АВТОМАТИЧЕСКИЙ АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТОВ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Котельников Е. В. (kotelnikov.ev@gmail.com),
Клековкина М. В. (klekovkina.mv@gmail.com)

Вятский государственный гуманитарный университет,
Киров, Россия

В статье представлены методы автоматической обработки текстов и машиного обучения, использованные авторами для решения задачи анализа мнений в рамках семинара РОМИП-2011. Обсуждаются вопросы выбора оптимального варианта векторной модели представления текстов и наиболее подходящего метода машинного обучения. Рассматриваются варианты построения векторной модели на основе подхода TF.IDF без использования обучающей информации о принадлежности текста тому или иному классу (unsupervised TF.IDF) и с использованием этой информации (supervised TF.IDF). Приведены данные о результатах применения следующих методов машинного обучения: наивного байесовского классификатора, метода Rocchio, метода k ближайших соседей, машин опорных векторов (SVM), метода на основе ключевых слов и его комбинации с SVM. Эксперименты показали, что наилучшие результаты показывает бинарная модель с косинусной нормализацией без обучения и метод, комбинирующий использование ключевых слов и SVM. Результаты экспериментов приводятся и анализируются в статье в сравнении с результатами другими участниками РОМИП-2011.

Ключевые слова: анализ тональности, машинное обучение, метод опорных векторов, метод Байеса

SENTIMENT ANALYSIS OF TEXTS BASED ON MACHINE LEARNING METHODS

Kotelnikov E. V. (kotelnikov.ev@gmail.com),
Klekovkina M. V. (klekovkina.mv@gmail.com)

Vyatka State University of Humanities, Kirov, Russian Federation

In this article the authors present the methods of text processing and machine learning which they used to fulfill the tasks of the tracks for the sentiment analysis on the seminar ROMIP-2011.
The questions of the choice of the optimal variant of text vector model and the most suitable machine learning method are raised.
Unsupervised and supervised TF.IDF methods of text representation are used. The authors apply such classification methods as: Naïve Bayes, Rocchio's method, k-Nearest Neighbors, Support Vector Machines (SVM), the method based on keywords and the method which combines SVM and keywords' method.
The experiments proved that the best way of text representation is unsupervised binary model with cosine normalization. The combination of SVM and keywords' method showed the best results for classification.
The authors give the analysis of the results in comparison with other participants of ROMIP-2011.

Key words: sentiment analysis, machine learning, support vector machines, Naïve Bayes

1. Введение

Автоматическая классификация текстов по тональности (анализ мнений, sentiment analysis) становится все более важной задачей, как с теоретической, так и с прикладной точек зрения [11]. На семинаре РОМИП-2011 впервые были предложены дорожки анализа отзывов пользователей по трем группам товаров — цифровые фотокамеры, книги и фильмы. Требовалось построить классификаторы для трех шкал оценок: двухбалльной, трехбалльной и пятибалльной.

Целью нашего участия в РОМИП-2011 являлось тестирование и сравнение, во-первых, различных подходов к представлению текста в рамках векторной модели, во-вторых, нескольких методов машинного обучения, в том числе метода опорных векторов (Support vector machine, SVM), наивного байесовского классификатора, метода классификации на основе ключевых слов и его комбинации с SVM.

В начале исследования мы ставили перед собой следующие вопросы:

1. Какой вариант векторной модели лучше подходит для решения задачи анализа мнений?
2. Какой метод машинного обучения лучше подходит для решения задачи анализа мнений?

3. Каким образом влияет размер оценочной шкалы (количество классов) на качество классификации?
4. Влияет ли тематика отзывов на качество классификации?

Для оценки качества классификации в процессе исследования использовались различные наборы данных. До того момента, когда тестовые данные, размеченные экспертами РОМИП, стали доступны, мы применяли скользящий контроль (cross-validation) на обучающих данных, предоставленных организаторами, и использовали подмножество тестовых данных, размеченных нами самостоятельно (по 100 отзывов по каждой группе товаров). После получения отзывов с экспертными оценками мы проверяли на них предварительные результаты — степень совпадения оказалась очень высокой.

