Выявление ошибок разметки данных с помощью моделей классификации для небольших наборов данных

Ф.В.Краснов

***Аннотация*— Разметка данных для задач классификации – сложный процесс, сопровождаемый неизбежно возникающими ошибками. Ручная или автоматическая разметка текстов для классификации включает системные ошибки, которые могут быть выявлены с помощью системных подходов на основе статистики и моделей машинного обучения. Рассмотрены именно небольшие наборы данных, так как в них последствия ошибок разметки наиболее заметны. Однако в условиях небольших наборов данных, ввиду недостаточности примеров, возникает проблема разряженных распределений, препятствующая обучению моделей с высокой сложностью. Автор использует эффект переобученности модели для минимизации ограничений, накладываемых недостаточностью данных. В рамках исследования проведено несколько экспериментов. Эксперимент на большом публичном наборе данных показал, что при классификации коротких текстов переобученная модель способна выявлять ошибки разметки данных. В эксперименте с формированием фасетов на основании пользовательских описаний товаров определена взаимозависимость наличия ошибок определения класса и работой асессоров по разметке текстовых данных на основании разных правил. В силу своей переобученности модель классификации способна на выявление существенных ошибок, которые драматически влияют на инженерное применение машинного обучения в высоконагруженных интернет-системах. В результате исследования автор приводит методы и критерии достижения моделью состояния «продуктивной переобученности». Лучший результат по метрике f1-score weighted (98 %) показала модель классификации на основе EmbeddingBag.**

***Ключевые слова* — переобучение, энтропия, классификация.**

I.Введение

В настоящее время значительное количество текстовых данных размечается асессорами для реализации разнообразных задач – классификация, выявление сущностей, определение частей речи, ранжирование. Процесс ручной разметки, в силу своей важности и трудоемкости, вызывает большой интерес со стороны исследователей. Так, в работах [1, 2] предпринята попытка упростить определение меток класса за счет подсказок с использованием уже размеченных данных. Значительного успеха добились авторы исследований [3, 4], применившие предобученные языковые модели для определения меток классов.

Ручная разметка производится по определенным правилам, но существует «человеческий фактор», и зачастую правила понимаются по-разному. Одним из методов работы с «человеческим фактором» считается разметка одних и тех же данных несколькими асессорами с дальнейшей перекрестной проверкой. Но при включении в процесс дополнительных асессоров также появляются и новые разночтения правил разметки, и новые роли, такие как, «асессоры асессоров». Распространенным примером ошибки разметки является ситуация, когда, одинаковые по смыслу тексты могут быть отнесены к различным классам. Заметим, что для модели машинного обучения размеченные данные являются истинными. Модель машинного обучения будет пытаться аппроксимировать ошибочно-размеченные данные и, возможно, не придет к оптимуму, не сойдется или переобучится. Причиной данного явления будет тот факт, что модель не может полностью заменить истинные (размеченные) данные. Возникают ошибки, не зависящие от модели, а потому их можно выявить и рассматривать как ошибки данных. Это не полностью автоматическое определение ошибок, человек остается включенным в процедуру оценки. При этом эффективность проверки разметки значительно увеличивается, когда с помощью модели выделяются эксперименты, наблюдаемые данные. Так как, размечаемые данные составляют иногда сотни тысяч наблюдений, то не автоматизированные способы проверки разметки мало эффективны. Таким образом, все более актуальной становится задача независимой автоматизированной проверки разметки данных. Далее автор рассматривает эффективность моделей в работе с ошибками разметки.

II.Методика

Эффективность модели текста для классификации зависит от всех шагов обработки текста, но в первую очередь от метода извлечения признаков из текста. Различают «неглубокие признаки» и «глубокие». К «неглубоким признакам» относят длину текста, количество слов, количество знаков препинания, и другие. «Глубокие признаки» основываются на авто-репрезентации текста, автоматическом извлечении признаков из текста. Существует множество методов

Статья получена 19 февраля 2023. Ф.В.Краснов, Исследовательский центр ООО "ВБ СК" на базе Инновационного Центра Сколково. krasnov.fedor2@wb.ru, http://orcid.org/0000-0002-9881-7371.

