УДК 004.912

Преодоление деградации результатов классификации текстов по тональности в коллекциях, разнесенных во времени

Рубцова Ю.В. (Институт систем информатики СО РАН)

В данной работе представлены подходы для решения задачи улучшения классификации текстов по тональности в динамически обновляемых текстовых коллекциях. Предлагается три метода решении обозначенной задачи, принципиально различающихся между собой. В данном случае для классификации текстов по тональности используются методы машинного обучения с учителем и методы машинного обучения без учителя. Приведены сравнения методов и показано в каких случаях какой метод наиболее применим. Описываются экспериментальные сравнения методов на достаточно представительных текстовых коллекциях.

Ключевые слова: *корпусная лингвистика, классификация текстов, анализ тональности текстов, машинное обучение, анализ данных социальных сетей*

1. Введение

Большая часть информации, содержащейся в сети, представлена в текстовом виде на естественном языке. Это усложняет ее обработку и требует привлечения методов компьютерной лингвистики. Поэтому в настоящее время возрастает актуальность лингвистических исследований, разработок новых эффективных программных систем извлечения фактов из неструктурированных массивов текстовой информации и классификации и кластеризации информации, нацеленных как на анализ самих сообщений в сети, так и на выявление источников распространяемой информации. На протяжении последних десяти лет, задачей автоматического извлечения и анализа отзывов и мнений из социальных медиа занимается много ученых и исследователей по всему миру. При этом в качестве одной из главных задач рассматривается задача классификации текстов по тональности.

Тема автоматической классификации текстов по тональности актуальна в России и за рубежом. Одна из первых задач классификации текстов по тональности, которой занимались исследователи, была задача классификации всего документа целиком [25, 30]. Подобный

уровень классификации предполагает, что весь документ выражает всего одну тональную оценку или мнение по поводу некоторого объекта или сущности. Если документ содержит описание нескольких объектов или сравнение нескольких объектов, то классификация текстов на уровне документов не даст корректного представления о тональности документа. Чуть позже, классификацию на уровне коротких фраз и выражений, а не на уровне абзацев или целых документов, проводили Wilson, Wiebe и Hoffmann [31]. В своей работе авторы показали, что важно определить окраску (положительная или отрицательная) отдельно взятого предложения, а не всего текста целиком. В длинном документе мнение автора об объекте может меняться с положительного на отрицательное и наоборот; также автор может отрицательно высказываться о мелких недочетах, но в целом оставаться положительно настроенным по отношению к описываемому в тексте объекту. Другими словами, не всегда длинный документ или отзыв однозначно можно классифицировать как положительно или отрицательно окрашенный.

Сообщения микроблогов не превосходят 140 символов, что дает нам возможность отнести классификацию постов микроблогов к классификации на уровне фраз или предложений. Несмотря на то, что микроблоги достаточно молодое явление, исследователи активно занимаются анализом тональности сообщений блогов в целом и Твиттера в частности [7, 10, 16, 24].

Сообщения микроблогов достаточно короткие, чтобы описывать все различные аспекты продукта или услуги и в то же время насыщены мнениями и эмоциональными оценками, поэтому задачу тоновой классификации коротких сообщений решают не только на уровне фраз и предложений, но в том числе и относительно заданного объекта [17, 19].

Большой научный и практический интерес к задаче автоматического извлечения и анализа текстов связан с тем, что пользователи ежедневно публикуют сотни тысяч мнений в социальных сетях, блогах, форумах, специализированных площадках, которые необходимо обрабатывать в полном объеме. Поэтому системы, автоматически распознающие тональность сообщений и умеющие вычленять мнение в текстах, востребованы специалистами, разрабатывающими рекомендательные системы, экспертные системы; маркетологами и аналитиками, проводящими маркетинговые исследования; политологами, которые оценивают тональность новостей и настроение населения и др.

Одной из сложных проблем в разработке и использовании систем, определяющих тональность сообщений, является то, что с течением времени качество их работы постоянно ухудшается. Это происходит, главным образом, из-за того, что со временем меняется словарный состав сообщений. В статье предлагаются подходы к решению данной проблемы.

Статья организована следующим образом, во второй главе обозначается и обосновывается проблема ухудшения качества классификации текстов по тональности на коллекциях одинаковых по составу и характеристикам, но разнесенных во времени, для этого описываются коллекции на которых проводились эксперименты, описываются метрики оценки качества результатов классификатора и приводятся результаты экспериментов работы классификатора на текстовых коллекциях собранных с разницей в полгода-полтора года. В третьей главе предлагаются подходы к решению этой задачи. Последний раздел состоит из выводов и заключения.

2. Снижение качества классификации текстов по тональности из-за изменения тональной лексики

Пользователи социальных сетей одни из первых начинают использовать новые термины в повседневном общении. Среди 40 новых слов, включенных в словарь Оксфорда в 2013 году были термины, пришедшие из социальных сетей, например: Srsly (сокращение от англ. seriously – серьезно), selfie (фотографирование самого себя, русский аналог – себяшка, селфи). Таким образом, активный лексикон постоянно дополняется, следовательно, автоматические классификаторы должны учитывать это в своих моделях. Если мы говорим о машинном обучении, то обучающие коллекции текстов должны пополняться. Если речь идет об использовании правил и словарей, то для улучшения качества классификаторов необходимо учитывать сленг, которым насыщены социальные сети. Так как активный словарный запас регулярно пополняется новыми терминами, в том числе и терминами, выражающими эмоции, следовательно, и словари тональной лексики также должны регулярно обновляться.

2.1. Коллекции коротких сообщений

Работы и эксперименты по автоматической классификации текстов показывают, что результаты классификации, как правило, зависят от обучающей текстовой выборки и предметной области, к которой относится обучающая коллекция. На сегодняшний день, многие работы сводятся к построению вектора признаков (англ. feature engineering) и подключению дополнительных данных, таких как внешние текстовые коллекции (не пересекающиеся с обучающей коллекцией) или тональные словари. Дополнительные данные позволяют снизить зависимость от обучающей коллекции и улучшить результаты классификации.

Для качественного решения задачи классификации текстов по тональности необходимо иметь размеченные коллекции текстов. Более того, для решения задачи улучшения классификации по тональности в динамически обновляемых коллекциях, необходимо иметь несколько текстовых коллекций, которые были собраны в разные временные промежутки.

