Тимощук А.А. Способ многоуровневой грануляции текста для проведения автоматического определения тональности текста // Международный журнал информационных технологий и энергоэффективности. – 2017. – Т.2 №1(3) с. 8-15

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.896

# СПОСОБ МНОГОУРОВНЕВОЙ ГРАНУЛЯЦИИ ТЕКСТА ДЛЯ ПРОВЕДЕНИЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ТОНАЛЬНОСТИ ТЕКСТА

**Тимощук А.А.**

*Филиал федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет МЭИ» в г. Смоленске, Россия (214013, г. Смоленск, Энергетический проезд, дом 1); email: timoshchuk.anastasia@mail.ru*

**Статья посвящена проблеме автоматического анализа тональности текста. Предлагается новый способ многоуровневой грануляции текста для автоматического определения тональности, комбинирующий результаты работы классификатора на основе векторного представления текста (Doc2Vec) и метода ключевых слов. Doc2Vec - алгоритм обучения без учителя, учится получать распределенные векторы для частей текстов. Метод ключевых слов основан на подсчете весов входящих в текст признаков.**
**Способ многоуровневой грануляция текста включает две фазы – обучение и распознавание. На первой фазе на основе размеченной коллекции текстов происходит обучение Doc2Vec-классификатора и классификатора на базе ключевых слов. На второй фазе результаты распознавания нового текста обоими классификаторами объединяются и формируется итоговое решение; при этом учитываются степени уверенности классификаторов в своих результатах.**

Ключевые слова: анализ тональности текста, текстовая классификация, метод ключевых слов, машинное обучение, Doc2Vec.

# METHOD OF MULTILEVEL TEXT GRANULATION FOR TEXT TONALITY AUTOMATIC IDENTIFYING

**Timoshchuk A.A.**

*Smolensk Branch of Federal state budgetary educational institution of higher education "National research University Moscow power engineering Institute", Russia (214013, Smolensk, Energeticheski proezd, 1); e-mail: timoshchuk.anastasia@mail.ru*

**The article is devoted to the problem of automatic text sentiment analysis. The new method of multilevel text granulation for text tonality automatic identifying which combines the results of work of classifier based on vector text representation (Doc2Vec) and Keywords method is suggested in the article. Doc2Vec is an unsupervised learning algorithm, learning to receive distributed vectors for parts of texts. The Keywords method is based on the counting of weights of features which the text contains.**
**The method of multilevel text granulation contains two phases – training and recognition. On the first phase Doc2Vec -classifier and Keywords-classifier are trained on the labeled collection of texts. On the second phase the results of recognition by both classifiers for a new text are combined and the final decision is formed; confidence levels of both classifiers are taken into account.**

Keywords: text sentiment analysis, text classification, keywords method, machine learning, Doc2Vec.

Мнения занимают основное место почти во всех областях человеческой деятельности. Наши убеждения и представления о реальности и выбор, который мы делаем, в значительной мере зависит от того, как другие видят и оценивают мир. По этой причине, когда нам нужно принять решение, мы часто ищем чужие мнения. Это справедливо не только для людей, но и для организаций. Мнения и связанные с ними понятия, такие как чувства, оценки, отношения и эмоции являются предметом изучения анализа настроений (sentiment analysis). Зарождение и быстрое развитие этой области связано с интересами людей в Интернете, как правило, спрос всегда порождает предложение. В качестве примера можно привести различные отзывы, форумы, обсуждения, блоги, социальные сети. Впервые в человеческой истории имеется огромный объем мнений, записанных в цифровом формате. С начала XXI века сфера анализа тональности данных стала одной из наиболее активно развивающихся и исследуемых направлений в области обработки естественных языков.

Сентимент анализ (анализ тональности текста) – это обработка естественного языка, классифицирующая тексты по эмоциональной окраске. Такой анализ можно рассматривать как метод количественного описания качественных данных, с присвоением оценок настроения. Например, необходимо выявить автора текста – определить субъект, затем то, о чем ведется речь – объект разговора, и, наконец, отношение первого ко второму – определение тональности.