Статья состоит из следующих разделов: в разделе 2 приводятся сведения о предварительной обработке текстов, в разделе 3 обсуждаются итоги исследования различных способов построения векторной модели текста. Раздел 4 посвящен используемым методам машинного обучения. В разделе 5 результаты экспериментов анализируются и сравниваются с результатами других участников. В разделе 6 обсуждаются выводы, сделанные на основе проведенных исследований, и направления дальнейшей работы.

2. Предварительная обработка

В наших исследованиях все используемые тексты подвергались единообразной предобработке. Из каждого текста исключались англоязычные и русскоязычные «стоп-слова» (частицы, предлоги, местоимения), удалялись слова длиной менее трех символов. Все слова преобразовывались к словарной форме (лемме) при помощи морфологического анализатора *mystem* от компании Яндекс. При этом из рассмотрения исключались все леммы, которые встречались менее чем в трех документах.

Полученная совокупность лемм обучающей коллекции составляет множество признаков для методов классификации и формирует словарь коллекции. Кроме лемм, в качестве признаков в словарь были добавлены различные варианты положительных и отрицательных смайликов — графических символов эмоционального отношения.

3. Векторная модель текста

Для ответа на первый вопрос («какой вариант векторной модели лучше подходит для решения задачи анализа мнений?») использовались два подхода к построению векторной модели — без использования обучающей информации о принадлежности текста тому или иному классу (*unsupervised*) и с использованием этой информации (*supervised*) [2].

В обоих подходах вес слова в тексте определяется по схеме *TF.IDF* [13]:

$$t_{ik} = L_{ik} \cdot G_i \cdot D_k \quad (1)$$

где t_{ik} — вес i-го термина в k-м документе,

L_{ik} — локальный вес i-го термина в k-м документе, отражающий значимость термина для данного документа,

G_i — глобальный вес i-го термина, отражающий значимость термина для всей коллекции,

D_k — нормализация для k-го документа.

Выражение (1) задает общую схему взвешивания, при подстановке в которую формул для всех трех компонентов получаются конкретные схемы вычисления весов. Для unsupervised TF.IDF мы исследовали следующие варианты [1]:

1) для локального веса: бинарный (BNRY), частотный (FREQ), логарифм частоты (LOGA);
2) для глобального веса: константный единичный (ONE), инвертированная документная частота (IDF), глобальный частотный IDF (GFIDF), логарифм GFIDF (IGFL). Кроме того, исследовался вариант вычисления глобального веса по методу TextRank [10];
3) для нормализации: отсутствие нормализации (NONE), косинусная нормализация (COSN).

Всего для unsupervised TF.IDF было протестировано $3 \times 5 \times 2 = 30$ способов вычисления весов терминов и получены следующие результаты (на основе метрики *macro F1* для бинарной классификации методом опорных векторов):

1) для разных групп товаров лучшими оказались разные способы вычисления локального веса: для фотокамер — FREQ, для фильмов — BNRY, для книг — LOGA и BNRY, причем для фотокамер отличие BNRY от FREQ не превышало 1 %;
2) во всех случаях лучшие результаты показал метод вычисления глобального веса ONE (присвоение всем терминам единичного глобального веса);
3) во всех случаях оказалось эффективнее вычислять косинусную нормализацию, чем обходиться без неё.

Для подхода supervised TF.IDF был выбран метод TF.RF, показавший по данным [6] наилучшие результаты в задаче тематической классификации. При этом в качестве локальных весов использовались методы взвешивания BNRY, FREQ и LOGA, осуществлялась косинусная нормализация, а глобальный вес подсчитывался по методу RF, предложенном в [5].

В методе RF (Relevance Frequency — релевантная частота) для вычисления глобального веса термина используется информация о распределении этого термина по документам обучающей коллекции с учетом принадлежности документов к классам.

Обозначим a — количество документов, содержащих i-й термин и относящихся к классу C, b — количество документов, содержащих термин

и не относящихся к классу *C*. Тогда, значимость *i*-го термина для класса *C* будет выражаться формулой [6]:

$$RF_i^C = \log_2\left(2 + \frac{a}{\max(1, b)}\right) \quad (2)$$

Результаты экспериментов показали, что метод вычисления глобальных весов RF показывает сходную эффективность с методом ONE — лучшим для unsupervised TF.IDF, — иногда незначительно превосходя его. Однако вычислительная сложность метода RF (как и всех других supervised методов) делает его применение нецелесообразным.