Сбор первого корпуса текстов проходил в декабре 2013 года – феврале 2014 года, для краткости будем называть ее коллекцией 2013 года. В соответствии с письменным обозначением эмоций был произведен поиск позитивно и негативно окрашенных сообщений. Таким образом, из коллекции 2013 года сформировано две коллекции: коллекция положительных твитов и коллекция негативных твитов. Нейтральная коллекция была сформирована из сообщений новостных и официальных аккаунтов twitter. С помощью метода [13] и предложенной автором фильтрации [8] из текстов 2013 года была сформирована обучающая коллекция.

Далее, необходимо собрать и подготовить тестовые коллекции текстов. Сбор второго корпуса, который состоит из около 10 миллионов коротких сообщений, проходил в июле-августе 2014 года. Третий корпус, состоящий из около 20 млн. сообщений, был собран в июле и ноябре 2015 года.

Из текстов 2014 и 2015 гг., сформированы две тестовые коллекции. Тексты 2014 и 2015 годов подверглись идентичной фильтрации, что и обучающая коллекция 2013 года. Формирование тестовых коллекций по классам тональности происходило аналогично обучающей коллекции, с помощью метода [13]. Распределение количества сообщений по классам тональности в коллекциях представлено в Таблица 1. Все три коллекции являются предметно независимыми, то есть не относятся ни к какой заранее определенной предметной области.

Таблица 1. Распределение сообщений в коллекциях по классам тональности

	Положительные сообщения	Отрицательные сообщения	Нейтральные сообщения
2013 год	114 911	111 922	107 990
2014 год	5 000	5 000	4 293
2015 год	10 000	10 000	9 595

Собранные коллекции текстов послужили основой для создания обучающей и тестовой коллекций твиттер-сообщений для оценки твита по тональности относительно заданного объекта на соревновании классификаторов SentiRuEval [3, 19, 20] в 2015 и 2016 годах. В 2015

году описанные коллекции использовались в бакалаврской работе студента МГУ на тему «Анализ тональности предложений на материале сообщений из Твиттера». В 2016 году, коллекции были использованы в качестве дополнительных коллекций для извлечения лексикона для алгоритма автоматической классификации текстов по тональности [9]. Более того, так как текстовые коллекции выложены в открытый доступ, на основе этих коллекций были разработаны веб приложения «Настроение России online» [6] и «Мониторинг тональности твитов о ВУЗ'ах в режиме реального времени» [5].

Ранее в работе [7] автором было показано, что собранные коллекции являются полными и достаточно представительными.

2.2. Метрики оценки качества классификатора

Оценка качества системы классификации текстов по тональности происходит путем сравнения результатов, полученных от автоматической системы классификации и эталонных размеченных результатов.

Основываясь на разнице значений эталонной коллекции и коллекции, автоматически размеченной оцениваемым алгоритмом, вычисляют следующие общепринятые метрики: accuracy, формула 1; точность (англ. precision), формула 2; полнота (англ. recall), формула 3; и F–мера, формула 4 [21].

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}, \quad (1)$$

где,

- TP – истинно положительное решение, количество текстов, правильно отнесенных к классу P;
- FP – ложно положительное решение, количество текстов, не правильно отнесенных к классу P;
- FN – ложно отрицательное решение, количество текстов, не правильно отнесенных к классу N;
- TN – истинно отрицательное решение, количество текстов, правильно отнесенных к классу N.

для получения стабильных результатов при оценке качества систем классификации используют более устойчивые метрики, такие как полноту, точность и гармоническое среднее между полнотой и точностью – F-меру.

Precision (точность) – это доля объектов классифицированных как X, которые действительно принадлежат классу X или вероятность того, что случайно выбранный твит попал в тот класс, которому он на самом деле принадлежит Формула 2.

Recall (полнота) – это доля всех объектов класса X, которые согласно алгоритму классифицируются как принадлежащие классу X или вероятность того, что случайно выбранный твит из класса при классификации в него и попадёт, вычисляется по формуле 3.

$$\text{Precision} = \frac{TP}{TP+FP}, \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN}, \quad (3)$$

F-мера это гармоническое среднее между точностью и полнотой:

$$\text{F-мера} = 2 \times \frac{Precision \times Recall}{Precision+Recall}, \quad (4)$$

В данной работе F-мера считается как среднее значение между F-мерой по каждому из классов тональности. Аналогично, Precision и Recall – среднее значение Precision и Recall по каждому из классов тональности в отдельности:

$$F\text{-}measure = \frac{F_{positive} + F_{negative} + F_{neutral}}{3},$$

$$F_{positive} = 2\frac{\text{Pr}ecision_{positive} \times \text{Re}call_{positive}}{\text{Pr}ecision_{positive} + \text{Re}call_{positive}},$$

$$\text{Pr}ecision = \frac{P_{positive} + P_{negative} + P_{neutral}}{3},$$

$$\text{Re}call = \frac{R_{positive} + R_{negative} + R_{neutral}}{3}.$$

2.3. Описание проблемы снижения качества классификации текстов по тональности из-за изменения тональной лексики

Для моделирования реальной ситуации, когда со временем может видоизменяться язык или обсуждаемые в социальных сетях темы, подготовлены вторая и третья коллекции коротких сообщений. Разница во времени между сбором первой и второй коллекций около полугода, первой и третьей – полтора года. Несмотря на то, что на первый взгляд лексика не может так быстро измениться, тем не менее темы твитов, влияющие на общее настроение в целом и репутацию в частности, значительно зависят от происходящих позитивных или

негативных событий с участием целевого объекта и, как правило, такие события невозможно предсказать заранее. Например, в январе-феврале 2014 года около 12% всех сообщений twitter были про олимпиаду, в августе 2014 года упоминания олимпиады не превосходило 0,5% от числа всех сообщений.

Прежде необходимо показать снижение качества классификации на коллекциях, разнесенных во времени. Для это обучаем модель классификатора на коллекции 2013 года и применяем ее к коллекциям 2014 и 2015 годов. В качестве словарей для построения признакового пространства выбраны словари men_3, men_5 и BOW. Префикс men_N означает, что термин встречается не меньше чем N раз в одной из коллекции, соответствующей одному из классов тональности (положительной, отрицательной или нейтральной). Общее значение количества терминов в обучающей коллекции обозначено как BOW (англ. Bag of words).