Формализация задачи сентимент анализа выглядит следующим образом. Пусть $X$ – множество всех документов, $Y$ – множество меток размера $N$, $\varphi: X \to Y$ – целевая функция определения тональности документов, значения которой известны только для обучающего подмножества $X_{train}=\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. Необходимо найти алгоритм $\alpha: X \to Y$, который способен классифицировать произвольный документ $x \in X$, расставить метки.

Задача определения эмоциональной окраски текста является задачей классификации, она может быть бинарной (негативный, позитивный), тернарной (негативный, нейтральный, позитивный) или n-арной (например, для n=5: сильно негативный, умеренно негативный, нейтральный, умеренно позитивный, сильно позитивный).

Все способы сентимент анализа можно разделить на три группы:

1. способ, основанный на словарях;
2. способ, основанный на и правилах;
3. машинное обучение.

Для получения более точных результатов классификации многие исследователи предлагают комбинировать результаты различных способов. Так в статье [2] объединяется метод опорных векторов (англ. SVM, support vector machine) и метод ключевых слов (англ. KWM, key words method). Во время экспериментов на тестовой коллекции (набор отзывов пользователей портала Imhonet.ru на различные фильмы) были получены следующие значения точности классификации (процент текстов, по которым классификатор принял правильное решение):

1. SVM-классификатор – 0,869%;
2. KWM-классификатор – 0,875%;
3. Комбинированный метод – 0,888%.

Результаты классификации комбинированным методом превосходят результаты, полученные SVM-классификатором и классификатором на основе ключевых слов. И это

подтверждает целесообразность использования комбинаций различных способ для проведения сентимент анализа текста.

В данной работе предлагается способ многоуровневой грануляция текста автоматического определения тональности текста, комбинирующий результаты работы классификатора на основе векторного представления текста и метода ключевых слов.

В качестве векторного классификатора предлагается использовать алгоритм Doc2Vec, предложенный в 2014 году Томасов Миколовым [7].

**Схема способа**

Способ многоуровневой грануляция текста объединяет результаты работы классификатора на основе векторного представления текста (Doc2Vec-классификатор) и классификатора на базе ключевых слов (KWM-классификатор). До начала применения предлагаемого способа необходимо задать дискретную шкалу для измерения тональности, содержащую *K* классов. Способ включает в себя две фазы – обучение и распознавание (классификацию) (рисунок 1). Для фазы обучения необходим обучающий текстовый набор данных, каждый текст в котором отнесен к одному из *K* классов.

