Подход к формированию обучающей выборки для оценки эмоциональной окраски постов социальной сети с применением машинного обучения

А.А. Константинов[1]

[1]Ульяновский государственный технический университет, Северный венец 32, Ульяновск, Россия, 432027

Аннотация. В данной статье описывается система определения эмоциональной окраски постов социальной сети. В статье подробно описан разработанный алгоритм формирования обучающей выборки, а так же реализация программной системы и эксперименты.

1. Введение

Исследование социальных сетей с каждым годом приобретает все большую актуальность в связи с обостряющейся необходимостью обеспечения безопасности населения и мониторинга общественных настроений. Анализ сообщений и постов может помочь оценить изменения в настроениях многих пользователей и найти применение в политических и социальных исследованиях, в том числе и в исследованиях потребительских предпочтений.

В настоящее время нейронные сети применяются для решения различного рода задач в области интеллектуальной обработки данных. Развертывание нейронной сети проводится в два этапа:

- выбор архитектуры нейронной сети;
- формирование обучающей выборки.

Этап подготовки обучающей выборки занимает большое количество времени. Во многих случаях обучающая выборка формируется экспертом в ручном режиме, следовательно, эксперт тратит много времени, чтобы проанализировать данные и сформировать обучающую выборку.

Цель данной работы – разработка экспериментального образца программной системы определения эмоциональной окраски постов социальной сети на основе авторских символов выражения эмоций.

Основными задачами являются:

- анализ предметной области, включающий определение исходных данных для формирования обучающей выборки и определение классов эмоциональной окраски постов;
- обзор существующих решений и исследований, которые были предложены российскими исследователями;
- разработка методики формирования обучающей выборки, основанной на методах лингвистического анализа текстовой информации;
- программная реализация;

- проведение экспериментов, отражающих эффективность определения эмоциональной окраски поста нейронной сетью, обученной сформированной обучающей выборкой.

2. Аналоги

В настоящее время в работах российских исследователей предлагаются способы формирования обучающих выборок.

Первый метод формирования обучающих выборок описан в работе [1]. Суть метода состоит в том, чтобы свести объём обучающей выборки к минимуму. Но в тоже время обучающая выборка должна в полной мере описывать поведение модели. Для минимизации количества экспериментальных данных при обучении нейросети предлагается синтезировать недостающие обучающие пары из предварительно построенной математической модели.

В работе [2] описано несколько методов формирования обучающей выборки. Первым способом является программная генерация, суть которой состоит в том, чтобы в процессе формирования выборки варьировать как можно больше параметров.

Второй способ – сэмплирование, суть которого в том, чтобы задать распределение в пространстве объектов. Данный метод применяется, чтобы исследовать не все данные, а только осмысленные части.

Следующим методом является закономерная модификация базового объекта. Обучающее множество получается путём модификации параметров.

Четвёртым примером является выборка из базы объектов. Суть состоит в том, чтобы сгруппировать объекты по группам, соответственно, объекты некоторой группы будут ближе расположены друг к другу, а из разных групп – дальше.

В работе [3] описывается исследовательский прототип анализатора тональности текста, который реализует процесс обработки, состоящий из следующих этапов: на первом этапе текст разбивается на отдельные предложения, предложения – на отдельные слова. На втором этапе производятся морфологический анализ каждого слова, лемматизация и определение частей речи. Перечисленные этапы анализа предложений необходимы для точного сопоставления найденных слов тональному словарю. Используются тональные словари для русскоязычного текста объемом порядка 35000 слов. В словаре каждому слову соответствует тональная оценка. Такой показатель представляет собой набор из пяти значений. Каждое значение определяет степень принадлежности слова к одному из классов: крайне отрицательный, отрицательный, нейтральный, положительный, крайне положительный.

Так же были рассмотрены программные системы и модули, выполняющие сентимент-анализ текстов. Модуль SentiFinder [4] определяет три вида тональности русскоязычных текстов (позитивную, негативную и нейтральную) относительно заданного объекта тональности как в пределах одного предложения, так и усредненную по всему документу. Средняя точность по трем видам тональности около 87%.

Существует ряд тезаурусов, специально размеченных с учётом эмоциональной составляющей. Такие словари, описанные далее, необходимы компьютерным программам при анализе тональности текста. WordNet-Affect — это семантический тезаурус, в котором понятия, связанные с эмоциями, представлены с помощью слов, обладающих эмоциональной составляющей [5]. Также в WordNet-Affect используются дополнительные эмоциональные метки для того, чтобы разделять синсеты в соответствии с их эмоциональной валентностью. Для этого определяются четыре дополнительные эмоциональные метки: позитивная, негативная, неоднозначная и нейтральная.

