УДК 81’33

Маслова Н. Е.

НЕЙРОННАЯ СЕТЬ DOC2VEC В ЗАДАЧЕ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТОВ СОЦИАЛЬНО-СЕТЕВОГО ДИСКУРСА (НА МАТЕРИАЛЕ РУССКОГО ЯЗЫКА)

Статья освещает подходы к решению задачи автоматического определения тональности русскоязычных текстов, входящих в состав полилогов социально-сетевого дискурса экологической направленности. В рамках подхода к машинному обучению с учителем создан программный продукт с помощью модуля нейронных сетей Doc2vec. Программа отвечает требованиям, диктуемым особенностями коротких неформальных текстов.
Ключевые слова: *определение тональности текста, социально-сетевой дискурс, машинное обучение с учителем.*

Маслова Н. Є. Нейронна мережа DOC2VEC в завданні автоматичного визначення тональності текстів соціально-мережевого дискурсу (на матеріалі російської мови). – Стаття.
Стаття висвітлює підходи до вирішення завдання автоматичного визначення тональності текстів російської мови, що входять до складу полілогів соціально-мережевого дискурсу екологічного спрямування. У межах підходу до машинного навчання з учителем створено програмний продукт за допомогою модуля нейронних мереж Doc2vec. Програма відповідає вимогам, які диктуються особливостями коротких неформальних текстів.
Ключові слова: *визначення тональності тексту, соціально-мережевий дискурс, машинне навчання з учителем.*

Maslova N. Ye. The neural network DOC2VEC in the automatic sentiment analysis of texts of social media discourse (based on the Russian language). – Article.
The article highlights approaches to the solution of an issue of automatic sentiment analysis of Russian texts included in polylogues of the social media discourse of ecological direction. Within the approach of supervised machine learning, with the help of a module of neural networks Doc2vec, there is created a software program. The software program satisfies the requirements determined by features of short informal texts.
Key words: *text sentiment analysis, social media discourse, supervised machine learning.*

Проблема автоматического определения тональности текста приобретает свою актуальность в рамках исследования общественного мнения (англ. *opinion mining*). Анализ тональности текста – сфера активных исследований последних десятилетий. На сегодняшний день нужны инструменты для автоматической обработки огромных объемов текстов. Для поиска решений с минимальным участием эксперта разработаны такие подходы: а) опирающиеся на эмотивную лексику (т. е. анализ по словарям и правилам); б) машинное обучение с учителем; в) машинное обучение без учителя. Словарный метод классификации тональности получил достаточно широкое применение. Он лег в основу таких аналитических систем мониторинга СМИ, как Интегрум[1], Медиалогия[2], IQBuzz[3], PalitrumLab[4], SemanticForce[5]. Его суть сводится к тому, что создаются словари эмотивных слов с заведомо определенной тональностью, и в исследуемом тексте определяется, слова из какого словаря (негативного или позитивного) в данном тексте преобладают. Как правило, анализ с помощью словарей осуществляется по определенным правилам (учитываются синтагматические границы, отмеченные знаками препинания, особо обрабатываются указанные в поиске словосочетания, принимаются во внимание диминутивы и аугментативы, а также отрицание). Кроме того, разработана разновидность этого метода, исходящая из установки, что не все слова играют равную роль в формировании тональности текста (теоретико-графовые модели). Такие модели выстраивают графу исследуемого текста, ранжируют его вершины, определяют «вес» каждого слова на основе тонального словаря и ранга вершины-слова [5].

Тем не менее, данный подход не лишен недостатков. Во-первых, экспрессивно-оценочную окраску слово получает, становясь частью высказывания, т. е. элементом речи. Пока слово представляет собой элемент языка, оно не несет экспрессивности, даже если принадлежит к кругу эмотивной лексики. Подробнее о разграничении понятий «эмотивность» и «экспрессивность» будет сказано ниже. Во-вторых, возникают трудности с полисемией, омонимией и идиоматичностью. В-третьих, такой метод оставляет систему нечувствительной к новообразованиям (окказионализмам). Наконец, создание словарей требует больших вложений человекоресурсов, глубокой работы экспертов-лингвистов.

Метод машинного обучения без учителя работает на основе того принципа, что термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов во всей коллекции, имеют в тексте наибольший вес. Выделив такие термины, а затем определив их тональность, можно сделать вывод о тональности всего текста.

