Якушенкова Ариадна Дмитриевна
магистрант
Институт вычислительной математики
и информационных технологий
ФГАОУ ВПО «Казанский (Приволжский)
федеральный университет»
г. Казань, Республика Татарстан
Самсонов Антон Андреевич
магистрант
Институт вычислительной математики
и информационных технологий
ФГАОУ ВПО «Казанский (Приволжский)
федеральный университет»
г. Казань, Республика Татарстан
Ситдикова Фарида Бизянова
канд. филол. наук, старший преподаватель
Институт международных отношений,
истории и востоковедения
ФГАОУ ВПО «Казанский (Приволжский)
федеральный университет»
г. Казань, Республика Татарстан

ПРИМЕНЕНИЕ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА ДЛЯ БОРЬБЫ С КИБЕРПРЕСТУПНОСТЬЮ

***Аннотация**: в данной статье рассматривается проблема обеспечения безопасности пользователей в Интернете. Для решения обозначенной проблемы авторами предлагается применение анализа тональности текста. В работе рассматривается автоматический способ идентификации людей, склонных к педофилии, на основе компьютерного анализа текстовых сообщений в социальных сетях.*

***Ключевые слова**: анализ тональности текста, анализ текстовых сообщений, социальные сети, киберпреступность.*

В нашей статье описывается применение анализа тональности текста для решения такой актуальной проблемы, как обеспечение безопасности пользователей в Интернете. Среди различных видов киберпреступности одним из самых опасных является педофилия. Цель статьи – рассмотрение автоматического способа идентификации людей, склонных к педофилии, на основе компьютерного анализа текстовых сообщений в социальных сетях.

Начнем с основных понятий и терминов. *Анализ тональности текста* (англ. sentiment analysis) – область компьютерной лингвистики, которая занимается выделением из текстов эмоционально окрашенной лексики или эмоциональной оценки автора. Анализ тональности представляет собой текстовую классификацию, т. е. процесс присвоения естественно-язычным текстам тематической категории из определенного набора. Под *мнением* (*тональностью*) понимают выраженное в тексте эмоциональное отношение некоторого субъекта к определенному объекту

Проблема автоматического распознавания мнений в тексте оказалась предметом активных исследований за рубежом сравнительно недавно – в 2000-х гг. В России таких работ до последнего времени было крайне мало; только в 2012 году оценка тональности текста была выбрана одной из главных тем конференции по компьютерной лингвистике «Диалог-2012» [1].

Область применения анализа тональности текста обширна и поистине огромна. Суть метода заключается в изучении эмоциональной окраски текста, используемой в тексте по отношению к объекту исследования. Несмотря на то, что тон является лишь одним из признаков мнения, проблема классификации тональности является наиболее важной и исследуемой учеными всего мира. Наиболее простым способом является анализ текста в одномерном эмотивном пространстве, т. е. в пространстве из двух эмоций: хорошо или плохо. Используя словарь, собранный на основе анализа речи, можно выделить в исследуемом тек-

сте слова, относящиеся к той или иной категории эмоций, а затем на основе количества употребляемых слов и их эмоциональной окраске сделать определенные выводы о наклонностях исследуемого лица.

Основной проблемой метода анализа текста является составление словарей (в дальнейшем будем называть их тезаурусами), это связано со сложностями составления их и большим объемом полученного тезауруса. Вторая проблема заключается в содержании словаря. Разделение контента на хорошее и плохое в проблеме педофилов является серьезной проблемой, потому что разговор может идти на совершенно невинные темы. Мы можем проанализировать много источников, но некоторые фразы и слова будут звучать безобидно, хотя их реальное значение будет противоестественно и незаконно. Эксперты, анализирующие тексты вручную, больше подходят для решения этой проблемы, но они проводят больше времени, чем машина, при анализе одного сообщения. Поэтому важной задачей является автоматизация создания и улучшения качества тезаурусов оценочных слов и выражений. Другой проблемой является регулярное обновление словаря. Каждый день педофилы создают новые шифровки и сокращения, чтобы общаться и скрываться от полиции. Так что нужна регулярно обновляемая и компетентная база данных слов. В настоящее время для многих языков создаются такого рода ресурсы оценочной лексики [2; 3]. Для русского языка опубликован и свободно доступен лишь один автоматический порожденный словарь оценочных слов и выражений ProductSentiRus [4].