SentiWordNet — это лексический семантический тезаурус, первая версия которого была разработана в 2006 году [6]. Данная система является результатом процесса автоматического аннотирования каждого набора синонимов в соответствии с его степенью позитивности, негативности и объективности. Использование SentiWordNet дает более чем 20 % прирост точности по сравнению с первой версией

SenticNet представляет собой еще один семантический тезаурус для работы с наборами эмоциональных понятий [7]. SenticNet применяется для проектирования интеллектуальных приложений, предназначенных для анализа эмоциональной составляющей текста. Главным

назначением SenticNet является упрощение процедуры машинного распознавания концептуальной и эмоциональной информации, передаваемой с помощью естественного языка. Если сравнить другие лексические тезаурусы, такие как SentiWordNet и WordNet-Affect с SenticNet, то их главным различием будет то, что SentiWordNet и WordNet-Affect обеспечивают связывание слов и эмоциональных понятий на синтаксическом уровне, не позволяя выявлять смысловую составляющую.

В рассмотренных научных работах описываются лишь общие рекомендации для формирования обучающей выборки, но не приводится методик или алгоритмов, которые позволили бы сформировать качественную обучающую выборку для сентимент-анализа в автоматизированном режиме. Знания, накопленные при изучении исследований, могут быть использованы при выполнении данной работы.

3. Модели и алгоритмы

Наиболее часто используемым методом, с помощью которого формируется обучающая выборка, является отбор по ключевым словам и фразам. При использовании данного метода используются словари авторских символов выражения эмоций и словари ключевых фраз.

Словари авторских символов выражения эмоций были составлены экспертным путем. Каждый словарь составлен для определённой эмоции и содержит несколько авторских символов выражения эмоций. Словари ключевых фраз были найдены в сети интернет и дополнены путём анализа постов социальной сети.

На первом этапе выполняется отбор постов на основе словарей авторских символов выражения эмоций. В качестве входной информации берутся 2,5 млн. постов из базы данных. Если пост содержит авторский символ выражения эмоций, то он относится к конкретному классу и добавляется в соответствующий список.

На втором этапе выполняется отбор постов на основе словарей ключевых фраз. В качестве входной информации берутся списки, полученные на предыдущем этапе. На данном этапе выполняется лемматизация каждого слова поста. Затем пост проверяется на содержание каждого слова из словаря. Если пост содержит фразу, значит он принадлежит к конкретному классу эмоциональной окраски. На выходе данные записываются в текстовые файлы, каждый из которых содержит обучающую выборку, относящуюся к конкретному классу эмоциональной окраски.

Нейронная сеть работает только с числами, поэтому тексты необходимо представить в численном виде. Для представления обучающей выборки в виде векторов использовался алгоритм word2vec [8]. Первоначально составляется список всех встречающихся в постах слов, предварительно лемматизированных. Затем создаются вектора, размер которых равен размеру списка всех слов. После в векторе ставится 1, если слово встречается в посте, иначе 0, если отсутствует.

В качестве нейронной сети был использован многослойный персептрон с тремя слоями. Количество нейронов в первом слое – размер списка всех слов словаря. Количество нейронов во втором слое – размер первого, делённый на 50. Размер второго слоя был подобран путём проведения множества экспериментов. Для словаря размером 2000 слов размер второго слоя будет 400 нейронов. Количество нейронов третьего слоя равно трём, так как нам нужно определять семь эмоций.

После обучения нейронной сети на вход подаётся тестовая выборка, каждый пост которой также векторизуется на основе словаря, полученного при обучении нейронной сети.

3.1. Формальное описание системы

Формально процесс отбора постов можно представить блок-схемой, показанной на рисунке 1. Блок-схема описывает процесс отбора постов для формирования обучающей выборки. Каждый этап отбора, представленный в блок схеме, содержит в себе процессы отбора постов для каждой конкретной эмоции.