[1] www.integrum.ru/.
[2] www.mlg.ru/.
[3] www.iqbuzz.pro/.
[4] www.palitrumlab.ru/.
[5] www.semanticforce.net/.

Данное исследование выполнено в рамках метода машинного обучения с учителем[6]. Выбор обусловлен в первую очередь особенностями эмпирического материала, к которому применяется алгоритм анализа. Это обсуждения экологических проблем России в интернет-сообществах, т. е. сегмент неинституционального социально-сетевого дискурса. Такой вид дискурса относится к неформальному, а значит, в нем велика доля экспрессивно-оценочных окказионализмов. Этот факт существенно снижает степень точности в работе словарного метода классификации тональности. Кроме того, высказывания на форумах являются частью макрополилога (термин Р.К. Потаповой) [3]. В силу этого большинство высказываний – сравнительно малые тексты (1–9 предложений), что не позволяет применять машинное обучение без учителя.

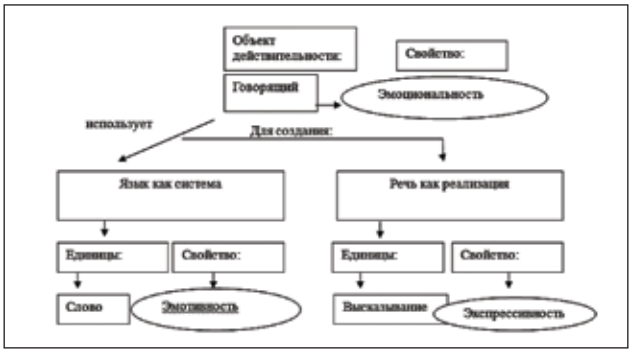

Рис. 1. Различие между терминами «эмоциональность», «эмотивность» и «экспрессивность»

В данном исследовании мы придерживаемся точки зрения Е.М. Вольф и М.М. Бахтина, согласно которой «экспрессивность рассматривается как свойство не отдельных слов, а высказывания в целом» [2, с. 37]. М.М. Бахтин выражает это следующим образом: «Экспрессивность – это конститутивный признак высказывания. В системе языка, то есть вне высказывания, экспрессивной интонации нет… Предложение как единица языка лишено экспрессивности» [1, с. 265]. Эмоциональность – параметр, описывающий состояние говорящего, в то время как эмотивностью обладают слова (т. е. элементы языка), которые описывают эмоции и оценки объектов действительности («радостный», «ужасный» и т. д.). Их можно назвать метаэкспрессивными – в том смысле, что они не несут экспрессивности как таковой, но описывают эмоциональность (категорию психических характеристик говорящего) или свидетельствуют об экспрессивности (характеристике высказывания в речи). Разграничение терминов *эмоциональный, эмотивный* и *экспрессивный* проиллюстрировано на рис. 1.

Разработка аналитических систем, базирующихся на словарях эмотивных слов, заведомо является непродуктивной, поскольку, во-первых, такая система может оказаться нечувствительной к смене указанного в словаре знака слова, употребленного в другом контексте, во-вторых, такая система не сможет зафиксировать экспрессивность высказывания, построенного из нейтральных слов.

Для решения исследовательской задачи был использован опыт последних лет. Ключевую роль в программе играет библиотека gensim языка python, разработанная Радимом Рекуреком [9] с применением модели Doc2vec, созданной Томасом Миколовым [8]. За основу был взят код, опубликованный в статье Михаила Черного «Современные методы анализа тональности текста» (авт. пер. с англ. «Modern Methods for Sentiment Analysis» by Michael Czerny – *Н. М.*) [6]. Также при создании продукта был использован опыт, изложенный в статье «Word Embeddings for Fun and Profit. Document classification with Gensim» («Векторы слов для забавы и пользы. Классификация документов с помощью библиотеки gensim» – авт. пер. – *Н. М.*) [11], опубликованной в качестве электронного ресурса на сайте github.com пользователем tmylk.

Прежде чем описывать алгоритм программы, необходимо вкратце объяснить суть метода Doc2vec. Более детальное описание можно найти в вышеупомянутой статье Т. Миколова «Распределенные векторы предложений и документов».