Таким образом, ответом на наш первый вопрос будТаким образом, ответом на наш первый вопрос будет утверждение, что с точки зрения эффективности и вычислительной сложности в качестве схемы взвешивания выгоднее всего использовать схему BNRY×ONE×COSN, т. е. бинарную модель с косинусной нормализацией. Такой вывод согласуется с результатами, полученными в [12].

4. Методы классификации

Для классификации текстов использовались известные методы машинного обучения [14]: наивный байесовский классификатор [7], метод Rocchio [3], метод *k* ближайших соседей [9], метод опорных векторов [4]. Кроме того, тестировался метод на основе ключевых слов и его комбинация с SVM.

В ходе предварительного тестирования на основе скользящего контроля по обучающим данным и размеченных самостоятельно тестовых документов выяснилось, что *методы Rocchio* и *k ближайших соседей* показывают существенно худшие характеристики качества, чем остальные. Поэтому было решено не отправлять на централизованное тестирование результаты, полученные этими методами.

Наивный байесовский классификатор был реализован традиционным образом [7], с учетом предварительной обработки текстов (см. раздел 2).

В качестве реализации *метода опорных векторов* была выбрана библиотека LIBSVM [8]. Проводился выбор ядра и подбор оптимальных параметров. Наилучшие результаты показало линейное ядро с регулирующим параметром $C = 1$.

Для задач с тремя и пятью классами использовалась стратегия «один против всех», когда обучается *N* классификаторов, где *N* — количество классов. Если несколько классификаторов «узнавали» тестовый документ, для окончательного решения выбирался наиболее положительный класс (при этом учитывалось неравномерное распределение количества обучающих отзывов по классам со смещением в сторону положительных оценок).

В *методе на основе ключевых слов* применялся лексико-статистический анализ и для каждого класса составлялся свой список ключевых слов. С этой целью для каждого слова из словаря коллекции (составленного после предварительной обработки, рассмотренной в разделе 2) вычислялся

вес для каждого класса по методу RF (2). В список заносилось подмножество слов с наибольшим весом, пороговый вес определялся экспериментально, на основе скользящего контроля и метрики *macro F1*, отдельно для каждого класса.

Определение класса документа из тестовой коллекции осуществлялось следующим образом. Для каждого класса на основе его списка ключевых слов подсчитывается суммарный вес входящих в документ слов, таким образом, получался вес класса. Решение об отнесении документа к тому или иному классу принималось на основе сравнения весов классов.

Подобная идея реализована, например, в [12], но без вычисления весов и порогов отбора слов; также слова отбирались в список на основе простой частоты встречаемости в документах соответствующего класса.

В *методе, комбинирующем SVM и метод ключевых слов,* сначала независимо вычислялись гипотезы обоих методов об отнесении тестового документа к тому или иному классу. Итоговое решение в различных ситуациях вырабатывалось на основе следующей стратегии:

1) *ни один из методов не определил класс* — относили отзыв к наиболее положительному классу в данной задаче;
2) *класс определен только в одном из методов* — относили отзыв к данному классу;
3) *оба метода определили классы* — здесь возможны следующие варианты:
   - есть совпадение ответов одного из классификаторов SVM с ответами метода ключевых слов — относили отзыв к этому классу;
   - вес класса в методе ключевых слов превышал заданный порог (определенный эмпирически) — относили отзыв к этому классу;
   - ни одно из предыдущих условий не выполнялось — приписывали отзыву наиболее положительную оценку SVM.

5. Результаты экспериментов

Результаты тестирования методов для бинарной классификации представлены на рис. 1–3. Приведены метрики *macro F1* и *Accuracy* наших методов и нескольких лучших участников при схеме оценки AND. В большинстве случаев оценки по схеме OR не изменяют относительного расположения результатов.

Обозначения рассмотренных нами методов: *SVM* — метод опорных векторов, *KW (Keywords)* — метод ключевых слов, *Comb* — комбинированный метод, *NB (Naïve Bayes)* — наивный байесовский классификатор;

yyy-N — коды наших результатов, *xxx-N* — коды результатов других (лучших) участников.