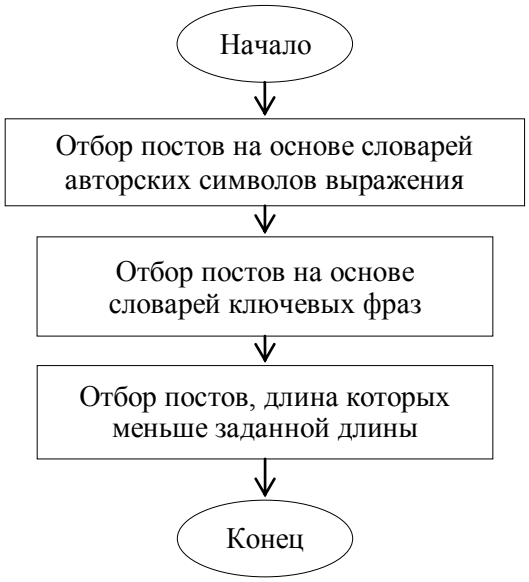

Рисунок 1. Процесс отбора постов.

На первом этапе отбираются посты на основе словарей авторских символов выражения эмоций для каждого класса эмоциональной окраски текста, затем, на втором этапе, происходит отбор постов на основе словарей ключевых фраз и на третьем этапе происходит отбор постов, длина которых меньше заданной длины. Ограничение длины было введено, из-за того, что обучение нейронной сети на больших постах понижает точность распознавания эмоциональной окраски текста.

Формально множество словарей, по которым происходит отбор постов можно представить формулой (1)

$$D = \{D^E, D^W\} \quad (1)$$

где D^E – множество словарей с авторскими символами выражения эмоций, D^W – множество словарей с ключевыми словами и фразами.

В свою очередь множество словарей с авторскими символами выражения эмоций можно представить формулой (2)

$$D^E = \{D^E_{joy}, D^E_{sad}, D^E_{surp}, D^E_{anger}, D^E_{disg}, D^E_{cont}, D^E_{fear}\} \quad (2)$$

где D^E_{joy} – словарь с эмоцией «радость», D^E_{sad} – словарь с эмоцией «грусть», D^E_{surp} – словарь с эмоцией «удивление», D^E_{anger} – словарь с эмоцией «злость», D^E_{disg} – словарь с эмоцией «отвращение», D^E_{cont} – словарь с эмоцией «презрение», D^E_{fear} – словарь с эмоцией «страх».

В свою очередь множество словарей с ключевыми словами можно представить формулой (3)

$$D^W = \{D^W_{joy}, D^W_{sad}, D^W_{surp}, D^W_{anger}, D^W_{disg}, D^W_{cont}, D^W_{fear}\} \quad (3)$$

где D^W_{joy} – словарь с эмоцией «радость», D^W_{sad} – словарь с эмоцией «грусть», D^W_{surp} – словарь с эмоцией «удивление», D^W_{anger} – словарь с эмоцией «злость», D^W_{disg} – словарь с эмоцией «отвращение», D^W_{cont} – словарь с эмоцией «презрение», D^W_{fear} – словарь с эмоцией «страх».

Каждому процессу отбора постов конкретной эмоции ставится в соответствие словарь с авторскими символами выражения эмоций D^E и словарь ключевых фраз D^W.

Процесс проверки обучающей выборки можно представить блок-схемой, показанной на рисунке 2.

На первом этапе формируется набор векторов с помощью алгоритма word2vec, затем происходит обучение нейронной сети и после этого происходит оценка точности определения эмоциональной окраски текста с помощью тестовой выборки.

4. Программная реализация

Для оценки эффективности разработанного подхода к формированию обучающей выборки была реализована программная система.

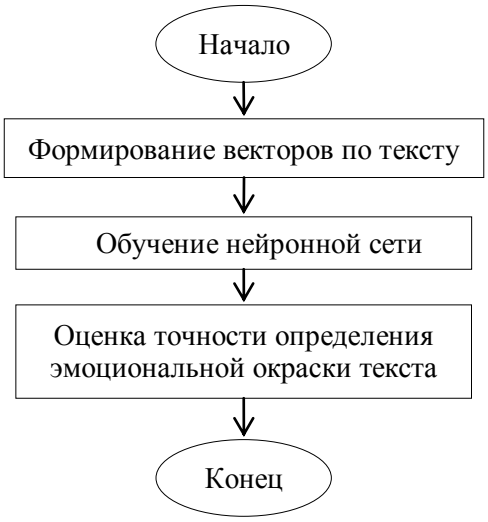

Рисунок 2. Процесс проверки обучающей выборки.

