проблемой, ему необходимо было отсортировать твиты, которые были про ИГИЛ и отличить твиты с пропагандой ИГИЛ. Чтобы решить проблему он использовал список учетных записей пользователей, которые были связаны с известными сторонниками террористов. Для классификации твитов автор использовал биграммы и буквенные биграммы. Для классификации направленности твитов как позитивный, негативный и нейтральный он использует для эксперимента три алгоритма: Naïve Bayes, Ada Boost, SVM.

В работе [6] представлены результаты исследования с помощью анализа настроений твитов в социальной сети Twitter. Он классифицировал твиты как отрицательный, положительный и нейтральный. Это основано на направленности слов, которое определяется

с помощью словарей как WordNet, DAL. Исходя из значений словарей к каждому слову присваивается оценка. Этим определяется направленность твита.

Для обнаружения и классификации внешних кибератак в реальном времени изучены и сравнены результаты четырех типов классификаторов машинного обучения: K-Nearest Neighbor Classifier, Naive Bayes classifier model, Random Forest Classifier, Logistic Regression [7]. Для проверки производительности модели использована перекрестная проверка 10 K-Fold. (Таблица 1)

Классификатор	Оценка модели (данные обучения)	Проверка модели (данные испытаний)	Средний балл
K-Nearest Neighbor Classifier	0.9993	0.9990	0.9991
Naive Bayes classifier model	0.6712	0.6697	0.6684
Random Forest Classifier	0.9998	0.9994	0.9994
Logistic Regression	0.9585	0.9585	0.9585

Таблица 1. Оценка моделей классификаторов по данным обучения и испытаний

Как видно на этой таблице, самый эффективный и надежный результат выдал классификатор Random Forest Classifier. Поэтому автор для решения проблемы высокопроизводительной IDS будет использовать данный классификатор.

Методы, адаптированные для анализа данных, использованы авторами работы [7]. С использованием веб-сервиса «Sentiment140» было извлечено для обработки 1,6 млн твитов и пользователей, которые затем были проанализированы с помощью Natural Language ToolKit для определения настроения твитов. Изначально набор данных «Sentiment140» имел три уровня настроений: положительный, отрицательны и нейтральный. Далее все предложения были разбиты на слова и все ненужные слова были удалены. Применены алгоритмы машинного обучения для обнаружения кибератак. Наиболее хорошие результаты 99.7% показал алгоритм Tree Decision. (Рисунок 1)

