- Research Methodology*]. Moscow: FLINTA: Nauka.
- Solganik, G.Ya. (1996). *Stilistika russkogo yazyka [Stylistics of the Russian language]*. Moscow.
- Xu, H. (2018). *The Application of Multi-modal Discourse in English Translation of Tujia Folk Song Long Chuan Diao in Western Hubei Province*. Retrieved from: <https://doi.org/10.1051/mateconf/201823202010>
- Zarubin, S.F., & Rozheckin, A.M. (1988). *Bol'shoj russko-yaponskij slovar' [Large Russian-Japanese dictionary]*. Moscow: Rus. yaz.

УДК 81.33

<https://doi.org/10.25076/vpl.37.03>

А.О. Кутняя

О.И. Максименко

Московский государственный областной университет

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ЭМОЦИИ «РАДОСТЬ»

Извлечение эмоций является одним из актуальных направлений автоматической обработки текста. В настоящее время проблема выделения эмоций из текста полностью не решена. В статье рассматриваются основные проблемы автоматического извлечения эмоций, связанные с языковыми особенностями, рассмотрены различные методы, используемые для их решения. В статье описан эксперимент по анализу работы системы автоматического сентимент-анализа текста Indico для русского и английского языков на примере автоматического извлечения из текста эмоции «радость». Тестирование проводилось на базе параллельных корпусов русско- и англоязычных текстов на политическую тематику. Был проведен сравнительный анализ результатов для обоих языков и выборочный ручной анализ текстов, выявлены основные ошибки в работе системы. В соответствии с результатами были сделаны выводы о недостаточно высоком качестве автоматического сентимент-анализа. В заключении были рассмотрены возможные причины ошибок системы.

Ключевые слова: обработка естественного языка, анализ

тональности, извлечение эмоций, эмоции, субъективность.

UDC 81.33

<https://doi.org/10.25076/vpl.37.03>

A.O. Kutnyaya

O.I. Maksimenko

Moscow Region State University

AUTOMATIC EXTRACTION OF EMOTION “JOY”

Emotion mining is one of the relevant areas of natural language processing. Nowadays the problem of emotion extraction from text is not completely solved. The paper deals with the main problems of automatic emotion extraction related to language features and the methods to solve them are considered. The article describes an experiment of analysis of the system for automatic sentiment analysis Indico for Russian and English languages on the example of automatic extraction of the emotion "joy". Testing was carried out on the parallel corpus of Russian and English political texts. A comparative analysis of results for both languages and selective manual analysis of the texts was conducted, and the main errors were described. According to the results, it can be concluded that the quality of the automatic sentiment analysis is not high enough. In conclusion, the possible causes of the system's errors were considered.

Key words: natural language processing, sentiment analysis, emotion mining, emotions, subjectivity.

Введение

В настоящее время стремительно развиваются новые области компьютерной лингвистики. Для выполнения определенных задач появляются и совершенствуются новые методы, позволяющие автоматизировать некоторые процессы, обычно совершаемые вручную. Одной из активно развивающихся областей является анализ тональности текста или сентимент-анализ. Сентимент-анализ нацелен на выявление в тексте эмоциональной окрашенности и субъективного мнения. На данный момент активно разрабатываются методы автоматического анализа текста, однако работа систем все еще нуждается в улучшении. Улучшение

автоматических методов анализа позволит ускорить работу и сделать ее более эффективной в будущем.

Материалом исследования были тексты, взятые из Цифровой библиотеки ООН, все тексты были проанализированы при помощи системы для автоматической обработки текста Indico.

Автоматический анализ тональности или сентимент-анализ – это одна из актуальных в настоящее время областей компьютерной лингвистики, развитию которой сейчас уделяется большое внимание. Это связано с быстрым темпом развития информационного общества и непрекращающимся ростом количества информации, что, в свою очередь, ставит перед современной лингвистикой задачу обработки больших массивов текстов. В этом случае ручной обработки текстов экспертами может стать недостаточно в силу больших затрат, в том числе временных ресурсов.

Невозможно представить себе коммуникацию без выражения эмоций, личного отношения говорящего к каким-либо объектам высказывания. Для выражения различных эмоций используется набор специфических лексических средств, и, в свою очередь, средства выражения эмоций становятся объектом исследования лингвистики. Предпринимаются попытки классификации эмоциональной лексики (Маслечкина, 2015). Поэтому, когда мы говорим об автоматическом анализе текста, мы рассматриваем не только факт выявления в тексте какой-либо объективной информации, но и эмотивной. Эмоциональный компонент также становится предметом изучения и подвергается классификации и формализации. В этом случае исследователи прибегают к анализу тональности текста.

В целом, автоматический сентимент-анализ можно определить как автоматическое выявление в тексте эмоционально окрашенной лексики и отношения говорящего к какому-либо объекту.

Этот тип анализа текста оперирует такими понятиями, как *opinion* (мнение), *sentiment* (настроение) и субъективность; соответственно, процедуры также могут быть известны под названиями *opinion mining*, *sentiment analysis* и *subjectivity analysis* (Pang, Lee, 2008). Также стоит понимать главное отличие этого типа анализа от поиска фактов (*fact mining*). В нашем случае критерии «объективности» высказанного мнения, «истинности»

или «ложности» сообщаемых фактов не принимаются во внимание. Языковая ситуация рассматривается только с позиции носителя высказывания. Таким образом, интерес для анализа в данной области представляет выраженное в тексте мнение и эмоциональное содержание высказываний.