извлечения признаков для классификации текста, таких как терм-документная матрица, модель счетчиков (term-frequency model), модели на основании ранжирующих функций семейства term-frequency и inverse document frequency (TF-IDF), извлечение именованных сущностей (NER), позиционные модели на основе глубоких нейронных сетях (RNN, LSTM и BERT). Все эти методы извлечения признаков основываются на разной информации о последовательностях символов - токенах, извлекаемых из текста. Такой информацией может быть количество токенов в документе, последовательность токенов, способ образования токена. Поэтому методы извлечения токенов непрерывно совершенствуются и развиваются. В модели классификации текста могут использоваться не сами токены, а более крупные структуры, содержащие несколько токенов. Такие структуры называются термами: на основании токенов собираются термы, терм может содержать несколько токенов. Таким образом, токен является наименьшей структурой, а терм более сложной, состоящей из последовательности токенов. Токенизация – это процесс разбиения текста на токены и формирования термов. В данной статье автор рассматривает несколько видов токенизации, в частности Byte Pair Encoding (BPE) [5], которая используется в моделях GPT и BERT и показывает SOTA результаты.

В исследовании [6] определяют веса термов как присвоение числового значения каждому терму для оценки его вклада, который способствует документу выделяться среди других документов коллекции. С другой стороны, выбор терма, также известный как выбор признаков (feature selection), определяется как выбор подмножества термов из всех термов, встречающихся в коллекции, для лучшего представления документа, либо для ускорения вычислений, либо для достижения большей эффективности классификации.

Процесс выбора термов также нуждается в еще одном параметре, чтобы настроить количество термов, которые необходимо выявить. Согласно исследованию [7] , алгоритмы взвешивания термов должны удовлетворять трем фундаментальным правилам, а именно:

- редкие термы не менее важны, чем частые термы.
- многократное появление терма в документе не менее важно, чем однократное появление.
- при одинаковом количестве совпадений термов длинные документы не менее важны, чем короткие.

Эти правила отмечены в статье [7], так же почти совпадают с тем, что приведено в работе [8]. Так же в [8] авторы определяют, что метод фильтрации при выборе терма должен соответствовать следующим требованиям:

- алгоритм должен присвоить признаку высокое значение скоринга, если он часто встречается в одном документе или в некоторых документах.
- алгоритм должен присвоить более низкую оценку признаку, если он редко встречается в одном документе или часто встречается во всех документах.

Поскольку взвешивание термов и выбор термов преследуют одну и ту же цель, существует множество алгоритмов взвешивания термов, вдохновленных алгоритмами выбора признаков. Существует множество алгоритмов смешивания при взвешивании термов, таких как «TF-коэффициент корреляции», «TF - отношение шансов», «TF - информационный выигрыш», TF-IDF, TFIDF-LTC, нормализованный LTC и основанный на вероятности TFIDF-LTC. В результате в [6] экспериментировали с такого рода алгоритмами взвешивания термов, чтобы определить, какие алгоритмы обеспечивают наилучшую точность классификации текста и получили лучшую оценку в 80% для метрики F1-score macro.

Автор обратил внимание на другую статью [8], в которой предложен новый алгоритм, названный «Частотно-обратный гравитационный момент» (TF-IGM). Эксперименты проводились с несколькими популярными наборами данных, такими как «20 групп новостей», Reuters-21578 и TanCorp. В [7] назвали это контролируемым взвешиванием термов (STW), взвешиванием термов путем использования известной категориальной информации в учебном наборе данных. TF-IGM был сравнен с TF-IDF, TF-RF, TF-IGM, RTF-IGM. Более того, в [8] также составили представление о том, сколько признаков использовалось в этих экспериментах, для набора данных «20 групп новостей» они использовали {500, 1000, 2000, 4000, 6000, 9000, 12000, 16000}, а{100, 300, 500, 700, 1000, 1500, 2000, 3000, 5000, 8436} использовались для набора данных Reuters-21578, и, наконец, {200, 500, 1000, 2000, 4000, 6000} для TanCorp. Так же важно отметить, что количество документов в коллекциях не велико, что делает результаты их исследования применимыми к рассматриваемым авторов наборов данных.