Модель Doc2vec представляет собой неглубокую (один скрытый слой) нейронную сеть. Общий принцип её работы сводится к тому, что к каждому слову модель строит свой вектор. При этом одно и то же слово в разных предложениях всей выборки имеет один и тот же вектор. На основе векторов слов модель строит векторы предложения, которые не повторяются в рамках одной выборки. При обработке тестовой выборки модель использует уже созданные вектора слов для формирования векторов тестовых предложений.

Модель Doc2vec представляет собой логическое развитие модели Word2vec и имеет два метода: distributed memory (далее – DM, распределенная память) и distributed bag of words (далее – DBOW, распределенный мешок слов). Метод DM прогнозирует слово по известным предшествующим словам и вектору абзаца (рисунок 2). Несмотря на то, что окно анализа пере-

[6] Исследование проводилось при поддержке Российского научного фонда (грант № 14-18-01059, руководитель проекта – Р.К. Потапова).

мещается по тексту, вектор абзаца не перемещается (отсюда название «распределенная память») и позволяет учесть порядок слов. Именно этим Doc2vec выгодно отличается от Word2vec: последняя не учитывает порядка слов, а потому пригодна для анализа тональности только очень коротких текстов (например, сообщений в Twitter) [10]. DBOW прогнозирует случайные группы слов в абзаце только на основании вектора абзаца (рисунок 3).

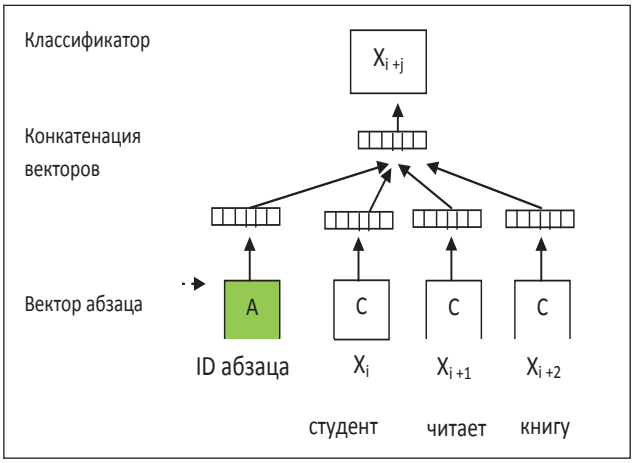

Рис. 2. Метод distributed memory (DM, распределенная память)

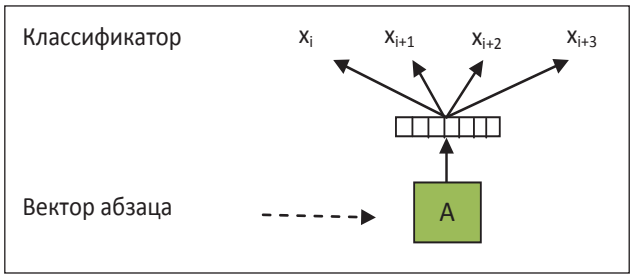

Рис. 3. Метод distributed bag of words (DBOW, распределенный мешок слов)

Эти векторы поступают в качестве параметров на вход SGD-классификатора, который выносит решение о «знаке» каждого высказывания. Таким образом, нейронные сети позволяют осуществлять анализ текста, не привязываясь к словарному значению слова (т. е. к языку как системе), а опираясь на реальное словоупотребление (речь как реализацию коммуникативной деятельности), что предельно важно.

Другим существенным преимуществом метода нейронных сетей является возможность быстрой обработки больших объемов текстов без предварительного анализа, причем увеличение объемов стимулирует увеличение точности классификации, хотя и не линейно. Эту тенденцию демонстрируют таблица 1 и рисунок 4.

Таблица 1

Зависимость показателей точности от объема высказываний

Общее число высказываний, n	Среднее значение точности, %
0	0
100	50
250	83
1715	85