Говоря о тональности текста, следует выделять три параметра: субъект тональности (автора текста), тональную оценку и объект тональности (предмет, о котором высказывается мнение).

Сентимент-анализ находит широкое применение в современном мире. Это касается теоретических исследований в области компьютерной лингвистики, а также ряда практических задач, связанных с жизнью информационного общества. Методы автоматической обработки текста используются в различных областях: в качестве примера можно привести применение подобных алгоритмов коммерческими организациями для сбора информации о мнениях клиентов для дальнейшего улучшения работы. Сентимент-анализ также может использоваться для мониторинга социальных медиа (Bandi, Fellah, 2019). Алгоритмы для анализа используются в социологических исследованиях, в различных областях дискурсивного анализа, и, в частности, могут использоваться в анализе политического дискурса, что также является актуальным в настоящее время. Тестирование программы автоматического анализа, выполненное в рамках работы, проведено на корпусе текстов, относящихся к политическому дискурсу.

Исследование эмотивного компонента высказывания ставит перед нами ряд меньших подзадач. Так, может возникнуть необходимость как характеристики тональности как положительной, отрицательной или нейтральной, так и вычленения из текста отдельных эмоций.

В данной работе рассматривается концептуализация эмоции «радость». Радость – это одна из базовых эмоций, названных К. Изардом (наряду с печалью, страхом, удивлением, гневом). Радость – это активная положительная эмоция, чувство внутреннего удовлетворения, психологического комфорта и благополучия, часто сопровождающееся приливом энергии и силы. Это единственная из перечисленных нами эмоция, которой может быть приписана четкая положительная оценка, в отличие от

удивления, которому может быть приписан знак «плюс» или «минус» в зависимости от контекста ситуации.

Радость связана с возможностью удовлетворить ту или иную актуальную потребность. Однако, как далее отмечает К. Изард, радость стоит отделять от чувственного (сенсорного) удовольствия (Изард, 2012). Положительный тон возникает на разных уровнях эмоциональной сферы. Радость сопровождается чувством удовлетворения собой и окружающим миром и возможным повышением альтруизма, то есть желанием делиться положительными эмоциями с окружающими.

Мы можем заключить, что данная эмоция – одна из наиболее простых и интуитивно понятных для человека, однако разработка алгоритма для выявления этого компонента в тексте может стать нетривиальной задачей. Задача лингвистической концептуализации той или иной эмоции состоит в том, чтобы выявить и систематизировать языковые средства, используемые носителем высказывания для ее выражения. В случае необходимости формализовать концепт, который большинству людей, говорящих на том или ином языке, понятен интуитивно, мы неизменно сталкиваемся с трудностями.

Сегодня существует много различных систем для автоматического анализа тональности текста. Они используют разные методы для решения проблемы концептуализации эмоций.

Особенности и проблемы сентимент-анализа

Сентимент-анализ в своей сути схож с другими методами анализа текста, которые, как правило, сводятся к категоризации документов.

Существует множество возможных категорий, определения которых могут зависеть от пользователя и приложения, и для конкретной задачи мы можем иметь дело с двумя классами (бинарная классификация) или тысячами классов (документы, относящиеся к сложной таксономии). В отличие от этого, в классификации мнения часто имеется относительно мало классов. Как и было сказано ранее, основная задача анализа настроения – это или бинарное разделение текстов на тексты положительной и отрицательной эмоциональной окраски, или выделение в тексте проявлений конкретных эмоций (Pang, Lee, Vaithyanathan, 2002). Тем не менее, есть и более сложные способы репрезентации

разных характеристик, например, показатели силы проявления какой-либо эмоции, степень позитивности или негативности высказывания.

Как уже было сказано ранее, то, что интуитивно понятно при восприятии текста, может создать определенные трудности при формализации.

Некоторыми проблемами при концептуализации эмоций в тексте являются:

- Количество объектов оценки.

Системы автоматического анализа часто используются при обработке таких текстов, как отзывы и рецензии, и способны показывать высокие результаты. Это объясняется тем, что в данных текстах очень легко выделить параметры объекта высказывания и отношения носителя высказывания к объекту. Такие тексты содержат одну конкретную оценку заранее известного нам объекта. В то же время более сложные по структуре тексты становятся трудны для обработки. Это наблюдается в том числе в исследованиях новостных и политических текстов. В текстах со сложной структурой может быть озвучено сразу несколько точек зрения по поводу разных объектов, что усложняет задачу: перед тем, как произвести оценку тональности, необходимо выделить все эти сущности из текста (Большакова, Клышинский и др., 2017).

Частным случаем этой проблемы является проблема референтности. Большое количество объектов, упоминаемых в тексте, также означает возможность большого количества референций, а значит, становится сложнее понять, какая отсылка относится к какой сущности. Сущности могут находиться в кореферентных отношениях, то есть отношениях между компонентами высказывания, которые обозначают один и тот же внеязыковой объект или ситуацию. Для решения этой проблемы необходимо обучить систему правильно находить связи между сущностями.

- Длина текста

Длина текста неразрывно связана с предыдущей проблемой. Может показаться очевидным, что более короткие тексты дадут меньше информации, поэтому их анализ станет менее точен; однако, чем длиннее текст, тем выше вероятность, что носитель

высказывания затронет несколько тем и объектов. Это актуально в том числе для рассматриваемых нами текстов политических выступлений. В дипломатическом дискурсе может быть важно озвучить разные точки зрения на проблему, что приводит к увеличению высказанных оценок. Также на конечный результат может повлиять цитирование.