Что на взгляд автора упускается в исследовании [9], это то, что токены могут быть самыми разными фрагментами текста: предложениями, словами, знаками препинания, набором символов. Для моделей классификации на уровне слов замечено, что из-за морфологического разнообразия в русском языке связь между представлением текста и метками класса ослабевает. Простой интуицией тут может служить то, что множественное число и падежи слова увеличивают в разы разнообразные сочетания слова и класса. Отдельно стоит отметить существенно большую устойчивость субсловарных токенов к опечаткам, отмеченную в исследованиях субсловарных моделях токенизации fastText [10] и Byte Pair Encoding [5].

Задача определения ошибок разметки решается автором в рамках постановки задачи о классификации. Ошибки разметки могут быть нескольких классов, так же необходимо определить модель классификации, способ выделения признаков и способ поиска оптимальных условий для наиболее точного выделения классов

ошибок разметки.

Переходя к рассмотрению моделей классификации, рассмотрим подробнее феномен переобученности. Формального математического определения для переобученности модели нет, но в машинном обучении и статистике это явление проявляется, когда построенная модель хорошо объясняет примеры из обучающей выборки, но относительно плохо работает на примерах, не участвовавших в обучении (на примерах из тестовой выборки). Связано это с тем, что при построении модели («в процессе обучения») в обучающей выборке обнаруживаются некоторые закономерности, которые отсутствуют в генеральной совокупности. Так же способность переобучаться зависит от количества параметров модели, другими словами от сложности модели. Модели векторного представления текста могут иметь огромное количество параметров из-за размера словаря термов и и количества документов в коллекции. В процессе обучения модель может запоминать огромное количество всех возможных примеров вместо того, чтобы научиться подмечать особенности. Исследование метрик переобученности предпринято в работе [11], в которой введена метрика Скорость переобучения (RO, Rate of Overfitting), позволяющая оценивать степень переобученности в динамике по формуле:

$$RO = -\left(\frac{dE_{test}}{dT}\right) \cdot e^{-\frac{dE_{train}}{dT}}, \quad (1)$$

где E_{test} , E_{train} - ошибки обучения на тестовом и тренировочном наборах данных соответственно, а T – время (эпохи) обучения. Знак RO определяется ошибкой тестирования E_{test}. Когда RO < 0, это означает, что с увеличением обучающего набора ошибка тестирования также увеличивается. А чем меньше значение RO, тем выше переобучение. Чтобы снизить ошибку тестирования E_{test} и избежать переобучения, некоторые алгоритмы могут увеличить ошибку обучения E_{train}, добавив штраф в целевую функцию. Когда RO = 0, это означает, что с увеличением обучающего набора ошибка тестирования больше не меняется. Когда RO больше 0, это означает, что ошибка тестирования E_{test} уменьшается. Если значение RO остается положительным, это означает, что модель не переобучилась. RO является комплексной метрикой, характеризующей переобученность модели. Недостаток метрики Скорость переобучения (RO, Rate of Overfitting) состоит в ее моментальном значении, в определенные моменты времени значение RO может быть отрицательным, а следующие положительным. Для оценки переобученности модели в целом автор ввел усовершенствованную метрику Cumulative Rate of Overfitting (ρ).

$$\rho = \int -\left(\frac{dE_{test}}{dT}\right) \cdot e^{-\frac{dE_{train}}{dT}} dT \quad (2)$$

Согласно определению (2) ρ будет характеризовать сколько переобучений и недообучений модели было в сумме к данному моменту времени, и таким образом более точно характеризовать степень переобученности к моменту времени, а не в момент времени, как это делала метрика RO.

С другой стороны, можно рассматривать переобучение, как проявление компромисса «отклонение-дисперсия» (Рис. 1).