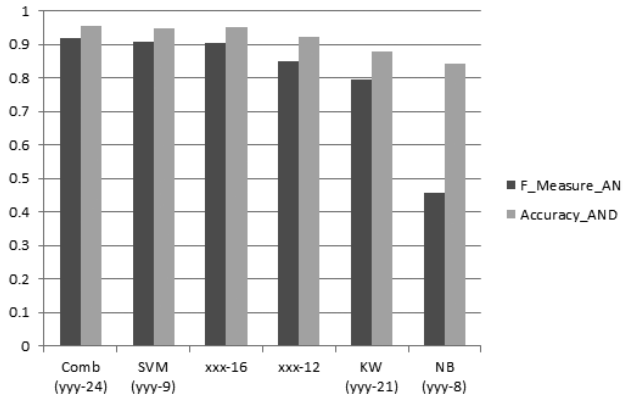

Рис. 1. Результаты классификации группы товаров «Фотокамеры» (AND)

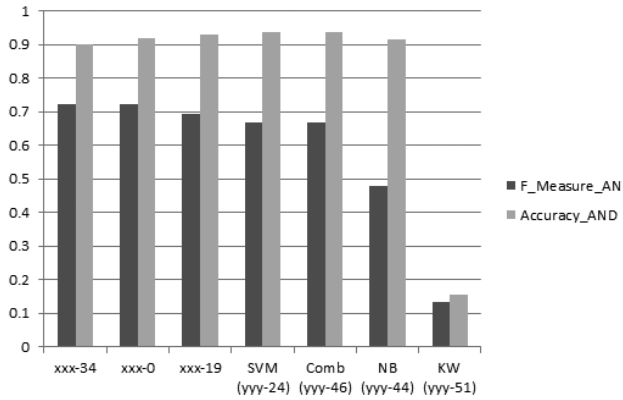

Рис. 2. Результаты классификации группы товаров «Книги» (AND)

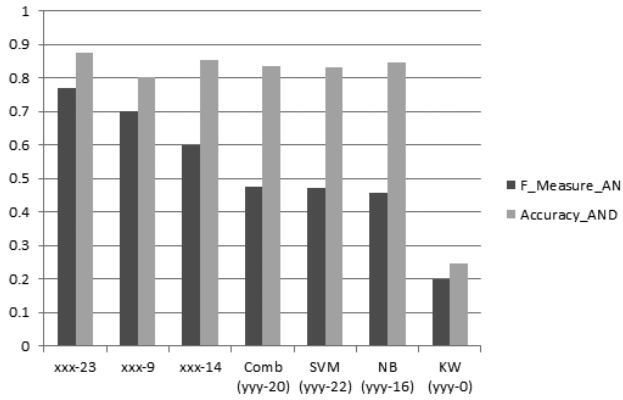

Рис. 3. Результаты классификации группы товаров «Фильмы» (AND)

Для задачи классификации с тремя и пятью классами результаты представлены в табл. 1 и 2. Приведены метрики *macro Precision*, *macro Recall macro F1* и *Accuracy* по схеме AND, обозначения аналогичны используемым на рисунках.

Таблица 1. Результаты классификации для трехбалльной шкалы (AND)

Run_ID	*Object*	*Macro_Prec*	*Macro_Rec*	*Macro_F1*	*Accuracy*
xxx-11	camera	0.745	**0.550**	**0.614**	0.787
xxx-3	camera	0.791	0.545	0.603	**0.812**
KW (yyy-12)	camera	0.753	0.514	0.574	0.778
Comb (yyy-6)	camera	**0.822**	0.515	0.566	0.797
SVM (yyy-1)	camera	0.590	0.377	0.412	0.720
xxx-43	book	**0.650**	**0.493**	**0.550**	**0.754**
xxx-3	book	0.641	0.492	0.536	0.715
Comb (yyy-37)	book	0.354	0.341	0.316	0.667
KW (yyy-47)	book	0.319	0.325	0.225	0.351
SVM (yyy-44)	book	0.232	0.293	0.259	0.636
xxx-10	film	**0.604**	**0.474**	**0.530**	**0.734**
xxx-19	film	0.598	0.471	0.527	**0.734**
Comb (yyy-4)	film	0.295	0.326	0.285	0.681
SVM (yyy-5)	film	0.233	0.309	0.265	0.662
KW (yyy-13)	film	0.300	0.285	0.206	0.312