- **Имплицитность**

Можно отметить, что значительная часть проблем проистекает из-за не эксплицитного, а имплицитного выражения эмоций в тексте. Имплицитное высказывание – это объективное высказывание, из которого следует некое оценочное суждение. Объективная информация, содержащаяся в таких высказываниях, высоко контекстуальна, и поэтому имплицитные высказывания тяжело поддаются формализации. Здесь сентимент-анализ пересекается со способами извлечением фактов из текста.

- **Контекстуальная эмоциональная окраска**

Некоторые средства выражения эмоций проявляют себя в контексте. Так, от окружения будет зависеть, имеет данное слово или конструкция положительную или отрицательную коннотацию (Большакова, Клышинский и др., 2017, с. 132).

- **Ирония**

Ирония и сарказм часто используются нами в коммуникации для придания эмоциональной окраски. Ирония подразумевает, что истинное мнение говорящего противоположно высказанному. Сарказм – это более грубая форма иронии, но по своей форме эти речевые приемы схожи. Существует множество исследований, нацеленных на то, чтобы научить компьютерный алгоритм понимать иронию, но в настоящий момент решить эту проблему крайне сложно. Есть возможность обучить систему на размеченном корпусе с высказываниями, содержащими иронию и сарказм, но индивидуальность речи каждого человека и многообразие средств выразительности показывает, что все еще нельзя выработать один рабочий алгоритм.

- **Слова-модификаторы**

Такие слова, как частицы отрицания, а также слова, отражающие степень (очень, слишком, немного) способны модифицировать слово, к которому они относятся, вплоть до изменения его тональности на противоположную.

- Ирреальность

Некоторые объекты оценки говорящего могут быть ирреальны и не иметь отношения к основному объекту оценки, например, когда речь идет о какой-либо нереальной ситуации. Соответственно, стоит разделять реальные и нереальные объекты. Как правило, языки располагают соответствующим набором маркеров, например, условные обороты, частицы *ли*, *бы* в русском языке, что может облегчить эту задачу.

При разработке методов анализа необходимо учитывать эти трудности.

Алгоритмов для решения задач извлечения мнений существует довольно много. Их можно разделить на две больших группы.

- Метод, основанный на правилах (rule-based)

Это самый простой метод анализа текста. Он оперирует двумя главными элементами: словарем и набором правил.

Эмоциональная окраска в тексте, как правило, выражена лексически. Поэтому в методах сентимент-анализа применяются различные словари. Существуют общие словари и словари для конкретных предметных областей, что позволяет улучшить работу с контекстом. Словарь выглядит как список слов с присвоенными каждому из них оценочными характеристиками (Посевкин, Бессмертный, 2015). Характеристики могут быть разнообразными: положительная или отрицательная эмоциональная окраска, сила тональности, конкретные категории, такие как категории эмоций. Могут использоваться готовые словари оценочной лексики, а также такие словари могут быть составлены на основе размеченного вручную корпуса текста какой-либо предметной области.

Далее вычисления проводятся с учетом правил. Правила имеют вид условий (Клековкина, Котельников, 2012). Они применяются для частичного разрешения указанных выше проблем, например, проблемы слов-модификаторов. Так, в правилах учитываются маркеры, усиливающие эмоциональную окраску (очень) и меняющие ее на противоположную (не).

Дальнейшее развитие словарного метода приводит к разработке новых моделей. Например, разрабатываются методы автоматического порождения словарей и извлечения оценочной лексики для определенной предметной области (Лукашевич,

Четверкин, 2013). Каждая предметная область может иметь свой набор характерной только для нее оценочной лексики, которая может иметь другой вес в текстах другой тематики или вообще будет являться немаркированной. Таким образом, более узкая специализация может улучшить качество анализа. В случае, если мы рассматриваем множество высказываний, принадлежащих одному субъекту, становится возможно выделить личные особенности речи носителя высказывания.

- Машинное обучение

Второй способ, который в перспективе может улучшить качество и упростить обработку больших массивов текста – это машинное обучение. Машинное обучение с учителем (supervised learning) подразумевает, что алгоритм обучается на предварительно размеченной коллекции текстов. Существует много способов представления текста и методов классификации. Текст может быть представлен в виде набор n-грамм (униграммов, биграммов). Может использоваться метод мешка слов (bag-of-words). Далее эти сведения передаются на обработку алгоритму классификации. Среди используемых методов можно выделить байесовский классификатор, метод опорных векторов (SVM), логистическая регрессия. Также используются методы для подсчета веса признака (tf-idf), методы расчета семантической близости для лексики (Araque, Zhu, Iglesias, 2019).

Машинное обучение без учителя (unsupervised learning) основано на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции (Клековкина, Котельников, 2012). На основе этого алгоритм расставляет веса для слов. Для правильной реконструкции связей между словами используются методы, основанные на графовых моделях, а также вероятностные модели (Xiea, Xiab, Lia, Lauc, 2017).

Машинное обучение является эффективным методом, несмотря на некоторые его проблемы. Так, для эффективной работы обучения с учителем требуется разметка очень больших массивов текста. Также системы, обученные на конкретном наборе текстов, могут оказаться хороши только в своей предметной области. Вместе с исследованиями в области машинного обучения

разрабатываются новые модели, основанные на сверточных нейронных сетях (Zhang, Li, Song, 2019).

Кроме того, используются гибридные методы. Так, при машинном обучении могут быть введены некоторые правила, корректирующие работу алгоритма.

Результатом анализа тональности является разделение документов на конечное число классов. Это могут быть бинарные категории (позитивные или негативные высказывания) или категории для каждой эмоции, а также могут быть представлены коэффициенты, показывающие интенсивность той или иной эмоциональной окраски. Нами была поставлена задача выделения из текста такой эмоции, как радость.