Результаты эксперимента, показывающие снижения качества классификации текстов, представлены в таблице 2. Из таблицы 2 видно, что за полтора года качество классификации текстов микроблогов согласно F-меры может упасть до 15-20% в зависимости от выбранного набора признаков.

Таблица 2 Метрики качества классификации текстов микроблогов по тональности на коллекциях, разнесенных во времени

BOW				**Men_3_tfidf**				**Men_5_tfidf**			
Acc	**P**	**R**	**F-мера**	**Acc**	**P**	**R**	**F-мера**	**Acc**	**P**	**R**	**F-мера**
Коллекция 2013 года											
0,7459	0,7595	0,7471	0,7505	0,6457	0,6591	0,6471	0,6506	0,6189	0,6542	0,6184	0,6223
Коллекция 2014 года											
0,6964	0,6984	0,7062	0,6933	0,5086	0,5829	0,5040	0,5026	0,5745	0,5823	0,5795	0,5808
Коллекция 2015 года											
0,6118	0,6317	0,6156	0,5996	0,4651	0,5218	0,4638	0,4549	0,5343	0,5337	0,5360	0,5344

3. Способы уменьшения деградации результатов классификации на текстовых коллекциях, разнесенных во времени

В качестве классификатора был использован метод SVM (support vector machine) и библиотека LIBLINEAR [12]. Библиотека LIBLINEAR – это реализация алгоритма SVM с линейным ядром. Как показывают эксперименты, обучение модели с помощью библиотеки LIBLENIAR существенно превосходит по скорости аналоги, поэтому в данной работе использовалась библиотека LIBLINEAR.

3.1. Использование весовой схемы с линейной вычислительной сложностью

В активный словарный запас постоянно входят новые слова и выражения. Первый вариант уменьшения устаревания словаря – это его постоянное обновление. Таким образом, можно будет следить за появлением новых терминов в языке и своевременно учитывать их при классификации. Постоянно обновлять словарь и пересчитывать веса терминов достаточно затратное действие с точки зрения вычислительной мощности. Следовательно, для постоянного обновления словаря надо подобрать вычислительно не затратную весовую схему. Так, например, для использования метода, основанного на мере TF-IDF:

$$tfidf = tf \times log\frac{T}{T(t_i)} \quad (5)$$

необходимо знать частоту встречаемости термина в коллекциях, следовательно, набор данных не должен меняться во время расчета весов. Это существенно усложняет вычисления при обновлении словаря, если требуется провести обсчет данных в реальном времени. При добавлении нового текста в коллекцию требуется пересчитать веса для всех терминов в коллекции. Вычислительная сложность перерасчета весов всех терминов в коллекции равна $O(N^2)$.

Для того, чтобы решить проблему поиска и расчёта весов терминов в режиме реального времени была использована мера Term Frequency – Inverse Corpus Frequency (TF-ICF) – формула 6. [14]

$$tf.icf = tf \times \log(1 + \frac{|C|}{cf(t_i)}) \quad (6)$$

Где C – это число категорий, cf – число категорий, в которых встречается взвешиваемый термин.

Для расчета TF-ICF не требуется информация о частоте использования термина в других документах коллекции, только принадлежность к категории, таким образом, вычислительная сложность меры TF-ICF равна *O(N)*.

Проверим применимость меры TF-ICF для взвешивания признаков для классификации текстов по тональности. Для этого поставим эксперимент на ранее описанных текстовых коллекциях коротких сообщений. Начнем с того, что получим базовые значения результатов классификатора, которые будем улучшать. Для этого из коллекции 2013 года создается словарь, на основе которого строятся вектора признаков. Для векторной модели признаки взвешиваются схемой TF-ICF, также используем булевскую модель (признак может принимать только два значения 0 – признак отсутствует или 1 – признак присутствует). Используем коллекцию 2013 года в качестве обучающей, на ее основе создается модель классификатора, коллекции 2014 и 2015 годов выступают в качестве тестовых коллекций. С целью выбора признаков, был проведен эксперимент для словарей, в котором термины взвешены мерой TF-ICF. На Рис. 1 представлены результаты работы классификатора согласно F-меры. Видно, что наилучшие результаты показывают словари из которых удалили термины, которые в одной из тональных коллекций встречаются менее трех раз (men_3) и менее пяти раз (men_5). В словарях с названиями 1_0_0, 3_0_0, 5_0_0 удалены термины, которые встречаются во всей обучающей коллекции менее 1, 3 или 5 раз соответственно.

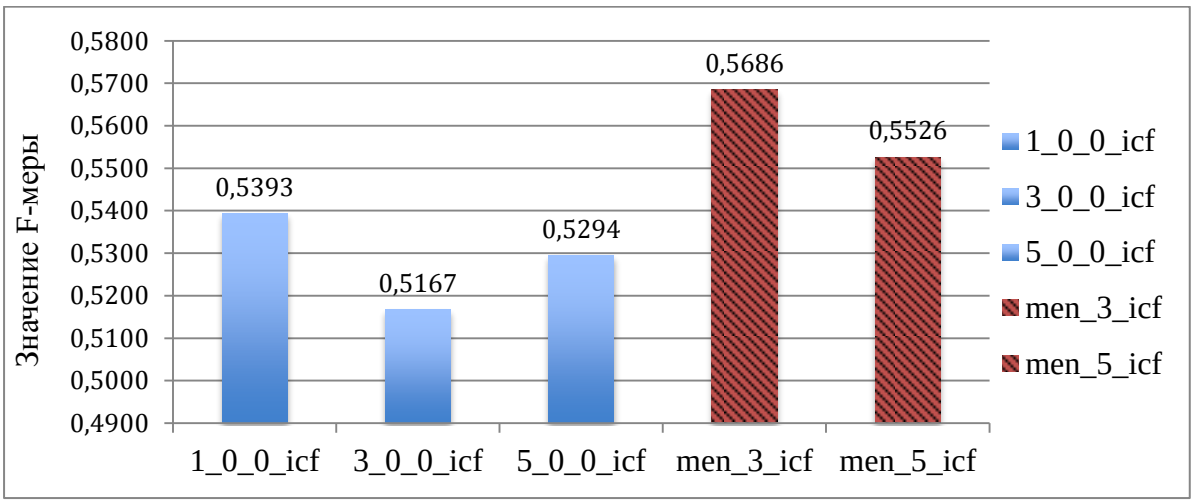

Рис. 1. Усредненные значения F-меры при перекрестной проверке на обучающей коллекции для каждого из словарей признаков взвешенных мерой TF-ICF.