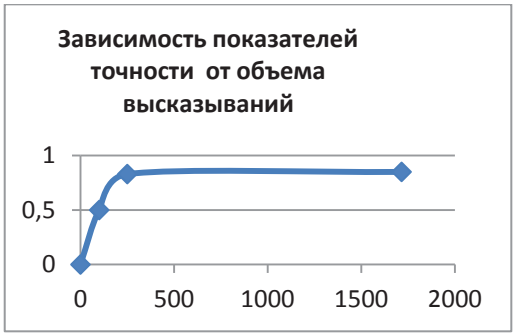

Рис. 4. Зависимость показателей точности от объема высказываний

Созданная программа уникальна тем, что позволяет учитывать параметры социально-сетевого дискурса (далее – ССД), разработанные Р.К. Потаповой [3, 4]. Все параметры высказывания делятся на четыре группы: а) **метаязыковые**, или паспорт реплики, – сайт, тема-стимул, дата, автор; б) **типы ССД по содержанию**: тональность высказывания, вид депривации (личностная – стратифицирующая): монотематический – политематический, информационно ненасыщенный (низкоконтекстуальный) – информационно насыщенный (высококонтекстуальный); не провоцирующий на полемику, конкретные действия, поступки – провоцирующий на полемику, конкретные действия, поступки; в) **типы ССД по форме:** тип дискурса (дистантный – прямой), канал связи (прямой – опосредованный), в реальном времени (online) – отложенный (offline), одновекторный – поливекторный (обращение к одному или нескольким адресантам), монохронный – полихронный (есть ли существенная временная дистанция между текстом-стимулом и текстом-реакцией); г) **типы ССД по функции:** информирующий – побуждающий с определенной целевой установкой к совершению конкретных действий, поступков (в частности, деструктивных, реализующийся по схеме «стимул → прагматическая реакция в виде конкретного деструктивного действия»); рассчитанный на целевую ограниченную группу пользователей или на неограниченное число пользователей.

Программа получает на вход xlsx файл, который представляет собой аннотированную базу дан-

ных с указанием паспорта реплики и параметров, на основе определенных показателей формирует списки высказываний. Эти списки получают соответствующие ярлыки, необходимые впоследствии для оценки эффективности классификатора. Затем выборка разделяется на обучающую и тестовую группы. Модель Doc2vec проходит обучение (формирует словарь выборки, строит векторы высказываний, сравнивает их с ярлыками), а затем применяет выработанные векторы на тестовой выборке. Классификация может осуществляться с помощью разных техник; в данном исследовании были взяты для сравнения логистическая регрессия и метод k средних соседей. Далее осуществляется оценка эффективности работы классификатора. Это также можно выполнить разными способами, в данном исследовании был использована матрица неточностей (confusion matrix). Программный продукт позволяет сравнивать воздействие разных пар параметров на эффективность обучения модели Doc2vec.

Первым параметром, на основе которого формируются все списки, является тональность высказывания. В зависимости от значения параметра каждое высказывание относится к списку положительных или отрицательных (они формируют обучающую выборку). Если предложение отрицательное, то оно выражает депривированное состояние говорящего. Поэтому вторым параметром был взят тип депривации. При разработке этого программного продукта впервые формулируется **классификация видов депривации**. Депривация может быть *личностной* (нацеленной на другого человека или на самого себя) или *стратифицирующей* (когда депривацию вызывает состояние человека как члена какой-то группы, занимающего определенную ячейку общества). Для лингвистов имеет значение тот факт, что указанное разделение влияет на результат речевой деятельности коммуникантов. Так, в полуформальном общении высказывания стратифицирующей депривации чаще будут формального стиля, а личностной – неформального. Высказывания первого типа могут чаще сопровождаться призывами к изменению существующего положения в обществе. Последние также могут тяготеть к такой тактике, как переход на личности, употребление слов сниженного регистра и т. д. При решении задачи классификации тональности текстов в рамках проблемы мониторинга общественного мнения в фокусе внимания находится именно стратифицирующая депривация как отражающая проблемы, более существенные для социума.

Как уже упоминалось выше, были опробованы два классификатора: логистическая регрессия и метод k средних соседей. Наибольшая точность классификации была получена с помощью логистической регрессии (98%). Второй метод достиг уровня точности в 70%. Результаты анализов показали, что метод k средних соседей наиболее эффективен для определения отрицательных предложений с личностным типом депривации, а метод логистической регрессии – для определения стратифицирующего типа депривации (см. таблицу 2).

Для визуализации работы классификаторов была использована матрица неточностей (см. таблицу 2), поскольку она оптимально показывает распределение верно и неверно определенных классов высказываний. Матрица неточностей – это таблица размером n × n, в которой ряды представляют собой истинное распределение экземпляров по классам, а колонки – прогнозируемое (см. рисунок 5). Тогда количество экземпляров, для которых программа правильно определила класс, отображается по «главной» диагонали (от левого верхнего угла к правому нижнему). В таблице 2 эти ячейки выделены цветом.