При выделении эмоций из текста мы сталкиваемся с определенными трудностями. Некоторые особенности нашей речи, которые без труда могут считаться человеком, тяжело подвергаются формализации. Такими факторами являются, например, большое количество объектов, упоминаемых в тексте, и сложная система референции, из-за чего становится сложно установить связь между объектом и высказыванием. Также большую сложность представляет факт, что многие эмоции выражаются в тексте не эксплицитно, а имплицитно; говорящие могут прибегать к таким средствам, как ирония и сарказм, сравнения, а также большую роль в определении эмоциональной окраски может сыграть контекстуальная информация, модифицирующая и трансформирующая высказывание. Также роль в интерпретации текста могут сыграть специфичная тематика текста и личные особенности речи носителя высказывания.

Нами также были рассмотрены основные методы, помогающие справляться с этими проблемами. Отчасти метод, основанный на правилах, помогает решить проблему, например, со словами-модификаторами. Однако более эффективным, хоть и более сложным в реализации, является метод машинного обучения на коллекции текстов, и на данный момент исследования анализа текста продвигаются в этом направлении.

Ход эксперимента

В данной работе рассматривается извлечение концепта «радость» из текста при помощи работы автоматического алгоритма.

Ход работы можно представить в виде последовательности действий:

1. Собран корпус текстов на русском и английском языках, состоящий из текстов выступлений советского дипломата Олега Александровича Трояновского.

2. Русские и английские тексты были по отдельности обработаны модулем для автоматического сентимент-анализа. Обработка производилась при помощи системы Indico.

3. Полученные результаты для тональности текста и для содержания в тексте эмоций были занесены в таблицу.

4. Были сделаны выводы о работе системы путем сопоставления результатов автоматического анализа между собой и с результатами выборочного ручного анализа.

Общие результаты

Результат анализа текста состоит из двух основных аспектов: позитивной/негативной окраски и показателей вероятности того, что в тексте выражена эмоция: радость, удивление, злость, печаль, страх (joy, surprise, anger, sadness, fear). Показатель позитивной/негативной окраски высказывания представляет собой число в интервале от 0 до 1, где нижняя граница – крайне отрицательное высказывание, верхняя – крайне положительное.

Общие результаты для анализа, включающие в себя средние показатели доли эмоции «радость» и общей тональности высказывания представлены в таблице 1.

В общем и целом, можно сразу отметить, что система лучше работает для английского языка, чем для русского, что легко объяснить тем, что в первую очередь она разрабатывалась и тестировалась на английском языке, в то время как на русских текстах она была менее натренирована и располагала меньшим количеством данных.

	Средний показатель тональности	Средний показатель радости (joy)
Русский язык	0,429	0,493
Английский язык	0,645	0,169

Таблица 1

При сравнении можно увидеть, что показатель тональности для английского языка несколько выше, чем для русского, однако разрыв между ними намного меньше, чем между показателями проявления радости в текстах. По причине разницы между результатами для русского и английского было принято решение рассмотреть их отдельно.

Анализ обработки русского языка

Можно сразу отметить, что практически для всех текстов система поставила «радость» на первое место в списке эмоций, определив ее как наиболее выраженную. 49 из 50 документов имеют самый высокий показатель радости относительно других эмоций. Это, разумеется, далеко не всегда согласуется с реальной эмоциональной окраской высказываний.

Показатели для радости колеблются в пределах между 0,737 и 0,376, преимущественно стремясь к среднему результату.

Так как «радость» фигурирует как главная эмоция практически во всех документах, сравнительный анализ вызывает некоторые затруднения. Для примера можно рассмотреть документ 46, имеющий наиболее высокий показатель радости (0,737).

Рассматривая его с лексической точки зрения, мы наблюдаем ряд слов и конструкций, обладающих негативной эмоциональной окраской, среди которых: «конфликты», «жертвы», «напряженность», «отрицательная реакция», «агрессия», «угроза».

В противовес этому, содержатся и положительно окрашенная лексика, например: «содействие», «поддержка», «добрые услуги», «удовлетворение», «положительный», «мирный». Эта лексика может использоваться для выражения радости.

Однако необходимо рассматривать документ целиком и учитывать контекст. При рассмотрении документа можно ясно видеть, что радостный тон не является в нем преобладающим. Также лексический состав сам по себе не должен являться единственным фактором, который влияет на результаты разбора. Например, рассмотрим контекст упомянутой нами положительной лексики: *«Очевидно, что правительство Великобритании не встало бы на путь применения вооруженной силы в целях сохранения своего колониального господства... если бы не содействие и прямая военная поддержка со стороны Соединенных Штатов Америки».*

Здесь мы видим положительно окрашенные слова «содействие», «поддержка», однако ясно, что общий тон высказывания не выражает радости говорящего, а эти слова в контексте приобретают отрицательную коннотацию. Конструкцию усложняют отрицания и условность предложения, а также контекстуальная информация («применение вооруженной силы», «колонияльное господство» - понятия с негативной коннотацией в данном сообщении).

Очевидно предположить, что для русского языка система менее обучена, но сложно однозначно сказать, почему система выделяет радость как преобладающую эмоцию почти во всех текстах. Возможно, признаки с положительной окраской имеют в системе оценивания больший вес. Проблема с восприятием контекста может частично объяснить такую работу системы.