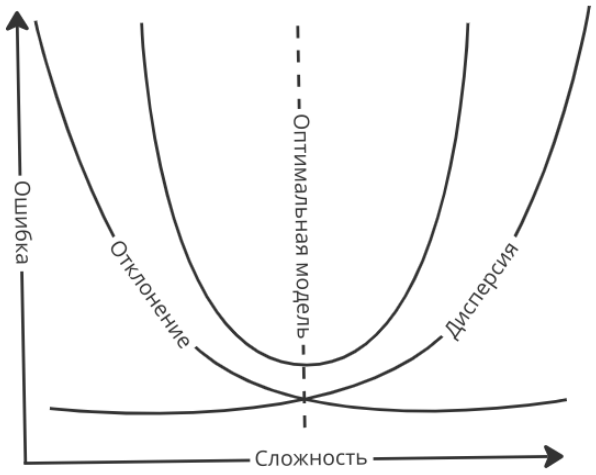

Рисунок 1 - Иллюстрация компромисса «отклонение-дисперсия»

В машинном обучении это свойство моделей предсказания, когда модели с меньшим отклонением от имеющихся данных имеют более высокую дисперсию на новых данных (то есть подвержены переобучению), и наоборот. Компромисс «отклонение-дисперсия» – это конфликт при попытке одновременно минимизировать эти два источника ошибки, которые мешают алгоритмам обучения с учителем делать обобщение за пределами тренировочного набора данных. Дисперсию тогда можно рассматривать, как ошибку чувствительности к малым отклонениям в тренировочном наборе. При высокой дисперсии алгоритм может как-то трактовать случайный шум в тренировочном наборе, а не желаемый результат (переобучение). С другой стороны, у модели может не доставать сложности, для того чтобы переобучиться. Например, модель классификации на основе «ближайших соседей» или «дерева решений» не достаточно сложны, чтобы переобучиться. Таким образом, если модель должна обладать достаточной сложностью чтобы быть способной переобучиться на тренировочном наборе данных, но даже если такая модель не достигает переобучения, то это свидетельствует от том, что в данных есть ошибки, которые и не позволяют переобучиться.

Одним из проявлений переобученности модели может быть неоднородность в параметрах модели,

возникающая при обучении. Некоторые параметры модели могут принимать аномально большие значения. Для борьбы с такой ситуацией используют регуляризацию. В случае оптимизации с помощью метода обратного распространения ошибки автор обратил внимание на основании какой функции потерь происходит оптимизация. В рассматриваемом нами случае текстовой классификации используется ошибка на основе энтропии. Для задачи мультиклассовой классификации для C классов функция энтропии прогноза $\hat{y}$ и фактических значений y равна:

$$L = -\sum_i^C y_i \log \hat{y}_i \quad (3)$$

Функция потерь на основе энтропии для классификации для n наблюдений будет средним значением L по всем n. Из формулы (3) видно, что если какой-либо прогноз $\hat{y}_k$ очень мал, то $\hat{y}_k$ будет чрезмерно влиять на потери, особенно если истинное значение y_k ближе к 1. То есть, при $y \in [0,1]$, имеем при $y \to 0$, $\log(y) \to -\infty$. Это поведение L приводит к тому, что определенные наблюдения могут чрезмерно влиять на ошибку, так что (небольшое) их количество в обучающем наборе (а не в тестовом наборе) чрезмерно повлияет на оценку потери и, следовательно, на обновление параметров обучения, вызывая переобучение. К тому же стоит отметить, что, когда функция потерь неограниченна, то появляется возможность больших значений параметров модели классификатора. А без предварительного знания истинного распределения это всегда риск.

Причина наличия ошибок разметки в обучающем наборе текстовых данных лежит в области человеческого восприятия текста и интерпретации правил разметки. Сами ошибки могут быть разделены на следующие два типа. Если X – это множество наблюдений, а C — это класс, $X_i^C \in X$ и $X_j^C \in X$ и $X_i^C \cap X_j^C = \emptyset$, то

Тип 1. Два или более подмножеств наблюдений отнесены $X_i^{C_a}, X_j^{C_b} \to C$ к одному классу,

Тип 2. Одно подмножество наблюдений частями отнесено к двум или более разным классам $X_i^C \to C_a$ и $X_j^C \to C_b$.