Система выполняет чтение данных из БД, чтение словарей с авторскими символами выражения эмоций и ключевыми словами для каждой эмоции, лемматизацию, формирование обучающей выборки и обучение нейронной сети. Диаграмма последовательности разработанной системы представлена на рисунке 3.

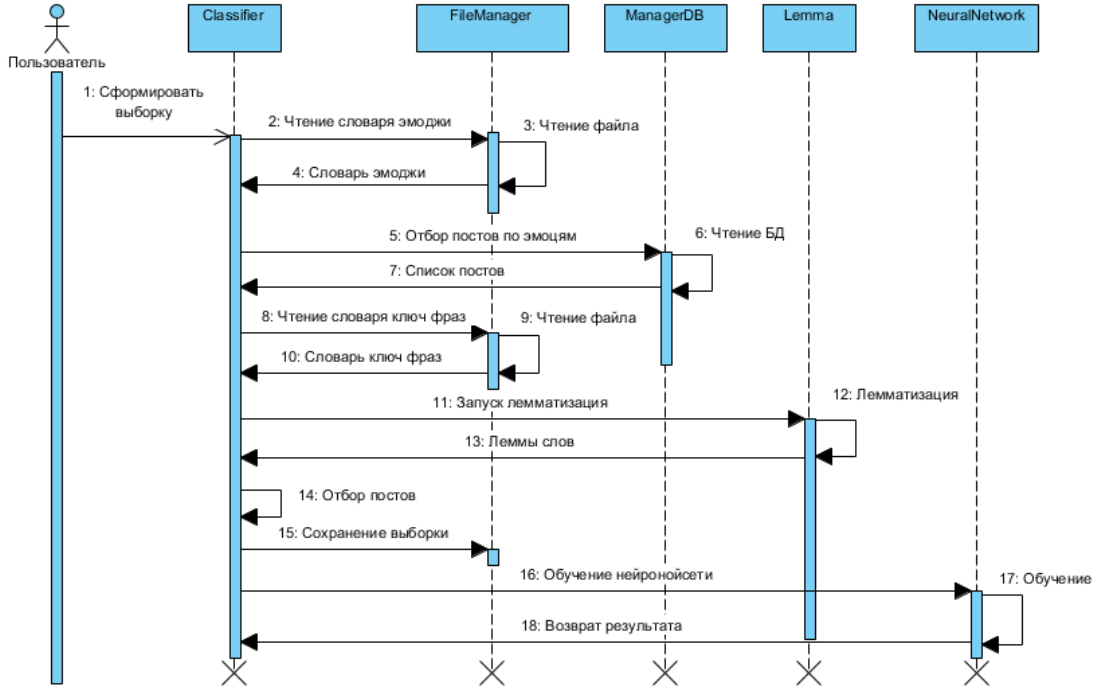

Рисунок 3. Диаграмма последовательности действий.

Первоначально происходит считывание словарей с авторскими символами выражения эмоций, а затем происходит отбор постов. После этого считываются словари с ключевыми фразами, посты лемматизируются и происходит отбор по ключевым фразам. Затем отобранные посты сохраняются в текстовые файлы. После формирования обучающей выборки происходит обучение и тестирование точности определения эмоциональной окраски постов нейронной сетью.

При построении программной системы были использованы следующие библиотеки:

Lucene Russian Morphology – библиотека морфологического разбора [9]. Данная библиотека выполняет морфологический разбор слова. Библиотека позволяет выполнить лемматизацию исходного слова на русском языке и получить информацию о части речи. Lucene использует словарную базовую морфологию с некоторой эвристикой для неизвестных слов и поддерживает омонимы.

Encog Machine Learning Framework – это библиотека машинного обучения [10]. Библиотека поддерживает различные алгоритмы обучения. Основное преимущество библиотеки заключается в алгоритмах нейронной сети. Библиотека содержит классы для создания широкого спектра сетей, а также поддерживает классы для нормализации и обработки данных для этих нейронных сетей. Многопоточность используется для обеспечения оптимальной производительности обучения на многоядерных машинах.

PostgreSQL JDBC Driver – библиотека, обеспечивающая доступ к БД PostgreSQL [11]. Библиотека обеспечивает подключение к БД и взаимодействие с ней. В качестве параметров библиотека принимает на вход адрес и порт БД, а так же логин и пароль для подключения. Далее библиотека принимает на вход SQL запросы к БД и возвращает данные.

5. Эксперименты

Качество сформированной обучающей выборки будем оценивать как точность определения эмоциональной окраски текста нейронной сетью.