Таблица 2. Результаты классификации для пятибалльной шкалы (AND)

Run_ID	*Object*	*Macro_Prec*	*Macro_Rec*	*Macro_F1*	*Accuracy*
xxx-4	camera	**0.591**	**0.223**	**0.259**	**0.520**
xxx-7	camera	0.393	0.195	0.246	0.493
Comb (yyy-3)	camera	0.582	0.206	0.225	0.473
KW (yyy-1)	camera	0.546	0.192	0.223	0.459
SVM (yyy-5)	camera	0.237	0.102	0.103	0.311
xxx-7	book	**0.510**	**0.225**	**0.253**	**0.574**
xxx-4	book	0.468	0.219	0.247	0.564
Comb (yyy-8)	book	0.285	0.184	0.204	0.468
SVM (yyy-6)	book	0.156	0.070	0.097	0.319
KW (yyy-2)	book	0.194	0.134	0.090	0.229
xxx-1	film	**0.325**	**0.194**	**0.230**	**0.531**
xxx-5	film	**0.325**	**0.194**	**0.230**	**0.531**
Comb (yyy-8)	film	0.201	0.095	0.110	0.258
KW (yyy-7)	film	0.197	0.091	0.081	0.191
SVM (yyy-4)	film	0.171	0.027	0.044	0.113

Проанализируем результаты классификации для двухбалльной шкалы (см. рис. 1–3), а также для трехбалльной и пятибалльной шкал (см. табл. 1, 2).

1. Из приведенных диаграмм видно, что метод опорных векторов показывает высокие значения метрики *F1* (за исключением группы товаров «Фильмы») и *Accuracy* (лучший результат для группы товаров «Книги»).

Для количества классов больше двух результаты метода опорных векторов существенно снижаются и он оказывается примерно в середине таблицы участников.

2. Наивный байесовский классификатор во всех случаях двухклассовой классификации показал низкие результаты по *F1*, но сопоставимые с лучшими результаты по *Accuracy*.

По техническим причинам в многоклассовой классификации наивный байесовский классификатор не был задействован.

3. Метод ключевых слов в бинарной классификации почти всегда показывает плохие результаты по обеим метрикам (за исключением группы товаров «Фотокамеры»).

В многоклассовых задачах ситуация неоднозначная, иногда метод ненамного отстает от лидеров и имеет преимущество перед SVM, в других случаях оказывается внизу таблицы результатов.

4. Результаты комбинированного метода для бинарной классификации практически идентичны методу опорных векторов, но в некоторых случаях (группа товаров «Фотокамеры») помогает скомпенсировать ошибки SVM и за счет этого выходит на первое место.

В случае трехбалльной и пятибалльной шкал комбинированный метод всегда показывает существенно лучшие результаты, чем метод опорных векторов и метод ключевых слов, и для группы товаров «Фотокамеры» имеет незначительную разницу по сравнению с лидерами.

В целом можно сделать следующие выводы.

1. SVM и комбинированный метод имеют, как правило, высокую точность (Precision), но низкую полноту (Recall), что в целом дает не слишком хорошую метрику F1. В свою очередь, например, для бинарной классификации низкая полнота получается из-за плохого распознавания отрицательных примеров. Связано это, возможно, с гораздо меньшим объемом обучающей выборки для негативных отзывов.

2. При увеличении количества классов результаты классификации всех участников семинара серьезно ухудшаются (например, для фотокамер при переходе от двух классов к пяти лучший результат по *F1* снижается с 92% до 26%). С другой стороны и оценки экспертов оказываются гораздо сильнее несогласованными в случае количества классов больше двух. В [11, стр. 27] высказывается мнение, что в отличие от многоклассовой тематической классификации в задаче анализа тональности текста, возможно, следует использовать регрессионные методы.