Таким образом, уточненный исследовательский вопрос состоит в выборе метода извлечения информации из коротких описаний товаров и метода классификации, обладающих оптимальным соотношением «отклонение-дисперсия» для решения задачи нахождения ошибок разметки фасетов.

III.Эксперимент

Для проверки изложенной выше методики был выбран публичный набор данных [12]. В ходе цифрового эксперимента опробованы различные методы обработки и классификации для получения переобученных моделей, у которых с помощью метрик оценивалось качество выявления ошибок разметки. Полученные результаты были применены для решения прикладной задачи поиска ошибок разметки в наборе данных D сопоставления коротких описаний товаров и названий динамических фильтров поиска (фасетов). Набор данных D состоит из 3000 записей, размеченных асессорами на 7 классов с ошибками. Для определения ошибок разметки были распределены 500 наблюдений с ошибками разметки первого и второго типов. Особенность решаемой задачи заключается в рассмотрении только размеченных асессорами данных и отмечании повторно ошибок в этих данных для обучения классификатора в той степени, чтобы проявилось максимально точное количество ошибок разметки. В качестве метрики используем f1-score. В качестве методов токенизации были выбраны – whitespace, subword 2-4 grams, BPE (таблица 1).

Таблица 1 -- Методы токенизации с примерами

Метод токенизации	Результат
Subword 2,4-grams	' д', ' да', ' дам', ' дл', ' для', ' м', ' ма', ' ман', ' н', ' на', ' на ', ' п', ' пы', ' пыш', 'а ', 'а м', 'а ма', 'ам', 'ан', 'анж', 'анже', 'гг', 'гге', 'ггер', 'ге', 'гер', 'геры', 'да', 'дам', 'дж', 'джо', 'джог', 'дл', 'для', 'для ', 'е ', 'е д', 'е дл', 'ер', 'еры', 'еры ', 'ет', 'ете', 'ете ', 'же', 'жет', 'жете', 'жо', 'жог', 'жогг', 'ля', 'ля ', 'ля п', 'ма', 'ман', 'манж', 'на', 'на ', 'на м', 'нж', 'нже', 'нжет', 'ны', 'ных', 'ных ', 'ог', 'огг', 'огге', 'пы', 'пыш', 'пышн', 'ры', 'ры ', 'ры н', 'те', 'те ', 'те д', 'х ', 'х д', 'х да', 'шн', 'шны', 'шных', 'ы ', 'ы н', 'ы на', 'ых', 'ых ', 'ых д', 'ыш',

	'ышн', 'ышны', 'я ', 'я п', 'я пы'
BPE	джоггеры на манжете для п@@ ы@@ шн@@ ы@@ х да@@ м
whitespace	джоггеры на манжете для пышных дам

Для обучения модели выбрано несколько методов классификации с различной сложностью и отношением «отклонение-дисперсия». Выбранные методы классификации приведены в Таблице 2.

Таблица 2 – Соотношение «отклонение-дисперсия» для различных методов классификации

Название метода	Отклонение / Дисперсия	Кол-во параметров
LogisticRegression	Высокое / Низкая	N * M
EmbeddingBag	Высокое / Низкая	M*Emb_Dim_Term
BERT + FF	Высокое / Низкая	N*Emb_Dim_Doc

В Таблице 2 за N обозначено количество документов в коллекции, M – размер словаря, Emb_Dim_Doc-размерность компактного векторного представления для документа, Emb_Dim_Term-размерность компактного векторного представления для терма.

В качестве предобученной модели на основании архитектуры BERT была использована модель RuBERT (Russian, cased, 12-layer, 768-hidden, 12-heads, 180M параметров) [13]. Для построения модели классификации на основе BERT созданы компактные векторные представления документов и добавлен линейный слой для определения вероятностей классов. Компактные векторные представления документов не изменялись в процессе обучения, поэтому переобучение происходит только на выходном линейном слое.