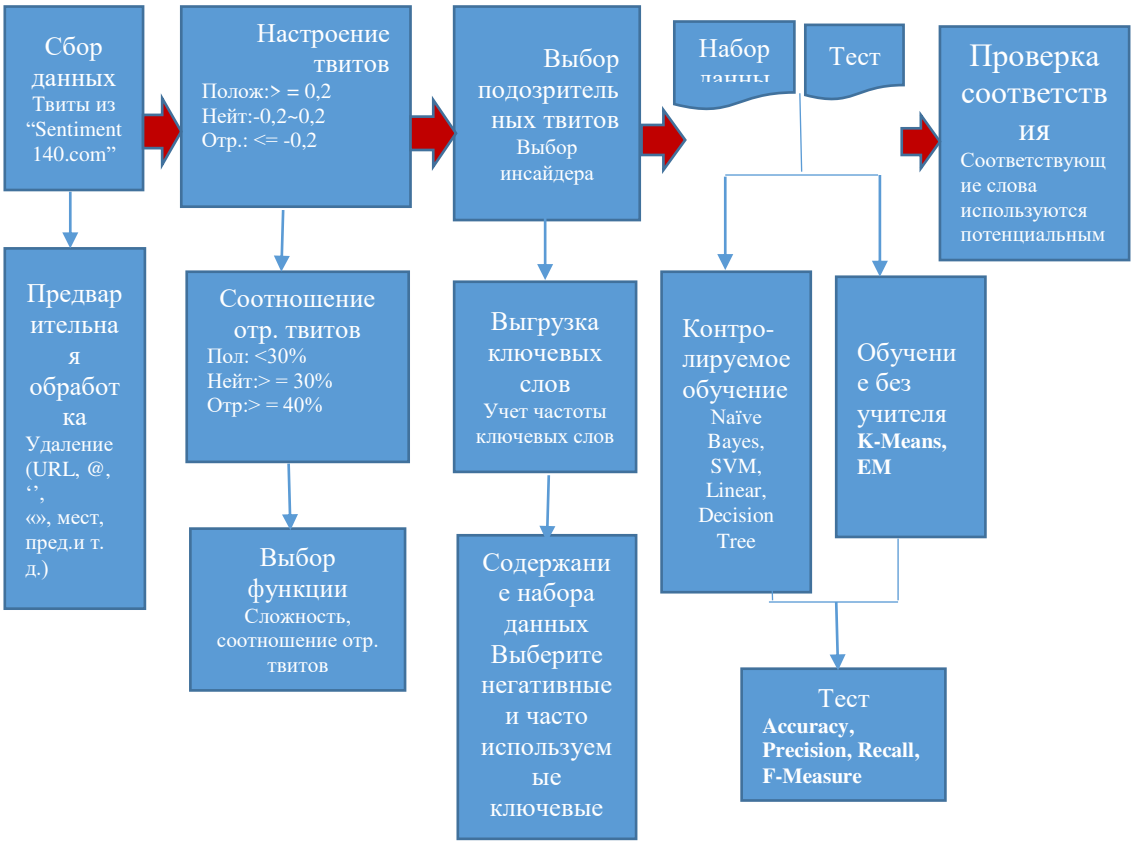

Рисунок 1. Общий процесс мониторинга и анализа

Методам обнаружения киберугроз на английском и голландских языках в социальных сетях посвящена работа [8]. Для мониторинга и анализа была выбрана социальная сеть Askfm. Данные социальной сети состоят из вопросов и ответов на них, которые опубликованы в профиле пользователя. Выгрузка данных получена с помощью программного обеспечения GNU Wget путем сканирования профилей, из которых были удалены слова не подлежащие английскому и голландскому языку. Набор данных составил 78 387 и 113698 постов на двух языках, соответственно. Для автоматического обнаружения кибератак были проведены эксперименты по методу SVM, реализованной на LIBLINEAR для бинарной классификации. В процессе предварительной обработки данных, использованы PoS-метки, токенизация и лемматизация данных (Lets Preprocess Toolkit), а так же очистку данных (удаление и фильтрация лишних пробелов, хэштегов, слов с ошибками). Для классификации слов по настроению, то есть положительные, отрицательные и нейтральные для голландских лемм были использованы лексикон Duoman и Pattern, а для английских лемм - AFINN и MSOL. Для английского и голландского языка в словари Linguistic Inquiry и Word Count добавлена частота всех 68 психометрических категорий. Это привело к появлению 871 296 и 795 072 функций для обоих языков.

Авторы работы [9] так же для своих экспериментов выбрали платформу Twitter для арабского языка. Для больших наборов данных они использовали функцию PCA и SGPLVM, чтобы сравнить результаты. Для классификации наборов данных была использована функция K-средних. Исходный набор данных классифицируется на семь классов насилия: преступность, насилие, нарушение прав человека, политические взгляды, кризис, несчастные случаи и конфликты. Существует дополнительный класс «другой», который содержит ненасильственные твиты, где упоминались некоторые слова насилия. (Таблица 2)

Таблица 2. Детали набора данных для анализа твитов на арабском языке

Класс	Обучение	Тестирование	Всего	%
Насилие	5673	2759	9332	57.5
Ненасилие	4790	2112	6902	42.5
Всего	11363	4871	16234	

Арабский язык имеет сложную морфологическую структуру и диалекты. Поэтому они применили инструмент MADIMARA для извлечения некоторых морфологических признаков. Частотность определена с помощью функции tf-idf. Результаты эксперимента можете увидеть ниже. (Таблица 3)

Таблица 3. Результаты эксперимента для анализа твитов на арабском языке [12]

Функция				
Модель	Dim	P	R	F
K-means (в данных)	14,621	0,46	0,65	0,54
PCA+K-means	11,000	0,47	0,66	0,55
PCA+K-means	8000	0.46	0.63	0.54
SGPLVMx+K-means	8000	0.56	0.60	0.58
Функция Токена				
Модель	Dim	P	R	F
K-means (в данных)	44,163	0,50	0,75	0,60

PCA+K-means	35,000	0,56	0,98	0,71
PCA+K-means	8000	0,49	0,72	0,58
SGPLVMx+K-means	8000	0,58	0,55	0,56

В настоящей работе рассмотрено небольшое, но растущее число решений для обнаружения аномалий в поведении пользователей и выявления их негативной направленности в социальных сетях. Несмотря на то, что эти подходы различаются, все основаны, главным образом, на методах обработки естественного языка. Именно эта основа дифференцирует обнаружение аномалий в социальных сетях по морфологическим и синтаксическим особенностям того или иного языка. Сравнение результатов праткического использования методов (Naïve Bayes, Random forest, Logistic regression, Support Vector Machine, TF-IDF, Ada Boost, K-Nearest Neighbor Classifier, Tree Decision), позволило сделать выводы об их применимости к данному классу задач, выявить оптимальные алгоритмы и подходы для будущих исследований, связанных с анализом контентного содержания социальных сетей на казахском языке.

Список использованных источников

1. M. Rybnicek, R. Poisel and S. Tjoa, "Facebook Watchdog: A Research Agenda for Detecting Online Grooming and Bullying Activities," *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, 2013, pp. 2854-2859. doi: 10.1109/SMC.2013.487

2. https://www.statista.com

3. Bolatbek M. A., Mussiraliyeva Sh. Zh., Tukeyev U.A., «Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language», Вестник КазНТУ им. Аль-Фараби 1(97) 2018: стр.134-141.

4. Enghin Omer, «Using machine learning to identify jihadist messages on Twitter», Master degree thesis, Upsala University (2015).

5. M. Ashcroft, A. Fisher, L. Kaati, E. Omer and N. Prucha, "Detecting Jihadist Messages on Twitter," *2015 European Intelligence and Security Informatics Conference*, Manchester, 2015, pp. 161-164. doi: 10.1109/EISIC.2015.27

6. Ilia Vovsha, Owen Rambow, Rebecca Passonneau, Apoorv Agarwal, Boyi Xie, «Sentiment analysis of twitter data», Proceedings of the Workshop on Language in Social Media (LSM 2011), Portland, Oregon, 23 June 2011. Pp. 30–38.

7. Won Park, Youngin You, Kyungho Lee, «Detecting Potential Insider Threat: Analyzing Insiders' Sentiment Exposed in Social Media», Hindawi Security and Communication Networks. Volume 2018, Article ID 7243296, 8 pages, https://doi.org/10.1155/2018/7243296

8. Van Hee C, Jacobs G, Emmery C, Desmet B, Lefever E, Verhoeven B, et al. (2018) Automatic detection of cyberbullying in social media text. PLoS ONE 13(10): e0203794. https://doi.org/10.1371/journal.pone.0203794

9. Kareem E Abdelfatah, Gabriel Terejanu, Ayman A Alhelbawy, «Unsupervised detection of violent content in Arabic social media», University of South Carolina, University of Essex, Fayoum University, pp. 01– 07, 2017. DOI: 10.5121/csit.2017.70401