Также даже показатели для отдельных слов могут быть контекстуальны. Например, часто встречающееся в документах, в том числе в данном документе, слово «мир» может быть в ряде случаев воспринято как «общество» или как «согласие, отсутствие конфликтов». Однако, если система рассматривает только вариант, которому стоит приписать более высокую позитивную коннотацию, то это влияет и на общий показатель для радости.

В выборке оказался всего один документ (20), приписанный которому показатель «радость» стоит не на первом месте. При рассмотрении его можно действительно отметить, что радость в нем не выражена ни эксплицитно, ни имплицитно, и лексика, использованная в нем, преимущественно несет негативный оттенок («агрессивные действия», «возмущение», «негодование», «осуждение», «жертвы агрессии», «ущемление прав») и сильно окрашена эмоционально.

Этот документ достаточно короткий, если сравнить с другими, что, возможно, улучшило эффективность работы системы, так как в тексте нет противоречащих друг другу оценок, и есть одна яркая негативная окраска. Однако можно отметить, что даже при выраженной негативной окраске сообщения и наборе лексики оценка для радости составляет ~3,8, а для преобладающей эмоции «страх» ~3,9, то есть разрыв между этими оценками очень невелик. Вероятно, признаки, позволяющие классифицировать тон сообщения как радостный, и в этой ситуации имеют слишком

большой вес для системы.

Нами был рассмотрен еще ряд документов с различными показателями радости.

Документ 27, в котором показатель для радости равен 0,653, в действительности совмещает в себе несколько видов эмоций. В первой половине сообщения преобладают негативные эмоции («с сожалением», «ущерб», «угроза», «напряженность»), однако количество ярко эмоционально окрашенных слов невелико. Вторая половина сообщения имеет более радостную эмоциональную окраску. Говорящий использует такие понятия, как «мирное, справедливое решение», «позитивный отклик», «укрепление мира», «поддержка (проекта)». Конструкции, в которых встречается данная лексика, действительно не имеют других коннотаций и не усложнены отрицаниями и другими модификаторами. По этой причине можно понять, почему система отнесла данный текст к «радостным».

Однако есть и тексты, присутствие в которых компонента «радости» по подсчетам системы нельзя объяснить. Например, рассматривая документ 9 с показателем радости 0,504, то есть выше среднего, мы можем, во-первых, с уверенностью сказать, что данное сообщение имеет ярко выраженную негативную окраску и не имеет в себе эксплицитного или имплицитного выражения радости. Рассматривая лексический состав, мы можем подтвердить это. Говорящим используется такая лексика, как: «агрессия», «гневное возмущение», «смерть и разрушения», «устрашить», «уничтожить», «катастрофические последствия», «конфликт», «угроза».

Лексика, использованная говорящим, не только негативная, но и ярко эмоционально окрашенная. Кроме этого, используются метафорические выражения («пламя военного конфликта») для того, чтобы дополнительно усилить эмоциональную окраску. Экспрессивные обороты в целом характерны для речи автора высказывания. По сравнению с негативными элементами, лексика, имеющая какие-то положительные коннотации, крайне незначительна. Слова «мир», «справедливость» имеют положительные коннотации, но в контексте сообщения их значение и эмоциональная окраска несравнимо малы, и никак нельзя сказать, что какая-либо положительная лексика здесь

может рассматриваться как средство выражения радости.

Также есть ряд сообщений, в которых четко прослеживаются элементы, выражающие радость говорящего. На примере документа 4 можно рассмотреть много средств, передающих эту эмоцию. Говорящий использует по отношению к адресатам высказывания такие конструкции, как: «я хотел бы сердечно приветствовать вас», «надежда», «отношения сотрудничества», «позитивное развитие», «искренняя признательность», «уважаемый представитель», «искусное руководство», «дипломатический талант», «высокие дипломатические качества», «установление справедливого и прочного мира». Эти слова и конструкции сами по себе или в контексте (например, здесь положительное в контексте «сотрудничество», которое в иных случаях может быть рассмотрено как негативное, как было в примере выше с «поддержкой» и «содействием»), передают радостную эмоцию говорящего. Здесь также присутствуют модификаторы, усиливающие эмоциональную окраску («искренняя признательность»), употребляются и эксплицитное, и имплицитное выражение эмоций.

В тексте также есть и средства выражения других эмоций, однако их меньше, чем вышеперечисленных, и их общий тон стремится к более нейтральному. Несмотря на это, системой показатель радости отмечен как 0,399, что даже несколько ниже среднего, как и ниже показателя для предыдущего документа, в котором выражение радости отсутствует как таковое.

Анализ обработки английского языка

В то же время, результаты для английского языка показали, что эмоция «радость» превалирует в 6 документах из 50. Кроме того, что это намного меньше результатов для русского языка, показатели для радости также колеблются между ~0,063 и ~0,55, стремясь к среднему показателю, который заметно ниже результатов для русского.

Несмотря на то, что, как уже было сказано, на первый взгляд система заметно лучше работает на англоязычном материале, нежели на русскоязычном, нельзя назвать результаты анализа точными. В то время как система относит к «радостным» большее количество русских текстов, часто совсем неоправданно, для английского «ложных срабатываний» заметно меньше, что можно

наблюдать в таблице. Но сам по себе этот факт не может быть абсолютным критерием успешной работы системы.

Можно провести сравнительный анализ одних документов для русского и английского языка.

0,55 – это самый высокий результат для радости в английском языке, однако стоит отметить, что он определен у уже рассмотренного нами документа 9, не имеющего признаков выражения данной эмоции, и напротив, обладающего средствами для экспрессивного выражения эмоций гнева и страха, что наблюдается и в английском ('aggression', 'conflict', 'death and destruction' и др.) Тем не менее, и анализ для русского, и анализ для английского оба приписали ему высокую оценку по нашему параметру.