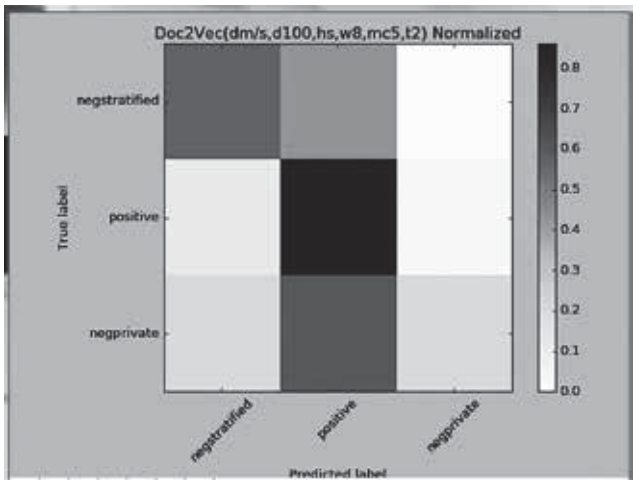

Рис. 5. Матрица неточностей для визуализации работы классификатора

Таблица 2

Сравнение двух классификаторов с помощью матриц неточностей (ЛД – личностная депривация, СД – стратифицирующая депривация)

Логистическая регрессия, %					**Метод k средних соседей, %**			
	Класс 1	**Класс 2**	**Класс 3**			**Класс 1**	**Класс 2**	**Класс 3**
Положительный	**25**	75	0		Положительный	**25**	75	0
Отрицательный, ЛД	0	**0**	100		Отрицательный, ЛД	43	**43**	14
Отрицательный, СД	2	1	**98**		Отрицательный, СД	16	14	**70**

Первоначально весь текст приводился к нижнему регистру (word.lower()). Однако, как нетрудно предположить, верхний регистр очень информативен при определении экспрессивности высказывания. На это также указывают некоторые исследователи. Действительно, при сохранении регистра без изменений замечается увеличение точности классификации именно положительных текстов и негативных текстов со стратифицирующей депривацией (что иллюстрирует рисунок 6).

Таблица 3

Влияние сленга русофобов на эффективность модели Doc2vec

Логистическая регрессия			
	Класс 1	**Класс 2**	**Класс 3**
Положительный	25	60	20
Отрицательный, ЛД	43	0	0
Отрицательный, СД	10	4	98

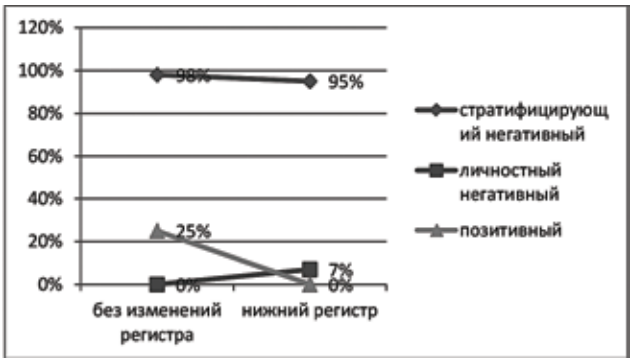

Рис. 6. Влияние регистра на качество классификации

Кроме того, получила подтверждение гипотеза о положительном воздействии присутствия «сленга русофобов» на качество классификации. «Сленг русофобов» – это еще одно достижение настоящего исследования. Анализ форумов экологической направленности оппозиционного характера позволил выделить некоторые тенденции словоупотребления, выражающиеся не только на лексическом уровне (определенные идеологические штампы), но также на графическом и морфологическом. Особенностью «сленга русфобов» является тематическая кроссдоменность – ее элементы встречаются в текстах совершенно различных тематик. Таким образом, составляющие «сленга русофобов» являются надежными маркерами отрицательной тональности текстов. Данные представлены в таблице 3.

Также было проверено, способствует ли классификации представление знаков препинания как отдельных слов. Для этого знаки препинания окружаются пробелами. И хотя другие авторы отмечают улучшение классификации при выполнении такой процедуры [7], только для определения негативных текстов с личностным типом депривации это можно считать оправданным (увеличение количества ~15%), в то время как для остальных типов текстов это приводило только к снижению показателей (см. рисунок 7).