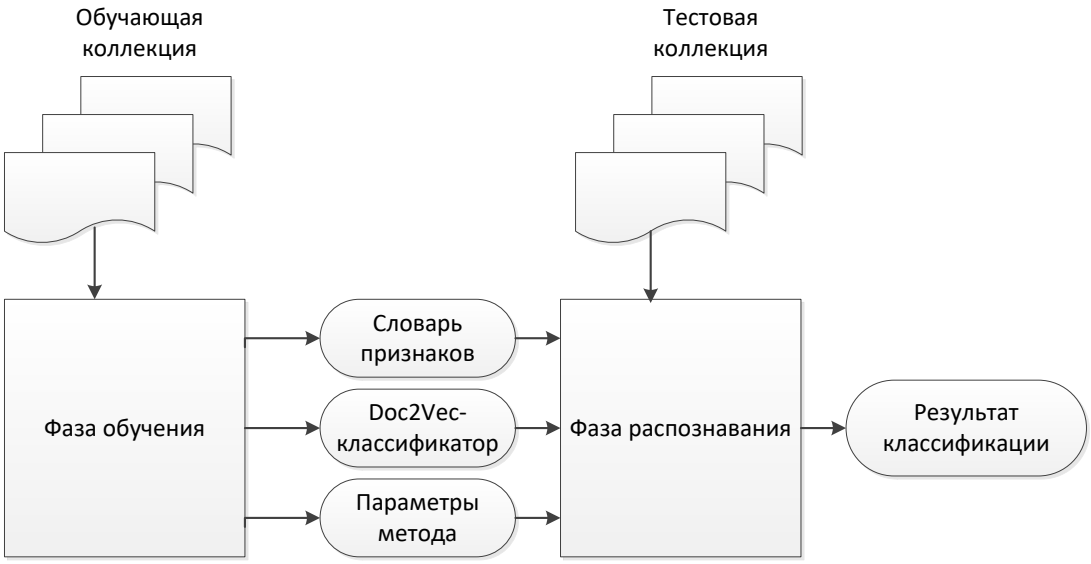


Рисунок 1 – Обобщенная схема способа многоуровневой грануляция документов для проведения автоматического определения тональности текста

Схема фазы обучения (рисунок 2) включает в себя следующие этапы [6]: предобработку текста, формирование словарей признаков (слов), взвешивание признаков методом ключевых слов, формирование векторной модели Doc2Vec, обучение векторного Doc2Vec-классификатора, настройку параметров для объединения результатов работы двух классификаторов.

Схема фазы распознавания (рисунок 3) также состоит из несколько этапов, первые два из которых совпадают с этапами предобработки и формирования векторной модели в фазе обучения.