3. Результаты классификации отзывов для различных видов товаров довольно сильно отличаются. В табл. 3 приведены максимальные и средние значения по всем участникам метрик Precision, Recall, F1 и Accuracy. Из таблицы видно, что классификация отзывов по фотокамерам оказалась существенно проще. Возможно, это отчасти связано с тем, что в отзывах по данному виду товаров отдельно выделяются преимущества и недостатки товара, что более четко его характеризует. Другие причины обсуждаются, например в [11, стр. 37].

Таблица 3. Максимальные и средние значения Precision, Recall, F1 и Accuracy для бинарной классификации (AND)

	Precision		Recall		F1		Accuracy	
Группа товаров	**Max**	**Avg**	**Max**	**Avg**	**Max**	**Avg**	**Max**	**Avg**
Фотокамеры	0,990	0,747	0,934	0,769	0,921	0,722	0,957	0,815
Книги	0,687	0,560	0,763	0,600	0,723	0,589	0,936	0,792
Фильмы	0,760	0,595	0,781	0,614	0,769	0,545	0,875	0,674

6. Заключение

Проведенное исследование позволило нам ответить на заданные в начале вопросы.

1. «Какой вариант векторной модели лучше подходит для решения задачи анализа мнений?» — бинарная модель с косинусной нормализацией без глобальных весов.
2. «Какой метод машинного обучения лучше подходит для решения задачи анализа мнений?» — среди исследованных нами методов наилучшие результаты показал метод, комбинирующий методы опорных векторов и ключевых слов.
3. «Каким образом влияет размер оценочной шкалы (количество классов) на качество классификации?» — при увеличении диапазона шкалы качество классификации существенно ухудшается.
4. «Влияет ли тематика отзывов на качество классификации?» — качество классификации в большой степени зависит от тематики отзывов.

В целом, наш первый опыт участия в семинаре РОМИП следует признать успешным: на предоставленных организаторами тестовых материалах удалось провести задуманное исследование, при централизованной оценке наши результаты по нескольким прогонам оказались на первом месте.

В дальнейшем предполагается совершенствовать рассмотренные методы за счет использования специализированных словарей эмоциональной лексики и применения других методов машинного обучения — регрессионного и структурированного вариантов SVM, Gradient boosting.

Хочется надеяться, что на будущих семинарах РОМИП проблема анализа тональности текста останется в центре внимания и в её рамках будут предложены новые интересные задачи.

Литература

1. *Chisholm E., Kolda T. G.* New term weighting formulas for the vector space method in information retrieval. Technical Report Number ORNL-TM-13756, Oak Ridge National Laboratory, Oak Ridge, TN, March 1999.
2. *Debole F., Sebastiani F.* Supervised term weighting for automated text categorization. Proceedings of the 2003 ACM symposium on Applied computing SAC 03, 2003, Vol. 138(Ml), pp. 784–788.
3. *Joachims T.* A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Proceedings of 14th International Conference on Machine Learning, Nashville, TN, 1997, pp. 143–151.
4. *Joachims T.* Text categorization with support vector machines: learning with many relevant features. Proceedings of 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 137–142.
5. *Lan M.* (2007) A New Term Weighting Method for Text Categorization. PhD Theses.
6. *Lan M., Tan C. L., Su J., Lu Y.* (2009), Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, no. 4, pp. 721–735.
7. *Lewis D. D.* Naive (Bayes) at forty: The independence assumption in information retrieval. Proceedings of 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 4–15.
8. *LIBSVM* — A Library for Support Vector Machines, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
9. *Masand B., Linoff G., Waltz D.* Classifying news stories using memory-based reasoning. Proceedings of SIGIR-92, 15th ACM International Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, 1992, pp. 59–65.
10. *Mihalcea R., Tarau P. Textrank*: Bringing order into texts. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004, pp. 404–411.
11. *Pang B., Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, no. 2, pp. 1–135.
12. *Pang B., Lee L., Vaithyanathan S.* Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp. 79–86.
13. *Salton G., Buckley C.* (1988), Term-weighting approaches in automatic text retrieval, Information Processing & Management, Vol. 24, no. 5, pp. 513–523.
14. *Sebastiani F.* (2002), Machine learning in automated text categorization ACM Computing Surveys, Vol. 34, no. 1, pp. 1–47.