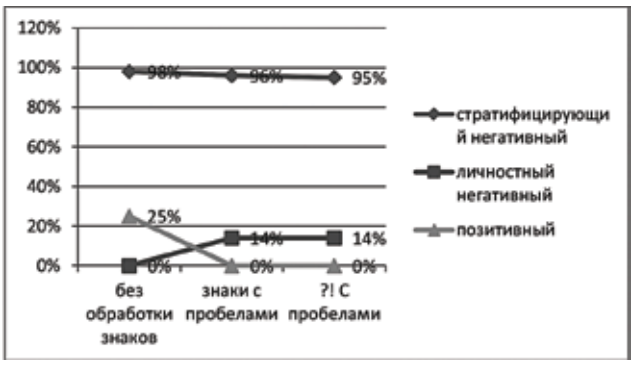

Рис. 7. Роль знаков препинания в классификации текстов

В дальнейшем планируется перебрать все параметры на созданной модели и выявить, при каких параметрах на одной и той же выборке достигается максимальный результат классификации. Это позволит математически проверить вес каждой пары признаков в создании негативной тональности высказывания. Другими словами, был разработан уникальный автоматический механизм исследования влияния различных аспектов текста на его тональность.

Таким образом, делаем вывод о том, что разработанный программный продукт имеет особую практическую значимость благодаря таким новейшим разработкам, как: а) параметры социально-сетевого дискурса, сформулированные Р.К. Потаповой; б) классификация видов депривации на личностную и стратифицирующую; в) разделение терминов «эмоциональный», «эмотивный» и «экспрессивный»; г) «сленг русофобов», встречающийся в текстах различных тематик и надежно маркирующий высказывания как отрицательные.

Исследование проводилось при поддержке Российского научного фонда (грант № 14-18-01059, руководитель проекта – Р.К. Потапова).

Литература

1. Бахтин М.М. Эстетика словесного творчества / М.М. Бахтин. – М. : Искусство, 1979. – 423 с.
2. Вольф Е.М. Функциональная семантика оценки / Е.М. Вольф. – М. : Едиториал УРСС, 2002 – 280 с.
3. Потапова Р.К. Социально-сетевой дискурс как объект междисциплинарного исследования / Р.К. Потапова // Материалы 2-й международной конференции «Дискурс как социально-сетевая деятельность». – М. : МГЛУ, 2014 – С. 20–32

4. Потапова Р.К. Депривация как базовый механизм вербального и паравербального поведения человека (на материале социально-сетевой коммуникации) / Р.К. Потапова // Речевая коммуникация в информационном пространстве. – М. : ЛЕНАНД, 2017, – С. 17–36.

5. Усталов Д.А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей (рус.) / Д.А. Усталов // УРФУ, Екатеринбург, Россия: конференция. – 2012.

6. Czerny M. Modern Methods for Sentiment Analysis / M. Czerny [Электронный ресурс].– Режим доступа : https://districtdatalabs.silvrback.com/modern-methods-for-sentiment-analysis#disqus_thread.

7. Liang H., Fothergill R., Baldwin T. RoseMerry: A Baseline Message-level Sentiment Classification System // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 551–555 [Электронный ресурс].– Режим доступа : http://alt.qcri.org/semeval2015/cdrom/pdf/SemEval092.pdf.

8. Mikolov T. Le Q. Distributed Representations of Sentences and Documents // Proceedings of the 31-st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

9. Rehurek R., Sojka P. Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, 2010, pp. 45–50 [Электронный ресурс].– Режим доступа : https://github.com/RaRe-Technologies/gensim#citing-gensim.

10. Tang D., Wei F., Yang N., Zhou M., Liu T., Qin B. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification // Proceeding of the 52th Annual Meeting of the Association for Computational Linguistics, ACL, 2014, – pp. 1155–1166. [Электронный ресурс].– Режим доступа : http://anthology.aclweb.org/P/P14/P14-1146.pdf.

11. Word Embeddings for Fun and Profit: Document classification with Gensim [Электронный ресурс].– Режим доступа : https://github.com/RaRe-Technologies/movie-plots-by-genre/blob/5a2d9157f9bf1bf908794051597b7851333dcfca/ipynb_with_output/Document%20classification%20with%20word%20embeddings%20tutorial%20-%20with%20output.ipynb#L1403.