Используя словари men_3_icf и men_5_icf, показавшие согласно F-мере наилучшие результаты классификатора при перекрестной проверке на коллекции 2013 года,

протестируем полученную модель классификатора на коллекциях 2014 и 2015 годов. В Таблица 3 приведены результаты F-меры при переносе модели 2013 года на коллекции 2014 и 2015 годов, для наглядности оставлены значения F-меры при перекрестной проверке на коллекции 2013 года.

Таблица 3 Значение F-меры и точности при классификации по тональности с использованием двух лексиконов взвешенных мерой TF-ICF

	Лексикон men_3_TF_ICF		**Лексикон men_5_TF_ICF**	
	F-мера	Accuracy	F-мера	Accuracy
2013	0,5686	0,5648	0,5526	0,5541
2014	0,4645	0,4833	0,4564	0,4971
2015	0,4109	0,4278	0,4143	0,4516

Наблюдается снижение качества классификации при тестировании классификатора на разнесенных во времени коллекциях. Согласно F-мере качество классификации снижается до 15%.

Несмотря на недостатки весовой схемы TF-ICF в виде относительно низкого значения F-меры, у этой схемы есть существенное преимущество в виде линейной вычислительной сложности, что особенно актуально, если речь идет о пространстве, состоящем из нескольких сотен тысяч признаков.

Следующий этап состоит в том, чтобы объединить словарь 2013 года со словарем 2014 года в один, пересчитать веса в полученном словаре и увеличенный словарь использовать для построения классификатора на объединенной коллекции 2013+2014 годов и тестировании на коллекции 2015 года. Для объединённых коллекций был применен метод перекрестной проверки с шагом 5, далее обученный на коллекции 2013+2014 классификатор тестировался на коллекции 2015 года. Ожидается, что значение F-меры для перекрестной проверки классификатора, построенного с помощью словаря признаков men_3_tficf будет в окрестности значения 0,5686 (см Таблица 3), а значение F-меры для коллекции 2015 года будет превосходить 0,4109. Также были проведены эксперименты на словаре признаков BOW. Результаты работы классификатора, согласно F-мере представлены в Таблица 4.

Таблица 4 Результаты классификатора при добавлении коллекций 2014 года к обучающей коллекции

	BOW				**men_3_TF-ICF**			
	Acc	P	R	F	Acc	P	R	F
2013+2014 (перекрестная проверка)	0,7205	0,7339	0,7215	0,7250	0,5539	0,5806	0,5550	0,5565
2015	0,6848	0,6889	0,6862	0,6872	0,5348	0,5571	0,5361	0,5334

Действительно, классификатор показал улучшение значений F-меры для коллекции 2015 года как с помощью использования словаря men_3_tf_icf, так и с помощью метода мешка слов, сохранив порядок результирующих значений при перекрёстной проверке классификатора на обучающей коллекции на уровне 0,55-0,57 для словаря men_3_tf_icf (Таблица 3) и на уровне 0,72-0,75 для мешка слов (Таблица 2).

На третьем этапе все три коллекции были объединены в одну. На основе объединенной коллекции был извлечен словарь для формирования вектора признаков. Термины словаря были взвешены мерой TF-ICF. Аналогично предыдущему эксперименту, задача классификатора состоит в том, чтобы сохранить результирующие значения классификатора не ниже 0,55 согласно F-мере при использовании в качестве признаков словаря men_3_tf_icf и не менее 0,72 при использовании мешка слов. Результаты работы классификатора на объединенных коллекциях 2013, 2014 и 2015 годов представлены на рисунке 2.

На Рис. 2 представлены графики, наглядно показывающие, что динамическое обновление лексикона позволяет сократить ухудшение качества классификации по тональности на разнесенных во времени коллекциях. Сплошная линия показывает значения результата F-меры при обновлении словаря и обучающей коллекции, пунктирная линия показывает результаты классификации при использовании коллекции 2013 года в качестве обучающей.

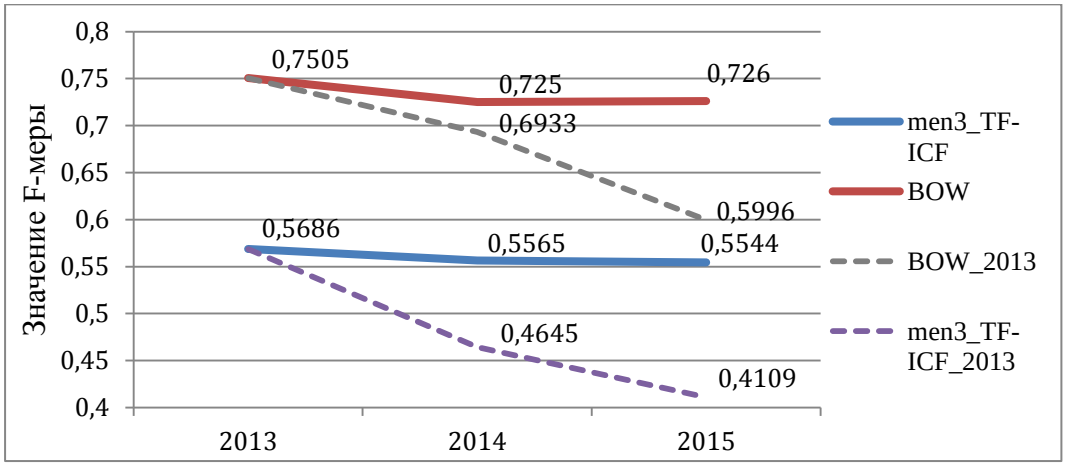

Рис. 2. Результаты F-меры при динамическом обновлении лексикона и тренировочной коллекции (сплошная линия) и без (пунктирная линия)

Классификатор ведет себя единообразно во всех проведенных экспериментах, что позволяет судить о достоверности результатов.