Следующий документ (10), имеющий второй по высоте показатель вероятности 0,435, в действительности имеет некоторые конструкции, передающие радость говорящего; в частности, это лексика и конструкции, относящиеся непосредственно к адресату.

Употреблены конструкции: '*I should like most sincerely to congratulate you*', '*willingness to co-operate with you*' '*to pay a due tribute*'. Также используются конструкции, которые имплицитно выражают положительное отношение говорящего к адресату: '*(who discharged his duties) in an extremely professional manner*'.

Однако нельзя сказать, что в общем тоне документа радость превалирует. Возможно, это связано с тем, что хоть дальнейшее сообщение и не выражает радость говорящего, негативные лексические средства и конструкции в нем не обладают высокой экспрессивностью.

Также можно снова обратиться к документу 4, приведенному нами в качестве примера сообщения с большим количеством конструкций, выражающих радость, и увидеть, что анализ для английского языка не распознал или не придал достаточный вес всем этим конструкциям, так как показатель для радости составил всего 0,148. Все это несмотря на употребленные конструкции: '*I should like warmly to congratulate you*' , '*positively advanced*', '*diplomatic talents*' '*sincere gratitude*', '*skillful leadership*', '*great diplomatic qualities*', которые уже были рассмотрены в русском варианте текста. В этих конструкциях ярко прослеживаются и

положительно окрашенная экспрессивная лексика, и положительные модификаторы.

Документы 13, 40, также получившие самые высокие показатели для радости, не имеют таковой ярко выраженной окраски, кроме незначительного количества позитивной лексики, которую, очевидно, перевешивает негативная.

В ручном анализе документа 13 мы снова сталкиваемся с проблемой контекстуальной информации. Несколько раз употребленное в сообщении слово 'peace' несет, очевидно, положительную коннотацию. Однако этого нельзя сказать о конструкциях с этим словом. Это можно увидеть на следующих примерах употребления: *'an act designed to undermine the sovereignty and territorial integrity of a peace-loving, independent Member State'*, где общее высказывание имеет негативную окраску; *'A great deal was said at that meeting about the need to enhance the effectiveness of this principal organ of the United Nations for the maintenance of international peace and security'*, где общий тон высказывания достаточно нейтральный. То же самое можно сказать о других лексических средствах ('support' в *'unconditional support of ... extreme manifestations of the adventurist policy'*). Можно заметить и имплицитно выраженное неодобрение говорящего относительно объектов высказывания.

Количество конструкций, выражающих радость, в тексте также крайне мало (*'I ... wish you success'*, *'The Soviet Union ... expresses its solidarity with and support'*), и экспрессивная положительная лексика в тексте не присутствует. Напротив, как вне контекста, так и в употребленных конструкциях очень ярко заметна негативная лексика (*'profound indignation'*, *'aggression'*, неоднократно употребленные в разных окружениях *'terror'* и *'terrorism'*).

Подобная ситуация наблюдается и со следующим упомянутым документом. Количество положительной лексики невелико (*'a reliable, trusty friend and ally'*), набор негативной представлен более широко, кроме того, на протяжении текста лексические средства склонны повторяться в различных окружениях, хотя относительно общей длины документа нельзя сказать, что текст и слова в нем обладают высокой экспрессивностью.

Выводы

Нами была протестирована работа системы для

автоматического анализа тональности текста на примере собранного корпуса текстов политических выступлений на русском и английском языках. Нашей задачей было рассмотреть то, как система автоматически вычленяет в тексте эмоцию «радость».

По итогу работы можно заметить большую разницу в результатах, выданных анализом для русского и английского языков. Если для средних значений тональности текста разница не так велика, то средние значения для радости/joy сильно расходятся – 0,493 для русского и 0,169 для английского.

Анализатор для русского языка выделил эмоцию «радость» как преобладающую в 49 случаях из 50, и этот вывод оказался справедлив далеко не для всех документов, иногда напрямую противореча их настоящей эмоциональной окраске. В определенных случаях можно предположить, что анализатор придал больше веса тем или иным лексическим средствам, выражающим радость говорящего, или придал меньше веса средствам, выражающим другие эмоции. В других случаях доля положительной лексики в текстах, отнесенных к «радостным», была минимальна, зато были заметны другие эмоции, которым анализатор приписал показатели ниже.

Анализатор для английского языка показал, на первый взгляд, более приближенные к реальности показатели, однако детальное рассмотрение привело к выводу, что и здесь его работу нельзя назвать качественной. Так, некоторые проявления радости были проигнорированы анализатором, и напротив, документы с негативной окраской были отнесены им в эту категорию. Поэтому, несмотря на то, что количество «ложных срабатываний» меньше, чем в анализе для русского, у анализа имеются схожие проблемы.

В целом, можно соотнести рассмотренные нами в теоретической части исследования проблемы анализа тональности и те ошибки, которые были обнаружены в результате работы системы. Их возможно классифицировать, выделив основные.

Наиболее частые проблемы были связаны со следующими факторами.

- Контекстуальная информация. В целом для анализа будет справедливо сказать, что необходимо рассматривать не отдельные лексические единицы, а целые конструкции. В рассмотренных

нами текстах и на русском, и на английском языке есть лексика, проявляющая себя в контексте. В приведенных примерах была рассмотрена положительная лексика, помещенная в контекст, в котором она меняет свою полярность. Для правильного восстановления семантики предложения необходимо рассмотреть всю конструкцию целиком, восстановить отношения между лексическими единицами и сравнить между собой веса различных лексических единиц.