Для проведения экспериментов были выбраны следующие параметры: различное количество постов в обучающей выборке и два метода обработки текста: стемминг и лемматизация. Точность работы системы измерялась на тестовых постах, каждый из которых однозначно относится к одной категории. Тестовые посты приведены в таблице 1.

Таблица 1. Тестовые посты.

Эмоция	Текст
Радость	я люблю прекрасные выходные
Грусть	грустно прощаться
Удивление	продолжаю удивляться многообразию России
Страх	сердце вырывается из груди
Злость	олени на дорогах
Презрение	а как ты думаешь
Отвращение	мама не разрешает

Качество обучающей выборки будем определять как количество верных выводов делённых на количество тестовых постов. Результаты экспериментов приведены в таблице 2.

Таблица 2. Результаты экспериментов.

Кол-во постов	Стемминг	Лемматизация
20	4/7	6/7
50	6/7	7/7
100	4/7	7/7
200	4/7	7/7
300	5/7	7/7

Проведённые эксперименты показывают, что обучающая выборка, сформированная с применением метода лемматизации, получается более качественная, чем с помощью метода стемминга. Из таблицы 1 видно, что точность распознавания постов нейронной сетью значительно выше, когда обучающая выборка формируется с применением метода лемматизации. Результаты эксперимента также представлены в виде графика на рисунке 4.

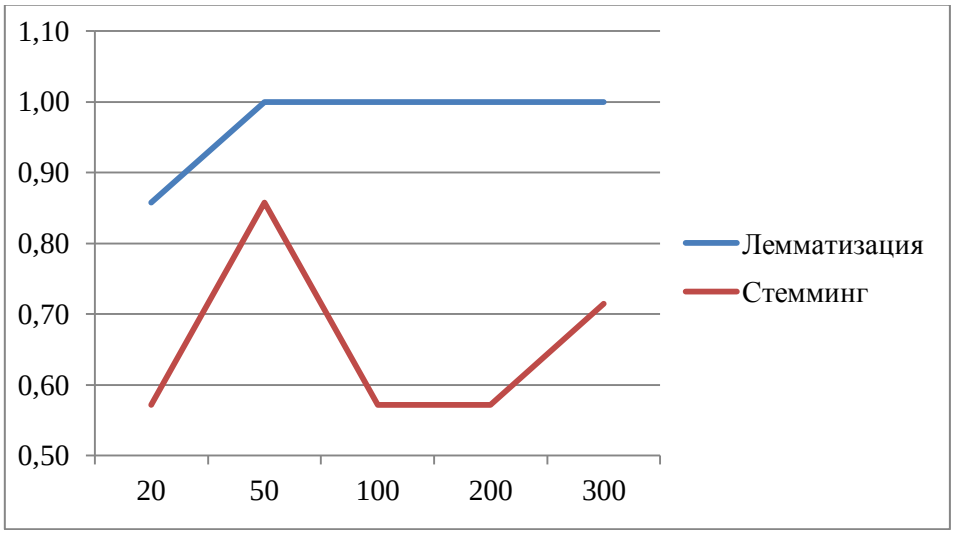

Рисунок 4. Эксперименты.

Дополнительно в нейронную сеть были поданы по 200 постов, относящихся к каждому классу эмоциональной окраски. Результаты эксперимента представлены в таблице 3:

Таблица 3. Результаты экспериментов.

Эмоция	Всего	+	-
Радость	200	148	52
Грусть	200	154	46
Злость	200	110	90
Удивление	200	126	74
Страх	200	101	99
Отвращение	200	151	49
Презрение	200	121	79
Сумма:	1400	936	464
Проценты:		0,669	0,331

Эксперименты показывают, что нейронная сеть правильно распознаёт эмоцию с точностью 67%. Лучше всего нейронная сеть определяет радость, грусть и отвращение с точностью около 75%. Результаты эксперимента также представлены в виде графика на рисунке 5.