При динамическом обновлении словаря постоянно увеличивается и размерность признакового пространства. Так, при объединении коллекции 2013 года и 2014 года, добавилось более семи с половиной тысяч новых терминов, часть из которых является набором символов, не несущих никакого смысла или встречающихся менее 3-х раз в объединенной коллекции, например, «ш__», «Х_м__с», «х_нд», «болчлоо» или «волчехь». Поэтому, кроме мешка слова, в работе рассматривается отфильтрованный словарь, который показал наилучшие результаты для поставленной задачи – словарь men_3. Это словарь в котором были отфильтрованы все термины, которые встречаются менее трех раз в одной из тональной коллекции. В Таблица 5 представлены значения увеличения размерности исходного словаря при добавлении в него терминов из коллекции 2014 и 2015 годов.

Таблица 5 увеличение размера лексикона, при расширении обучающей коллекции

	BOW	men_3
2013	219 280	41 295
2013+2014	226 964	42 867
2013+2014+2015	245 845	46 312

При накоплении большого количества текстов, становится затруднительно динамически пересчитывать веса терминов в режиме реального времени с использованием меры TF-IDF.

При использовании подхода мешка слов, размерность словаря (а, следовательно, и размерность вектора признаков) будет постоянно расти, потребляя вычислительные ресурсы, но не повышая качество классификации, которое зафиксировалось на уровне 0,72-0,75 согласно F-мере. Использование отфильтрованного словаря и меры TF-ICF сдерживает рост размерности векторов признаков и позволяет пересчитывать веса терминов в режиме реального времени, однако качество классификации согласно F-мере остается на уровне 0,55.

Использование описанного метода оправдано при ограниченных вычислительных ресурсах, а также при отсутствии внешних тональных словарей и дополнительных текстовых коллекций.

3.2. Использование внешних словарей оценочных слов и выражений

Следующая гипотеза состоит в том, что использование внешних словарей эмоционально окрашенной и/или оценочной лексики, повышает качество классификации текстов по тональности, а также сокращается зависимость классификатора от обучающей коллекции. Термины словаря могут быть использованы в качестве признаков в машинном обучении [26] или же использоваться в подходах, основанных на словарях и правилах [27]. Существуют работы, описывающие извлечение или настройку тональных словарей на определенную заранее заданную предметную область [1, 2]. Приводятся примеры терминов, которые могут описывать позитивные характеристики в одной предметной области и нейтральные или даже негативные – в другой. Однако, как показывают [26, 22], объединение обучающих данных из разных предметных областей улучшает качество классификации по тональности в каждой из выбранных областей. Следовательно, существует множество оценочных слов с ярко-выраженной тональной ориентацией, подходящих для разных предметных областей.

В качестве внешних подключаемых словарей в данной работе были использованы два обще тематических словаря тональной лексики, размеченные экспертами: РуСентиЛекс и Linis-crowd.

РуСентиЛекс [4] – лексикон был собран из нескольких источников: оценочные слова из тезауруса русского языка РуТез, сленговые слова из Твиттера и слова с позитивными или негативными ассоциациями (коннотациями) из корпуса новостей. Словарь содержит более десяти тысяч слов и словосочетаний русского языка. Лексикон включает в себя оценочные слова, автоматически извлеченные из текста и проверенные экспертами.

Другой словарь, который использовался в этой работе это **Linis-crowd** [11]. Несмотря на то, что авторы лексикона для формирования словаря использовали тексты социально политической тематики, отмечается, что в словаре присутствует лексика не специфичная для

социально-политической тематики, но передающая эмоциональную оценку, поэтому авторами словаря было решено включить ее в прототип Linis-crowd. Словарь содержит 9539 терминов. Каждый термин имеет вес от -2 (сильно-негативный) до +2 (сильно позитивный).

Подключение тональных словарей. Для тонального классификатора, основанного на методах машинного обучения, помимо признаков, порождаемых на основе обучающих данных, были добавлены словарные признаки. Для каждого термина *w* из словаря, обладающего полярностью *p* определено значение *(w, p)*:

$$(w, p) = \begin{cases} > 0, w - positive \\ < 0, w - negative \\ = 0, w - neutral \end{cases}. \quad (7)$$

В качестве признаков добавляются:

- общее количество терминов *(w, p)* в тексте твита;
- сумма всех значений полярностей слов лексикона $\sum_{w \in tweet}(w, p)$;
- максимальное значение полярности: $\max_{w \in tweet}(w, p)$.

Каждый из словарей подключался отдельно, сравнение результатов работы словарей можно увидеть в Таблица 6. Как видно из таблицы, оба словаря показывают схожие результаты на обучающей и тестовых коллекциях.

Таблица 6 результаты работы классификатора при подключении словарей РуСентиЛекс и Linis-Crowd

	РуСентиЛекс				**Linis-Crowd**			
	Acc	**P**	**R**	**F**	**Acc**	**P**	**R**	**F**
2013	0,7273	0,74	0,7284	0,7318	0,7272	0,7398	0,7283	0,7316
2014	0,7245	0,7387	0,7259	0,7295	0,7244	0,7386	0,7258	0,7294
2015	0,6724	0,6802	0,6733	0,6759	0,6725	0,6803	0,6733	0,6760

Таким образом, с помощью подключения внешних лексиконов удается приостановить снижение качества классификации на коллекциях, разнесенных во времени. Так как основные признаки порождаются обучающей коллекцией, тенденция к деградации классификатора все же сохраняется, однако, она сокращается с 15% при использовании мешка слов (таблица 2) до 5,6% при подключении словарей эмоциональной лексики.

Таким образом, видно, что при наличии внешних подключаемых тональных словарей имеет смысл использовать этот метод, так как он позволяет сдержать падение качества классификации текстов по тональности на коллекциях, разнесенных во времени.

3.3. Использование распределенных представлений слов в качестве признаков

В двух предыдущих методах, пространство признаков для обучения классификатора строится на основе обучающей коллекции, следовательно, сильно зависит от качества и полноты этой коллекции. Несмотря на хорошие результаты описанных выше моделей, между терминами нет семантических связей, а постоянное добавление новых терминов ведет к увеличению размерности вектора признаков. Еще одним способом преодоления устаревания лексикона является использование пространства распределенных представлений слов в качестве признаков для тренировки классификатора.

3.3.1. Пространство распределённых представлений слов

Распределённое представление слова (англ. distributed word representation, word embedding) – это k-мерный вектор признаков $w=(w_1,\ldots,w_k)$, где $w_i \in R$ это компоненты вектора [28]. Если сравнивать с бинарной моделью или моделью взвешенного вектора, то количество координат k такого вектора существенно меньше. Обычно это число не превосходит нескольких сотен, в случае бинарной модели оно измеряется десятками тысяч, в зависимости от размера исходного словаря.