- Модификаторы. Наличие модификаторов является частным случаем указанной выше проблемы контекста.

- Имплицитность. В ряде высказываний автор имплицитно выражает свое отношение к ситуации. Подкрепляя свою точку зрения и пытаясь воздействовать на адресатов высказывания, говорящий сообщает объективные факты, иллюстрирующие общий посыл его сообщения. Эта проблема касается не просто анализа средств выразительности, но полного анализа семантики текста.

- Длина текста. Многие рассмотренные нами речи были достаточно длинными, что действительно могло создать препятствия в работе системы. Проблемой является не столько сама длина текста, сколько количество высказанных в нем точек зрения и объектов референции.

- Цитирование. В ходе изучения текстов нами было отмечено, что автор прибегает к цитированию или озвучиванию иной точки зрения. При этом объектом цитирования может стать другой говорящий, имеющий точку зрения, противоположную носителю высказывания. Это перекликается с проблемой имплицитности, так как при ручном анализе тех же данных будет понятно, какую позицию занимает автор относительно тех или иных высказанных точек зрения. Для системы это не всегда представляется возможным. Поэтому цитируемый материал и его оценка также сказывается на общей эмоциональной оценке высказываний.

- Следующую проблему нельзя обозначить с абсолютной уверенностью. Однако, можно предположить, что большому количеству ошибок способствует расстановка весов у тех или иных признаков. Это может относиться к выявленной нами проблеме с русскоязычными документами, когда большинство из них были неоправданно отмечены как преимущественно

«радостные» по эмоциональной окраске. Такая проблема, на первый взгляд менее выраженно, присутствовала и в оценке англоязычных документов. Возможно предположить, что признаки, маркирующие эмоцию «радость», имеют в системе слишком большой вес относительно остальных. Таким образом, лексика, имеющая положительную коннотацию, воспринимается системой как более важная для общей оценки, даже если это неоправданно в контексте речи. При этом, однако, сложно выявить корреляцию между общей эмоциональной окраской текста и показателем радости в том случае, когда первый ниже среднего, а показатель для радости выше среднего.

- Другие проблемы, которые не удалось идентифицировать. Некоторые документы, которые система отнесла к тому или иному классу, при ручном анализе абсолютно не подходят под нужные критерии, и установить причину такой классификации не представляется возможным. Это характерно для некоторых текстов на русском языке. Возможно, мы имеем дело с совокупностью каких-то из проблем выше или техническими погрешностями в работе системы.

Способы научить искусственный интеллект распознавать эмоции совершенствуются, но все еще существует ряд проблем, которые мы рассмотрели в теоретической части работы. К наиболее частым проблемам относятся различные имплицитные средства выражения эмоций, такие как ирония, сарказм, сравнения, а также информация, которую можно понять только из контекста, и которая может изменить весь эмоциональный тон сообщения. К другим проблемам относятся проблема референции и высказывания различных точек зрения и оценок в одном сообщении, а также слова-модификаторы.

Мы оценили работу одного из таких анализаторов на практике. Нашей задачей было провести автоматический анализ корпуса текстов политических выступлений на русском и английском языках при помощи системы Indico для выделения из текста эмоции «радость». Нами было рассмотрено то, насколько система справляется со своей задачей и насколько показанные результаты соответствуют реальности.

Мы рассмотрели ошибки, допущенные системой. Некоторые примеры, приведенные нами в анализе результатов, показывают

уже рассмотренные нами общие проблемы извлечения эмоций. Так, возникали проблемы с текстами, в которых были высказаны разные эмоции, из-за чего становится сложнее установить общий тон сообщения. Также возникали проблемы с информацией, выраженной говорящим имплицитно. Эмоциональная окраска некоторых слов может меняться в зависимости от контекста и приобретать другие коннотации, однако это знание недоступно для системы. Можно также предположить, что ошибки происходят из-за неверного распределения весов для тех или иных параметров. Однако есть документы, результаты анализа которых сложно объяснить однозначно.

Все это приводит к тому, что компонент «радость» или приписывается сообщениям, в которых его нет или он очень мал, и игнорируется в других. Это справедливо и для русского, и для английского языков, несмотря на то что в числовых показателях есть большие расхождения.

Литература

1. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и анализ данных. – М.: Изд-во НИУ ВШЭ, 2017.
2. Изард К.Э. Психология эмоций. – СПб.: Питер, 2012.
3. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г. – С. 81-86. – URL <http://ceur-ws.org/Vol-934/paper15.pdf> (дата обращения: 10.11.2019).
4. Лукашевич Н. В., Четверкин И. И. Построение модели для извлечения оценочной лексики в различных предметных областях // Моделирование и анализ информационных систем. – 2013. – Т. 20. – № 2. – С. 70–79.
5. Маслечкина С.В. Выражение эмоций в языке и речи [Электронный ресурс] // Вестник БГУ. – 2015. – №. 3. – Режим доступа: <https://cyberleninka.ru/article/n/vyrazhenie-emotsiy-v-yazyke-i-rechi> (дата обращения: 10.11.2019).