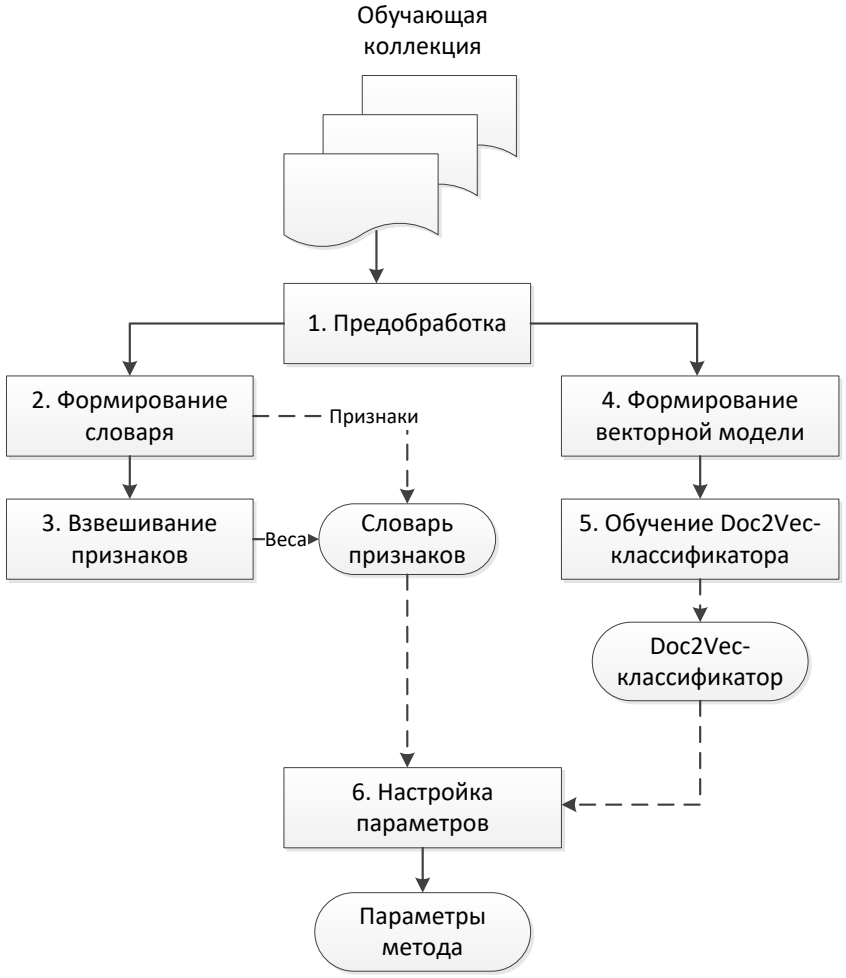


Рисунок 2 – Схема фазы обучения

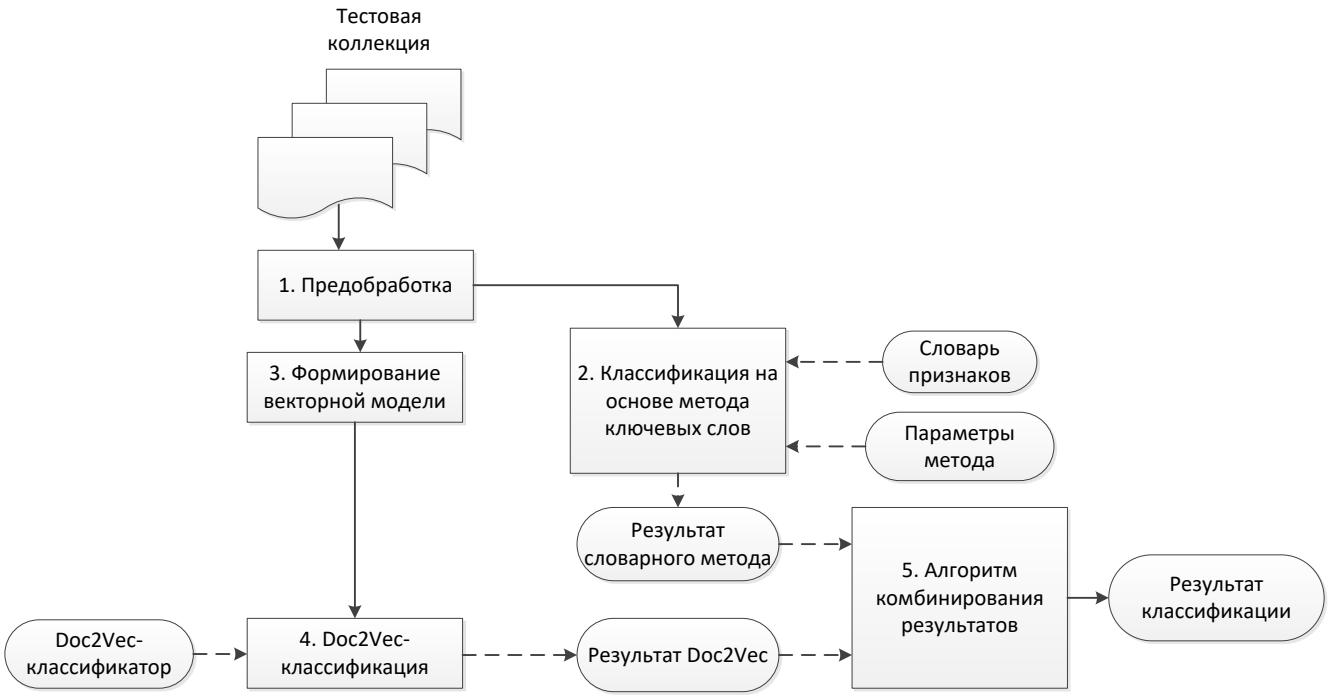


Рисунок 3 – Схема фазы распознавания

Затем для всех текстом из тестовой коллекции применяются классификации на базе метода ключевых слов и Doc2Vec, в результате чего каждый классификатор определяет к какому из *K* классов относится текст.

На завершающем этапе результаты классификации второго и третьего этапов объединяются, и полученные значения классификации составляют итоговый результат работы способа многоуровневой грануляция документов.

**Doc2Vec**

Многие методы машинного обучения требуют на вход данные, представленные в виде вектора признаков фиксированной длины. Когда дело доходит до текстов одним из самых распространенных способов представления векторов фиксированной длины является «мешок слов». Однако эта модель имеет много недостатков. Порядок слов теряется, разные предложения, имеющие одни и те же слова, могут иметь одинаковые представления. Алгоритм Doc2Vec решает эту проблему.

Алгоритм Doc2Vec (первоначальное название Paragraph Vector), алгоритм обучения без учителя, учится получать распределенные векторы для частей текстов [7]. Тексты могут быть переменной длины: от предложения до большого документа. Для простоты все входные тексты в данном разделе будут именоваться документами.