Концепция метода EmbeddingBag взята из программной библиотеки pytorch. Основное отличие EmbeddingBag от Embedding состоит в том, что обычный слой Embedding создает вектор значений (количество значений равно Emb_Dim_Term) для каждого терма. При обучении моделей типа Transformers или LSTM необходимо разбивать последовательность термов на одинаковые по размеру блоки [14]. Для этого исследователи должны дополнять (PAD) короткие предложения и обрезать (TRUNC) длинные документы. И обрезание, и дополнение документов изменяет распределения совстречаемости термов. Вместо того, чтобы каждый терм был представлен вектором значений в EmbeddingBag, каждый документ представлен вектором.

В таблице 3 приведены примеры ошибочной разметки классов текстов из набора данных [12], детектированные моделью.

Таблица 3 – Примеры ошибок разметки, обнаруженных с помощью модели классификации

Текст	**Класс разметки**	**Класс модели**	**Вероятность модели**
Как стать аналитикомextreme и сделать карьеру в крупной IT компании 5 6 мая в 16. 00 ждем вас на практическом интенсиве по Аналитике данных Регистрация по ссылке На интенсиве Кто такой аналитик данных чем он занимается и где работает. Какие инструменты нужны новичку в работе и как начать карьеру. Практика учимся выдвигать и проверять бизнес гипотезы на примере кредитного сектора. Что такое A B тестирование и почему это самое перспективное направление аналитики. Инструмент аналитика данных Googtokenoid Cotokenoid. Проводим A B тесты и делаем выводы на их основе. Регистрация по ссылке		*martial_arts*	*1.00*

Бонусы для участников Спикер Данила Елистратов Техлид факультета Аналитика данных в Skypro. Скорее регистрируйтесь и встретимся в прямом эфире			
Похудеть возможно даже если. 1. Любишь сладкое и жирное да бургер – бро сахар – не зло. 2. Ешь на ночь никогда не поздно поужинать. 3. Не исключаешь углеводы даже на ночь. 4. Не скупаешь жиросжигатели и прочую чушь несмотря на старания маркетологов. 5. Питаешься вкусно разнообразно и сытно диета – не равно голодовка. Не веришь? Переходи по ссылке Здесь нет места маркетинговой лапше. Я эксперт по питанию и сертифицированный тренер с 15 летним опытом научу тебя как управлять весом без тяжелых диет и волшебных таблеток. Пусть другие худеют на капустном листе и терпят фиаско на нелепых диетах. А ты узнаешь о 5 ти простых шагах которых достаточно чтобы начать делать форму. Скачивай книгу и действуй	*martial_arts*	*athletics*	*1.00*
Ищу способных	*autosport*	*martial_arts*	*1.00*
учеников на обучение трейдингу удаленной профессии набирающей популярность в интернете. Мне нужны трудоспособные мужчины и девушки у которых есть ноутбук или смартфон один свободный час в день и желание уделить его обучению и применению знаний на практике чтобы зарабатывать в 2 3 больше чем на основной работе. Что Вы узнаете за 5 дней Как стартовать в онлайн трейдинге. Подходит ли Вам профессия трейдера. Как совмещать трейдинг и основную работу. Как выйти на стабильный доход в 2 3 зарплаты РФ с трейдинга. Переходите по ссылке ниже и первый урок сразу придет в личные сообщения. Тренинг целиком бесплатный но буду благодарен если после оставите отзыв и расскажете об успехах.			

На рисунке 2 приведены зависимости Cumulative Rate of Overfitting (ρ) для нескольких моделей классификации, полученные в результате эксперимента.

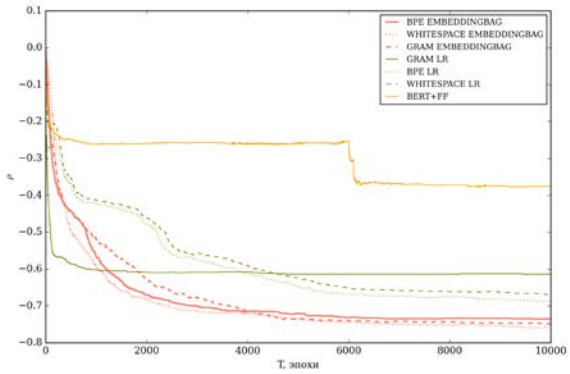

Рисунок 2 – Зависимости Cumulative Rate of Overfitting (ρ) от модели классификации

На рисунке 2 наглядно изображено, что все рассматриваемые модели выходят на асимптоту с течением времени и метрика переобученности ρ перестает уменьшаться. Для каждой модели предельное значение ρ имеет собственные границы и варьируется в зависимости метода токенизации в рамках этих границ.