6. Объединенные Нации. Цифровая библиотека. – [Электронный ресурс]. – Режим доступа: <https://digitallibrary.un.org/?ln=ru> (дата обращения: 19.10.2019)
7. Посевкин Р.В., Бессмертный И.А. Применение sentiment-анализа текстов для оценки общественного мнения // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – №1. – [Электронный ресурс]. – Режим доступа: <https://cyberleninka.ru/article/n/primenenie-sentiment-analiza-tekstov-dlya-otsenki-obschestvennogo-mneniya> (дата обращения: 10.11.2019).
8. Araque O., Zhu G., Iglesias C. A. A semantic similarity-based perspective of affect lexicons for sentiment analysis // Knowledge-Based Systems, 2019. Т. 165. – Р. 346–359. – [Электронный ресурс]. – Режим доступа: <https://www.sciencedirect.com/science/article/pii/S0950705118305095>, (дата обращения: 10.11.2019).
9. Bandi, A., Fellah, A. Socio-Analyzer: A Sentiment Analysis Using Social Media Data // EPiC Series in Computing. – Режим доступа: <https://easychair.org/publications/open/HJnm> (дата обращения: 10.11.2019).
10. Modeling Sentiment and Aspect Using Syntax: A Topic Model Approach / R. Xiea, M. Xiab, C. Lia, R.Y.K. Lauc // Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing, 2017. – Рз. 191–196. – [Электронный ресурс]. – Режим доступа: <https://pdfs.semanticscholar.org/12ef/2699787de36060e7f63afb802c676174459f.pdf>, (дата обращения: 10.11.2019).
11. Pang B., Lee L., Vaithyanathan S. Thumbs up? // Sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Volume 10. – Association for Computational Linguistics, 2002. – Pp. 79–86.
12. Pang, B., Lee, L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. – Boston: now Publishers Inc, 2008. – Т. 2. – No. 1–2. – Pp. 1–135
13. Zhang C., Li Q., Song D. Syntax-Aware Aspect-Level Sentiment Classification with Proximity-Weighted Convolution Network // Proceedings of the 42nd International ACM SIGIR

Conference on Research and Development in Information Retrieval. ACM, 2019. – Pp. 1145–1148. – [Электронный ресурс]. – Режим доступа: <https://arxiv.org/pdf/1909.10171>, (дата обращения: 10.11.2019).

Refereneces

- Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165, 346–359.
- Bandi, A., & Fellah, A. (2019). Socio-Analyzer: A Sentiment Analysis Using Social Media Data. *EPiC Series in Computing*, 64, 61–67
- Bolshakova, E.I., Klyshinsky, E.S., Lande, D.V., Noskov, A.A., Peskova, O.V., & Yagunova, E.V. (2017). *Avtomaticeskaya obrabotka tekstov na estestvennom yazyke i analiz dannyh [Automatic Text Processing in Natural Language and Data Analysis]*. The Higher School of Economics Publishing House.
- Chetviorkin, I.I., & Loukachevitch, N.V. (2013). Postroenie modeli dlya izvlecheniya ocenочноj leksiki v razlichnyh predmetnyh oblastyah [Construction of a Model for the Cross-Domain Opinion Word Extraction]. *Modeling and Analysis of Information Systems*, 20(2), 70–79.
- Izard, C.E. (2012). *Psihologiya emocij [The Psychology of Emotions]*. St. Petersburg: Piter Publ.
- Klekovkina, M.V., & Kotelnikov, E.V. (2012). Metod avtomaticheskoy klassifikacii tekstov po tonal'nosti, osnovannyj na slovare emocional'noj leksiki [The method of automatic classification of texts by tonality, based on the dictionary of emotional vocabulary]. In *Selected Papers of XVI All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections"*, (pp. 81-86).
- Maslechkina, S.V. (2015). Vyrazhenie emocij v yazyke i rechi [Expression of the Emotion in Language and Speech] *Vestnik Bryanskogo gosudarstvennogo universiteta*, (3), 231-236.
- Pang, B., & Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (pp. 79–86).
- Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis.

- Foundations and Trends in Information Retrieval*, 2(1-2).
- Posevkin, R.V., & Bessmertnyj, I.A. (2015). Primenenie sentiment-analiza tekstov dlya ocenki obshchestvennogo mneniya [Texts sentiment-analysis application for public opinion assessment]. *"Scientific and Technical Journal of Information Technologies, Mechanics and Optics"*, 15 (1), 169-171.
- United Nations Digital Library System. (2020). Retrieved from <https://digitallibrary.un.org/?ln=ru>
- Xiea, R., Xiab, M., Lia, C., & Lauc, R.Y.K. (2017). Modeling Sentiment and Aspect Using Syntax: A Topic Model Approach. In *Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing*, (pp. 191–196).
- Zhang, C., & Li, Q., & Song, D. (2019). Syntax-Aware Aspect-Level Sentiment Classification with Proximity-Weighted Convolution Network. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 1145–1148).

УДК 811.131

<https://doi.org/10.25076/vpl.37.04>

Соколова А.Г.

Хучбарова Д.М.

Национальный исследовательский Московский
государственный строительный университет
Военная академия Ракетных войск стратегического
назначения имени Петра Великого

ВЗАИМОДЕЙСТВИЕ НЕМЕЦКИХ И ИТАЛЬЯНСКИХ РЕАЛИЙ НА ТЕРРИТОРИИ ШВЕЙЦАРИИ НА ПРИМЕРЕ РОМАНА АЛЬБЕРТО НЕССИ «ВСЕ ВЫХОДЯТ»

Данная работа посвящена исследованию контакта языков и культур на территории Швейцарии на примере романа итальянского писателя Альберто Несси «Все выходят» (Несси, 1989). Авторами данной работы была предпринята попытка проанализировать взаимодействие немецких и итальянских реалий на территории Швейцарии на примере данного романа.