Основная идея векторного распределенного представления слов заключается в нахождении связей между контекстами слов согласно предположению, что находящиеся в похожих контекстах слова, скорее всего означают или описывают похожие предметы или явления, т.е. являются семантически схожими. Для этого каждый термин представляется в виде вектора из k координат в которых закодированы полезные признаки, характеризующие этот термин и позволяющие определять сходство этого термина с похожими терминами в коллекции. Формально это представление терминов является задачей максимизации косинусной близости между векторами слов, которые появляются рядом друг с другом в близких контекстах, и минимизация косинусной близости между векторами слов, которые не появляются в близких контекстах. Косинусная мера близости между векторами, $\cos(\theta)$, может быть представлена следующим образом (формула 8):

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}, \quad (8)$$

где A_i и B_i координаты вектора A и B соответственно.

Помимо сокращения размерности вектора признаков, распределенное представление слов учитывает смысл слова в контексте. То есть позволяет обобщить, например, «быстрый автомобиль» на отсутствующее в обучающей выборке «шустрая машина», тем самым снижается зависимость от обучающей выборки.

Для получения распределенного представления слов используют модели машинного обучения без учителя, напр. CBOW, Skip-Gram, AdaGram [29], Glove. Результаты недавних исследований показывают [18], что нейронная языковая модель Skip-gram превосходит другие модели по качеству получаемых векторных представлений. Поэтому в данной работе используется модель Skip-Gram.

3.3.2. Модель Skip-Gram

Модель Skip-Gram была предложена Томасом Миколовым с соавторами в 2013 году [23]. На вход модели подается неразмеченный корпус текстов, для каждого слова рассчитывается количество встречаемости этого слова в корпусе. Массив слов сортируется по частоте, редкие слова удаляются. Как правило, можно устанавливать порог встречаемости слова при котором слово можно считать редким и до которого все редко встречающиеся слова будут удалены. Для того, чтобы снизить вычислительную сложность алгоритма, строится дерево Хаффмана (англ. Huffman Binary Tree). Далее алгоритм проходит заранее заданным размером окна по выбранному отрезку текста. Размер окна задается как параметр алгоритма. Под окном подразумевается максимальная дистанция между текущим и предсказываемым словом в предложении. То есть если окно равно трем, то для предложения «Я смотрел хороший фильм» применение алгоритма Skip-gram будет проходит внутри блока, состоящего из трах слов: «Я смотрел хороший», «смотрел хороший фильм». Далее применяется нейросеть прямого распространения (англ. Feedforward Neural Network) с много переменной логистической функцией.

Схематически модель Skip-gramm представляется в виде нейронной сети (Рис. 3) [23]:

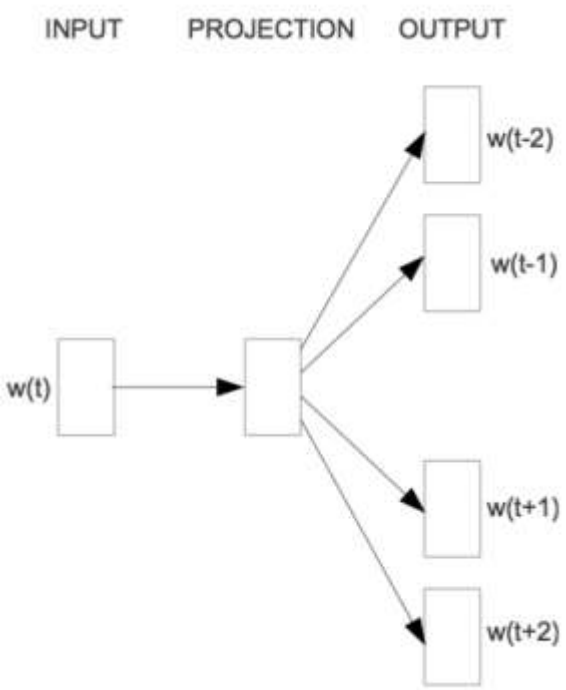

Рис. 3. Архитектура модели Skip-gram

Изображенная на Рис. 3 нейронная сеть состоит из трех слоев: входной (input), выходной (output) и скрытый (projection). Слово, подаваемое на вход, обозначено w(t), в выходном слове w(t-2), w(t-1), w(t+1) и w(t+2) – слова контекста которые пытается предсказать нейронная сеть. То есть модель skip-gram – предсказывает контекст при данном слове.

3.3.3. Использование модели Skip-Gramm для снижения зависимости от обучающей коллекции.

В работе [15] показано, что нейронные сети, с помощью векторных представлений слов, полученных при помощи алгоритма word2vec [29], могут эффективно решать задачи обработки текстов на естественном языке, в общем случае и задачу классификации текстов по тональности в частности. Описанный алгоритм показал лучшие результаты по сравнению с другими алгоритмами на выбранных текстовых коллекциях.

Для обучения модели Skip-Gramm произвольным образом было выбрано 5 миллионов текстов из первоначальной, не разделенной по классам тональности, коллекции 2013 года. Коллекции 2014 и 2015 годов в обучении не участвовали, так как делается предположение, что обученная модель должна быть переносима на более поздние коллекции.

В качестве программной реализации модели Skip-gram был использован Word2Vec [29] со следующими параметрами:

- size 300 – каждое слово представляется в виде вектора заданной размерности;
- window 5 – как много слов из контекста обучающий алгоритм должен принимать во внимание;
- negative 10 – число негативных примеров для негативного сэмплирования;
- sample 1e-4 – суб-сэмплирование, применение суб-сэмплирования улучшает производительность. Рекомендуемый параметр суб-сэмплирования от 1e-3 до 1e-5;
- threads 10 – количество используемых потоков;
- min-count 3 – ограничивает размер словаря для значимых слов. Слова, которые встречаются в тексте менее указанного количества раз, игнорируются. Стандартное значение – 5);
- iter 15 – количество обучающих итераций.

Одной из особенностей Word2Vec является то, что алгоритм не разделяет слово и следующий за ним знак препинания, поэтому, чтобы в файле-модели не было терминов со знаками препинания таких как: «например,», перед началом тренировки знаки препинания были отделены пробелом от идущего перед ними слова. Аналогично, чтобы «не + слово» не было разделено на два различных термина, пробел между частицами не и ни был заменен нижним подчеркиванием (напр. «ни_разу», «не_хотел»).