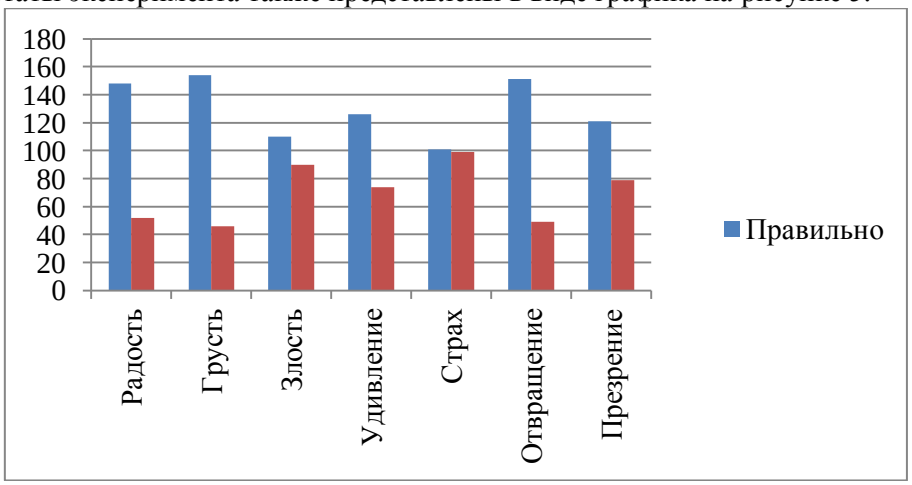

Рисунок 5. Эксперименты.

6. Заключение

В результате роботы была разработана экспертная система определения эмоциональной окраски постов социальной сети. Обучающая выборка создаётся в автоматизированном режиме с использованием словарей авторских символов выражения эмоций и словарей ключевых фраз. Нейронная сеть правильно определяет класс эмоциональной окраски поста с точностью 67%. Эмоции радости, грусти и отвращения нейронная сеть распознаёт с точностью 75%.

В будущем планируется совершенствовать алгоритм формирования обучающей выборки. Составленные словари будут расширены и уточнены. Для тестирования выборки будут использоваться нейронные сети различной архитектуры, например, глубокого обучения.

7. Благодарности

Работа выполнена при финансовой поддержке РФФИ. Проекты № 18-47-730035 и 18-47-732007.

8. Литература

[1] Гришелёнок, Д.А. Использование результатов математического планирования эксперимента при формировании обучающей выборки нейросети: статья / Д.А. Гришелёнок, А.А. Ковель – Красноярск: СибГАУ, 2010.

[2] Кафтанников, И.Л. Проблемы формирования обучающей выборки в задачах машинного обучения / И.Л. Кафтанников, А.В. Парасич // Вестник ЮУрГУ. Серия «Компьютерные технологии, управление, радиоэлектроника». – 2016. – Т. 16, № 3. – С. 15-24.

[3] Посевкин, Р.В. Применение сентимент-анализа текстов для оценки общественного мнения / Р.В. Посевкин, И.А. Бессмертный // Научно-технический вестник информационных технологий, механики и оптики. – 2015. – Т. 15, № 1. – С. 169-171.

[4] Модуль SentiFinder [Электронный ресурс]. – Режим доступа: eurekaengine.ru (Дата обращения: 20.11.19).

[5] Тезаурус WordNet [Электронный ресурс]. – Режим доступа: http://wndomains.fbk.eu/wnaffect.html (Дата обращения: 20.11.19).

[6] Тезаурус SentiWordNet [Электронный ресурс]. – Режим доступа: http:// sentiwordnet.isti.cnr.it (Дата обращения: 20.11.19).

[7] Тезаурус SenticNet [Электронный ресурс]. – Режим доступа: https://sentic.net (Дата обращения: 20.11.19).

[8] Алгоритм Word2Vec [Электронный ресурс]. – Режим доступа: https://neurohive.io/ru/ (Дата обращения: 20.11.19).

[9] Библиотека морфологической обработки Russian Morphology: Russian [Электронный ресурс]. – Режим доступа: https://github.com/AKuznetsov/russianmorphology / (Дата обращения: 20.11.19).

[10] Библиотека нейронной сети Encog Machine Learning Framework [Электронный ресурс]. – Режим доступа: https://www.heatonresearch.com/encog/ (Дата обращения: 20.11.19).

[11] Библиотека доступа к БД PostgreSQL JDBC Driver [Электронный ресурс]. – Режим доступа: https://jdbc.postgresql.org / (Дата обращения: 20.11.19).

An approach to the formation of a training sample for assessing the sentiment degree of social network posts using machine learning

A.A. Konstantinov[1]

[1]Ulyanovsk State Technical University, Severny Venets 32, Ulyanovsk, Russia, 432027

Abstract. This article describes a system for determining the emotional coloring of social network posts. The article describes in detail the developed algorithm for the formation of the training sample, as well as the implementation of the software system and experiments.