В данной модели векторные представления документов обучаются предсказывать слова в документе, точнее берется вектор документа и объединяется с несколькими векторами слов из него, и модель пытается предсказать следующее слово с учетом контекста. Векторы слов и документов обучаются с использованием метода стохастического градиентного спуска и метода обратного распространения ошибки. Векторы документов являются уникальными, а векторы одинаковых слов в разных документах совпадают.

Существует 2 архитектуры для построения векторных представлений документов:

1. DM

В данной конструкции каждый документ представлен уникальным вектором в виде столбца в матрице D, и каждый термин представлен уникальным вектором в виде столбца в матрице W. Вектор документ и векторы слов в нем объединяются или усредняются для предсказания следующего слова из контекста.

В этой архитектуре можно рассматривать токены как отдельные слова. Они действуют как память, которая помнит, что отсутствует в текущем контексте или теме документа. По этой причине модель называется Distributed Memory.

2. DBOW

Описанный выше метод рассматривает объединение вектора документа с векторами слов, входящих в него, для предсказания следующего слова в текстовом окне. Другой способ заключается в игнорировании слов из контекста на входе, но при этом модель должна предсказывать случайно отобранные слова для документа на выходе. Это означает, что на каждой итерации стохастического градиентного спуска просматривается текстовое окно, затем просматривается случайное слово в текстовом окне и формируется задача классификации с учетом вектора документа. Эта версия называется Distributed Bag-of-Words [7].

Для определения тональности текста Doc2Vec-классификатор сначала обучается на текстах неизвестной тональности для составления векторной модели, при этом создаются

экземпляры двух моделей Doc2Vec – DM и DBOW. Затем готовой модели передаются тексты с указанной тональностью, и модель обучается предсказывать тональность текста, используя логистическую регрессию. В результате таких действий получается готовая модель, которой на вход подается текст, а на выходе – вероятность принадлежности текста одному из *K* классов.

**Метод ключевых слов**

Идея метода ключевых слов заключается в том, что тональность текста определяется на основе подсчета весов входящих в него признаков (ключевых слов). Для вычисления весов признаков (слов) в данном методе будет использоваться RF (Relevance Frequency – релевантная частота), предложенная в [1, 3].

Обозначим *t* как количество текстов, содержащих *i*-й признак и принадлежащий классу *D*; *a* – это количество текстов, содержащих *i*-й признак и не принадлежащих классу *D*. Тогда вес *i*-го признака для класса *D* будет вычисляться формулой

$$RF_i^D = log_2\left(2 + \frac{t}{a}\right). \quad (1)$$

В процессе использования метода ключевых слов [5] для каждого класса *D* вес *i*-го признака $RF_i^D$ вычисляется независимо по формуле (1). Признаки с весами для всех классов сохраняются в словаре признаков.

Для каждого класса $D_l$ вычисляется вес текста $V_l$ путем простого суммирования весов входящих в текст признаков:

$$V_l = \sum_{i=1}^{K} v_i^l,\ l \in [1, \ldots, K], \quad (2)$$

где $v_i^l$ – вес *i*-го признака в тексте *T* по отношению к классу $D_l$.

Затем среди полученных весов находится максимальный:

$$V_{max} = max\{V_l\}, l = 1, \ldots, K. \quad (3)$$

Текст *T* относится к тому классу, для которого получен максимальный вес $V_{max}$.

**Алгоритм комбинирования результатов классификации**

После применения классификаторов будут получены следующие параметры:

- $K_{Doc2Vec}$ – класс, выданный Doc2Vec-классификатором на основе сформированной векторной модели текстов;
- $K_{KW}$ – класс, выданный KWM-классификатором;
- $Conf(K_{Doc2Vec})$ – степень уверенности Doc2Vec-классификатора;
- $Conf(K_{KW})$ – степень уверенности KWM-классификатора.

Алгоритм, объединяющий результаты двух классификаторов для текста *T* состоит из следующих шагов:

1. Текст *T* относится к классу *K*, если решения классификаторов совпадают, то есть $K_{Doc2Vec}=K_{KW}=K$;

2. Если один из классификаторов сильно уверен в своем решении, а другой слабо, то текст *T* относится к классу, который выдал классификатор с сильной уверенностью.