В таблице 4 приведены результаты моделей классификации на тестовой выборке.

Таблица 4 – Результаты моделей на тестовой выборке

Модель	**f1-score (weighted avg.)**	**Количество параметров**
BERT+FF	0.87	5 383
BPE EMBEDDINGBAG	0.92	15 015
WHITESPACE EMBEDDINGBAG	0.91	17 239
GRAM EMBEDDINGBAG	**0.98**	**111 127**
GRAM LR	0.97	55 223
BPE LR	0.90	6 524
WHITESPACE LR	0.92	7 497

Эксперимент проводился многократно, и результаты метрик приведены по среднему значению. Наилучший результат f1-score показала модель классификации на основе EmbeddingBag. Отметим, что эта модель обладает и наименьшим значением метрики переобученности ρ. Что в целом подтверждает исследовательскую гипотезу автора о продуктивности переобученных моделей классификации для обнаружения ошибок разметки в небольших текстовых коллекциях.

IV. ЗАКЛЮЧЕНИЕ

В статье приведены результаты изучения точности методов поиска ошибок в разметке. Точность разметки измерена с помощью дополнительной разметки случаев с неправильной разметкой. Поиск ошибок разметки с помощью разметки – это в некоторой степени «порочный круг», но такой подход позволил автору оценить эффективность различных методов машинного обучения в качестве кандидатов для автоматизированного контроля разметки. В практике существуют методы усреднения ошибок асессоров, которые требуют значительно больших затрат, чем предлагаемая автором частичная автоматизация.
Автор исследовал применение наиболее современных методов токенизации текста. Токенизация с помощью BPE не обнаруживает пропуски пробелов и опечатки. В силу того, что классифицируемыми документами являются Описания товаров, текст, созданный пользователями торговой онлайн-площадки, может содержать опечатки, которые иногда исправляются в рамках отдельных шагов процесса обработки текста. Токенизация на основе subword 2,4-grams более устойчива к опечаткам и пропущенным пробелам.

Автор проанализировал эффективность трех моделей классификации для определения ошибок разметки (таблица 2). Наилучший результат по точности f1-score weighted 98% показала модель EmbeddingBag с токенизацией subword 2,4-grams, в которой оптимально сочетаются сложность и ошибка. Модель классификации на основе предобученной языковой модели BERT показала худший результат для решения данной задачи.

Проведенный эксперимент выявил: влияние метода токенизации на результат переобученности распространяется только в рамках одной модели классификации. Иными словами, результат в меньшей степени зависит от токенизации, чем от выбора модели классификации.

В заключение необходимо подчеркнуть, что переобучение моделей напрасно считается априори отрицательным, не полезным и ошибочным. Установлено, что для классификации коротких текстов переобученная модель способна выявлять ошибки разметки данных. В эксперименте с формированием фасетов на базе пользовательских описаний товаров автор обосновал: наличие ошибок определения класса связано не с качеством модели, а с тем, что асессоры делали разметку на основании разных правил.

Библиография

[1] Hu, S., Ding, N., Wang, H., Liu, Z., Li, J., & Sun, M. (2021). Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035.

[2] Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. (2020). Text classification using label names only: A language model self-training approach. arXiv preprint arXiv:2010.07245.

[3] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv preprint arXiv:1909.00161.

[4] Yin, W., Rajani, N. F., Radev, D., Socher, R., & Xiong, C. (2020). Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. arXiv preprint arXiv:2010.02584.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1715–1725.