Из текстов, как и ранее, были отфильтрованы эмотиконы, так как они являются метками принадлежности текста к определенному классу тональности.

Каждый текст был представлен в виде усредненного вектора входящих в него слов (формула 9):

$$d = \frac{\sum w_i}{n}, \tag{9}$$

где w_i – векторное представление i-го слова, входящего в исследуемый текст, i=(1,..,n), n – число слов из лексикона, входящих в исследуемый текст.

Классификатор был обучен на коллекции 2013 года, далее обученная модель классификатора применялась для тестирования коллекций 2014 и 2015 годов. Результаты классификатора представлены в Таблица 7, результаты метрик качества для коллекции 2013 года оставлены для наглядности.

Таблица 7 Результаты классификации текстов по тональности с использованием векторов слов, полученных при использовании Word2Vec в качестве признаков

	Acc.	**Precision**	**Recall**	**F-мера**
2013	0,7206	0,7250	0,7221	0,7226
2014	0,7756	0,7763	0,7836	0,7787
2015	0,7289	0,7250	0,7317	0,7252

Рис. 4 наглядно показывает, что качество классификации на три класса не только не снижается на коллекциях, собранных с разницей полгода/год, но и держится на уровне лучших значений, полученных при использовании модели мешка слов при перекрестной проверке на коллекции одного года (Таблица 2, Таблица 4). При том, что число координат в векторе слова ровно 300 (задаваемый параметр), а не превосходит 200 тысяч, как в булевской или векторной моделях.

Данный метод хорошо подходит для применения в том случае, если у нас есть внешняя достаточно представительная коллекция текстов, которая схожа по лексике с обучающей и тестовой коллекциями, то есть здесь, как для других нейронных сетей требуется большая обучающая выборка текстов. Метод позволяет получить устойчивые и стабильные результаты.

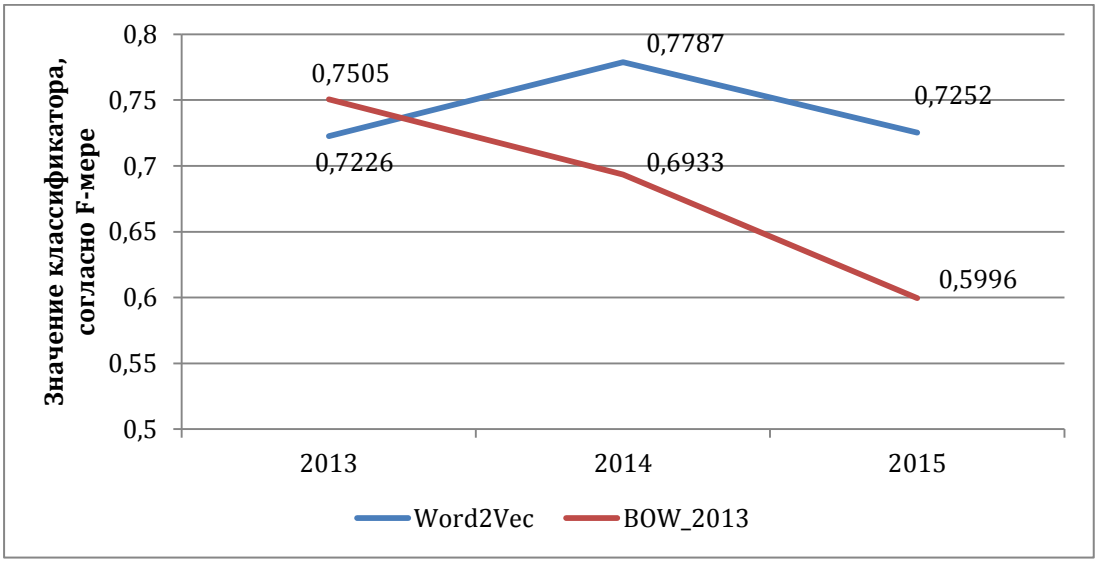

Рис. 4. Сравнение использование векторов слов Word2Vec в качестве признаков и лексикона, основанного на мешке слов коллекции 2013 года.

4. Заключение

В данной статье предложено три принципиально различных модели, позволяющие преодолеть ухудшение результатов классификации по тональности на коллекциях разнесенных во времени. В таблице 2 было показано, что качество классификации текстов по тональности согласно F-мере за полтора года может снизиться до 15%. Таким образом, цель предлагаемых в статье подходов – свести до минимума снижение качества классификации текстов коллекций, разнесенных во времени.

1. В качестве первого подхода, предлагается использовать весовую схему с линейной вычислительной сложностью. Таким образом можно динамически обновлять лексикон и переобучать классификатор. Зависимость от обучающей коллекции снижается потому, что обучающая коллекция постоянно обновляется. В этом случае, разница между работой классификатора на коллекции 2013 года и 2015 года составляет всего 2,4% согласно F-мере при использовании мешка слов и 1,42% при использовании меры TF-ICF. Несмотря на очевидные достоинства этого подхода, у него есть два недостатка:

    1. с обновлением лексикона, увеличивается размерность пространства признаков. Соответственно, с каждым обновлением лексикона, система требует больших ресурсов, вектор текста становится более разреженным.
    2. качество классификации с использованием меры TF-ICF существенно уступает качеству классификации при использовании мешка слов.

2. Второй подход основан на подключении словарей тональной лексики. Лексикон РуСентиЛекс и Linis-Crowd. Использование внешних словарей позволяет сократить разрыв в качестве классификатора между коллекциями 2013 года 2015 до 5,6% согласно F-мере. Разница согласно F-мере между результатами классификации коллекции 2013 и 2014 годов менее 1% и составляет всего 0,2%. При этом качество классификатора держится на уровне 0,68-0,73, что сопоставимо с лучшими результатами. Таким образом, порождение признаков на основе внешних словарей не влечет за собой масштабного увеличения пространства признаков и позволяет показывать хорошие результаты классификации. Несмотря на это, так как пространство признаков по прежнему зависит от обучающей коллекции, наблюдается незначительное снижение качества классификации на более поздних коллекциях.