В настоящее время существует несколько методов для анализа тона, но для данной задачи метод теоретико-графовых моделей является наиболее подходящим. Этот метод основан на предположении, что каждое слово в тексте представляется в виде вершины графа, а связь, соединяющая два слова на семантическом уровне, считается ребром графа. Каждое слово имеет определенный вес и различное влияние на общий тон текста. Анализ текста в таких условиях можно разделить на несколько этапов: построение графика по тексту, обозначение всех вершин, классификация найденных слоев (присвоение каждой вершине своего веса, положительного или отрицательного, на основе используемого тезауруса)

и расчет окончательного тона текста. Оценка текста рассчитывается по формуле T = P / N, где T – общая оценка текста, P – положительная оценка компонента текста, N – оценка негативной составляющей. Для того, чтобы получить оценку P, необходимо суммировать все положительные веса графа, для оценки N – получить сумму отрицательных весов графа [5]. Этот метод хорош для анализа коротких сообщений и переписки между педофилом и ребенком.

На момент написания данной статьи, российские ученые совместно с полицией Бельгии и Нидерландов создали работоспособную версию программы, анализируя отдельные сообщения. Эта программа уже дала некоторые результаты, которые показывают, что примерно 70% от анализа текстовых данных дает правильный результат, который можно использовать на практике. Во-первых, исследователи вводят некоторую коллекцию текстов и указывают их метаданные (подробности, теги). После этого автоматически строится объект, атрибутивно описывающий эти данные, который позволяет аналитику в интерактивном режиме визуализировать данные и сделать необходимые выводы. Полученное программное обеспечение было протестировано эмпирически. Для сравнения, исследование показало, что эксперты корректно оценивают эмоциональную окраску текста вручную примерно в 79% случаев [6]. Как мы видим, программа не уступает ручным исследованиям с точки зрения точности оценки и в ближайшем будущем таких алгоритмов будут использоваться все больше и больше во всех сферах общественной жизни, но теперь для большей эффективности должны быть использованы модераторы, чтобы помочь программным комплексам составлять словари и совершать предварительный анализ сообщений. При совместной работе экспертов и программ анализа текстов будут достигаться наилучшие результаты.

Краткие выводы. В статье отмечается актуальность исследований по оценке тональности текста в настоящее время как за рубежом, так и в нашей стране. Статья рассматривает применение анализа текста для борьбы с педофилией. Описан метод теоретико-графовых моделей, который используется для анализа тональности текста сообщений потенциального преступника. Были рассмотрены

практические результаты работы аналогичных программ и выявлена точность работы данных программ в сравнении с работой, выполненной вручную экспертами анализа эмоциональной окраски текста. Улучшить полученные результаты можно за счет изменения алгоритма наполнения словаря либо за счет ручной корректировки словаря экспертами, однако, уже сейчас можно сказать, что данный метод является актуальным, современным и подходящим для нашей конкретной задачи.

Список литературы

1. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference «Dialogue» (2012).

2. Computational Linguistics and Intellectual Technologies [Электронный ресурс]. – Режим доступа: http://www.spsl.nsc.ru/FullText/konfe/Dialog'2012-Vol.2.pdf

3. Steinberger J. Creating Sentiment Dictionaries via Triangulation. Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT-2011 / J. Steinberger. – Oregon, 2011. – P. 28–36.

4. Mihalcea R. Learning multilingual subjective language via cross-lingual projections. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics ACL-2007 / R. Mihalcea, C. Banea, J. Wiebe. – Prague, Czech Republic, 2007. – P. 976–983.

5. Herman I. Graph visualization and navigation in information visualization: a survey / I. Herman, G. Melançon, M.S. Marshall // IEEE Trans. on Visualization and Computer Graphics. – 2000. – Vol. 6.

6. Ogneva M. How Companies Can Use Sentiment Analysis to Improve Their Business. / M. Ogneva. – Mashable, 2012).

7. Bing Liu. Sentiment Analysis and Subjectivity / N. Indurkhya, F.J. Damerau // Handbook of Natural Language Processing. – 2010.