[6] Luo, Qun, Weiran Xu, and Jun Guo. "A Study on the CBOW Model's Overfitting and Stability." Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning. 2014. https://dl.acm.org/doi/abs/10.1145/2663792.2663793

[7] Debole, Franca and Fabrizio Sebastiani. “Supervised term weighting for automated text categorization.” ACM Symposium on Applied Computing (2003).

[8] Uysal, Alper Kursat. “An improved global feature selection scheme for text classification.” Expert Syst. Appl. 43 (2016): 82-92.

[9] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[10] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

[11] Feng, X., Liang, Y., Shi, X., Xu, D., Wang, X., & Guan, R. (2017). Overfitting reduction of text classification based on AdaBELM. Entropy, 19(7), 330.

[12] Mazurov M. Russian Social Media Text Classification. [Электронный ресурс] // https://www.kaggle.com : Наборы данных для конкурсов. М., 2022. URL: https://www.kaggle.com/datasets/mikhailma/russian-social-media-text-classification (дата обращения: 24.12.2023).

[13] Kuratov, Y., Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Identifying data labeling errors using classification models for small datasets

F.V. Krasnov

Abstract — Labeling up data for classification tasks is a complex process, accompanied by unavoidable errors. Manual or automatic labeling of texts for classification includes systematical errors that can be identified using system approaches based on statistics and machine learning models. It is small data sets that are considered, since the consequences of labeling errors are most noticeable in them. However, in case of small datasets, due to the lack of samples, the problem of sparse distributions arises, which prevents the training of models with high complexity. The author uses the effect of overfitting the model to minimize the limitations imposed by insufficient data. Several experiments were conducted as part of the study. An experiment on a large public dataset showed that when classifying short texts, the overfitted model is able to detect data labeling errors. In an experiment with the formation of facets based on user description of goods, the interdependence of the presence of class definition errors and the work of assessors on the labeling of text data based on different rules was determined. Due to its overfitting, the classification model is capable of identifying significant errors that dramatically affect the engineering application of machine learning in highly loaded Internet systems. As a result of the research, the author provides methods and criteria for achieving the state of "productive overfitting" by the model. The best result on the f1-score weighted metric (98%) was shown by the EmbeddingBag-based classification model.

Keywords — Overfitting problem, Entropy, Classification

REFERENCES

[1] Hu, S., Ding, N., Wang, H., Liu, Z., Li, J., & Sun, M. (2021). Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. arXiv preprint arXiv:2108.02035.

[2] Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. (2020). Text classification using label names only: A language model self-training approach. arXiv preprint arXiv:2010.07245.

[3] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv preprint arXiv:1909.00161.

[4] Yin, W., Rajani, N. F., Radev, D., Socher, R., & Xiong, C. (2020). Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. arXiv preprint arXiv:2010.02584.

[5] Rico Sennrich, Barry Haddow, and Alexandra Birch, “Neural machine translation of rare words with subword units,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1715–1725.

[6] Luo, Qun, Weiran Xu, and Jun Guo. "A Study on the CBOW Model's Overfitting and Stability." Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning. 2014. https://dl.acm.org/doi/abs/10.1145/2663792.2663793

[7] Debole, Franca and Fabrizio Sebastiani. “Supervised term weighting for automated text categorization.” ACM Symposium on Applied Computing (2003).

[8] Uysal, Alper Kursat. “An improved global feature selection scheme for text classification.” Expert Syst. Appl. 43 (2016): 82-92.

[9] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150.

[10] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5, 135-146.

[11] Feng, X., Liang, Y., Shi, X., Xu, D., Wang, X., & Guan, R. (2017). Overfitting reduction of text classification based on AdaBELM. Entropy, 19(7), 330.

[12] Mazurov M. Russian Social Media Text Classification. [Jelektronnyj resurs] // https://www.kaggle.com : Nabory dannyh dlja konkursov. M., 2022. URL: https://www.kaggle.com/datasets/mikhailma/russian-social-media-text-classification (data obrashhenija: 24.12.2023).

[13] Kuratov, Y., Arkhipov, M. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. arXiv preprint arXiv:1905.07213.

[14] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.