3. Если одновременно оба классификатора слабо или сильно уверены в своих решениях, но при этом решения не совпадают, выбирается решение Doc2Vec-классификатора, так как он выбран приоритетным.

В итоге данный алгоритм позволяет решать конфликты двух классификаторов, используя оптимальные параметры, найденные на фазе обучения.

**Тестирование**

Для обоснования целесообразности использования векторного Doc2Vec-классификатора сравним его по точности классификации текстов с векторным SVM-классификатором.

Тестирование и сравнительный анализ выбранных способов автоматического сентимент анализа текста производилось на набор данных IMDB, который часто используется в качестве ориентировочного набора обучающей и тестовой коллекции текстов для проведения сентимент анализа.

Набор данных состоит из 100 000 отзывов на различные фильмы, случайным образом взятых с сайта IMDB (крупнейший в мире веб-сайт о кинематографе). Одним из ключевых аспектов этого набора данных является то, что каждый отзыв на фильм имеет несколько предложений. 100 000 отзывов делятся на три набора данных: 25 000 размеченных текстов по тональности для обучения, 25 000 размеченных текстов по тональности для тестирования и 50 000 неразмеченных текстов по тональности для обучения.

В данном наборе данных задана дискретная шкала измерения тональности, содержащая два класса отзывов: положительный и отрицательный. Отзывы сбалансированы как в наборе для обучения, так и в тестовом наборе.

Для сравнения работы классификаторов будем использовать метрику – точность классификации (процент текстов, по которым классификатор принял правильное решение). Результаты тестирования представлены в таблице 1.

Таблица 1. Результаты работы классификаторов на наборе IMDB

| **Модель классификатора** | **Точность, %** |
|---|---|
| SVM | 86.95 |
| Doc2Vec | **90.24** |

Из таблицы 1 видно, что Doc2Vec-классификатор является более точным, что подтверждает целесообразность использования в качестве векторного классификатора.

Следовательно, на основе проведенных исследований предполагается, что способ многоуровневой грануляция текста для автоматического определения тональности текста, объединяющий результаты двух классификаторов (Doc2Vec-классификатора и KWM-классификатора), позволит добиться повышения точности классификации.

**Список литературы**

1. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по матер. ежегодн. Междунар. конф. «Диалог». 2012. № 11 (18). С. 753–762.
2. Котельников Е.В. Комбинированный метод автоматического определения тональности текста // Журнал Программные продукты и системы. 2012. № 3. С. 189–195.

3. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Annual International Conf. «Dialogue», CoLing&InTel, 2012, no. 11 (18), pp. 739–746.
4. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995, no. 2 (12), pp. 1137–1143.
5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management, 1988, Vol. 24, no. 5. pp. 513–523.
6. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 2002, Vol. 34, no. 1. pp. 1–47.
7. Tomas Mikolov, Quoc Le. Distributed Representations of Sentences and Documents. // In Proceedings of Workshop at The 31st International Conference on Machine Learning (ICML) – 2014.

**References**

1. Kotelnikov E.V., Klekovkina M.V., CoLing&InTel, 2012, no. 11 (18), pp. 753–762.
2. Kotelnikov E.V. Combined method of text tonality automatic identifying// Journal of Software products and systems, 2012, no. 3, pp. 189–195.
3. Chetviorkin I., Braslavskiy P., Loukachevitch N. Sentiment Analysis Track at ROMIP 2011 // Computational Linguistics and Intellectual Technologies: Annual International Conf. «Dialogue», CoLing&InTel, 2012, no. 11 (18), pp. 739–746.
4. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection // Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1995, no. 2 (12), pp. 1137–1143.
5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management, 1988, Vol. 24, no. 5. pp. 513–523.
6. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys, 2002, Vol. 34, no. 1. pp. 1–47.
7. Tomas Mikolov, Quoc Le. Distributed Representations of Sentences and Documents. // In Proceedings of Workshop at The 31st International Conference on Machine Learning (ICML) – 2014.