3. В основе третьего подхода лежит идея пространства распределенных представлений слов и нейронная языковая модель Skip-gram. Как и во втором подходе, здесь были использованы внешние ресурсы. Пространство распределенных векторов слов строилось на

не размеченной коллекции твитов, которая в разы больше автоматически размеченной обучающей коллекции. В качестве признаков были использованы усредненные вектора слов, входящих в один твит. Таким образом, размерность векторного пространства составила всего 300 – это первое преимущество подхода. Вторым преимуществом подхода являются результаты классификации: разница между F-мерой 2013 и 2015 годами составляет 0,26%, при чем результаты классификации на коллекции 2015 года выше. Аналогично с результатами классификации на коллекциях 2013 и 2014 годов, результаты классификации на коллекции 2014 года превосходят на 5,6% результаты классификации на коллекции 2013 года согласно F-мере. Это можно объяснить тем, что для получения результатов на коллекции 2013 года использовался метод перекрестной проверки, то есть коллекция делилась на обучающую и тестовую в отношении 4:5, а при обучении классификатора для тестирования на коллекциях 2014 и 2015 годов использовалась полная коллекция 2013 года.

Таким образом, все три предложенных подхода позволяют снизить ухудшение результатов классификации по тональности на разнесенных во времени коллекциях.

Список литературы

1. Клековкина М. В., Котельников Е. В. Метод автоматической классификации тек-стов по тональности, основанный на словаре эмоциональной лексики //Труды конференции RCDL. – 2012. – С. 118-123.
2. Лукашевич Н. В., Четвёркин И. И. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса //Вычислительные методы и программирование. – 2011. – Т. 12. – №. 4. – С. 73-81.
3. Лукашевич Н., Рубцова Ю. Объектно-ориентированный анализ твитов по тональ-ности: результаты и проблемы // Труды Международной конференции DAMDID/RCDL-2015. — Обнинск, 2015. — С. 499–507.
4. Лукашевич, Н. В., Левчик, А. В., Loukachevitch, N. V., & Levchik, A. V. (2016). Создание лексикона оценочных слов русского языка РуСентиЛекс.
5. Мониторинг тональности твитов о ВУЗ'ах в режиме реального времени [Элек-тронный ресурс]. – Режим доступа: https://tweets-about-universities.herokuapp.com/ (Дата обращения: 10.09.2016).
6. Настроение России online [Электронный ресурс]. – Режим доступа: http://twittermood-ru.appspot.com/ (Дата обращения: 10.09.2016).
7. Рубцова Ю. В. Разработка и исследование предметно независимого классификатора текстов по тональности //Труды СПИИРАН. – 2014. – Т. 5. – №. 36. – С. 59-77.
8. Рубцова Ю.В. Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов // Электронные библиотеки: перспективные методы и технологии,

электронные коллекции: Труды XV Всероссийской научной конфе-ренции RCDL’2013, Ярославль, Россия, 14-17 октября 2013 г. – Ярославль: ЯрГУ, 2013. –С. 269-275.

9. Русначенко Н. Л., NL R. Улучшение качества тональной классификации с ис-пользованием лексиконов.
10. Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau, R. Sentiment analysis of twit-ter data //Proceedings of the Workshop on Languages in Social Media. – Association for Computational Linguistics, 2011. – С. 30-38.
11. Alexeeva, S., Koltsov, S., Koltsova O. (2015), Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media, Computational linguistics and computa-tional ontology, pp 25–34.
12. Fan R.-E. , Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J. LIBLINEAR: a Library for Large Linear Classification. J. of Machine Learning Research. 2008. vol. 9. pp. 1871–1874.
13. J. Read. Using emoticons to reduce dependency in machine learning techniques for sen-timent classification. In Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2005.
14. Joel W. Reed, Yu Jiao, Thomas E. Potok, Brian A. Klump, Mark T. Elmore, Ali R. Hurson. TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams. — Proc. Machine Learning and Applications, 2006, ICMLA '06, pp. 258–263.
15. Kim Y. Convolutional neural networks for sentence classification // arXiv preprint arXiv:1408.5882. – 2014.
16. Kouloumpis E., Wilson T., Moore J. Twitter sentiment analysis: The good the bad and the omg! //ICWSM. – 2011. – Т. 11. – С. 538-541.
17. Lek H. H., Poo D. C. C. Aspect-based Twitter sentiment classification //2013 IEEE 25th International Conference on Tools with Artificial Intelligence. – IEEE, 2013. – С. 366-373.
18. Levy, O. Improving Distributional Similarity with Lessons Learned from Word Em-beddings / O. Levy, Y. Goldberg, I. Dagan // Transactions of the Association for Com-putational Linguistics. – 2015. – P. 211–225.
19. Loukachevitch N., Rubtsova Y. Entity-Oriented Sentiment Analysis of Tweets: Results and Problems //Text, Speech, and Dialogue. – Springer International Publishing, 2015. – С. 551-559.
20. Loukachevitch, N., and Rubtsova, Y. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis. // In Proceedings of International Confer-ence on Computational Linguistics and Intellectual Technologies Dialog-2016. – 2016. – С. 375-384.
21. Manning C. D., Schutze H. Foundations of Statistical Natural Language Processing // The MIT Press, 1999.
22. Mansour R. et al. Revisiting The Old Kitchen Sink: Do We Need Sentiment Domain Adaptation? //RANLP. – 2013. – С. 420-427.
23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

24. Pak A., Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining //LREC. – 2010. – T. 10. – C. 1320-1326.
25. Pang B., Lee L. Thumbs up? Sentiment classification using machine learning techniques. Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia: ACL. 2002. pp. 79–86.
26. Saif Mohammad, Svetlana Kiritchenko S. and Xiaodan Zhu. 2013. NRC-Canada: Build-ing the state-of-the-art in sentiment analysis of tweets. In: Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13).
27. Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Computational Linguistics, 2011. 37(2): p. 267-307.
28. Titov, I. Modeling Online Reviews with Multi-grain Topic Models / I. Titov, R. McDonald // Proceedings of the 17th International Conference on World Wide Web (WWW'08). – 2008. – P. 111–120.
29. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Dis-tributed Representations of Words and Phrases and their Compositionality. In Proceed-ings of NIPS, 2013. – Pp. 3111-3119.
30. Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 417–424.
31. Wilson T., Wiebe J. and Hoffmann P. Recognizing contextual polarity in phrase level sentiment analysis. In Proc: of Human Languages Technologies Conference/ Conference on Emperical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancou-ver, CA, 2005.