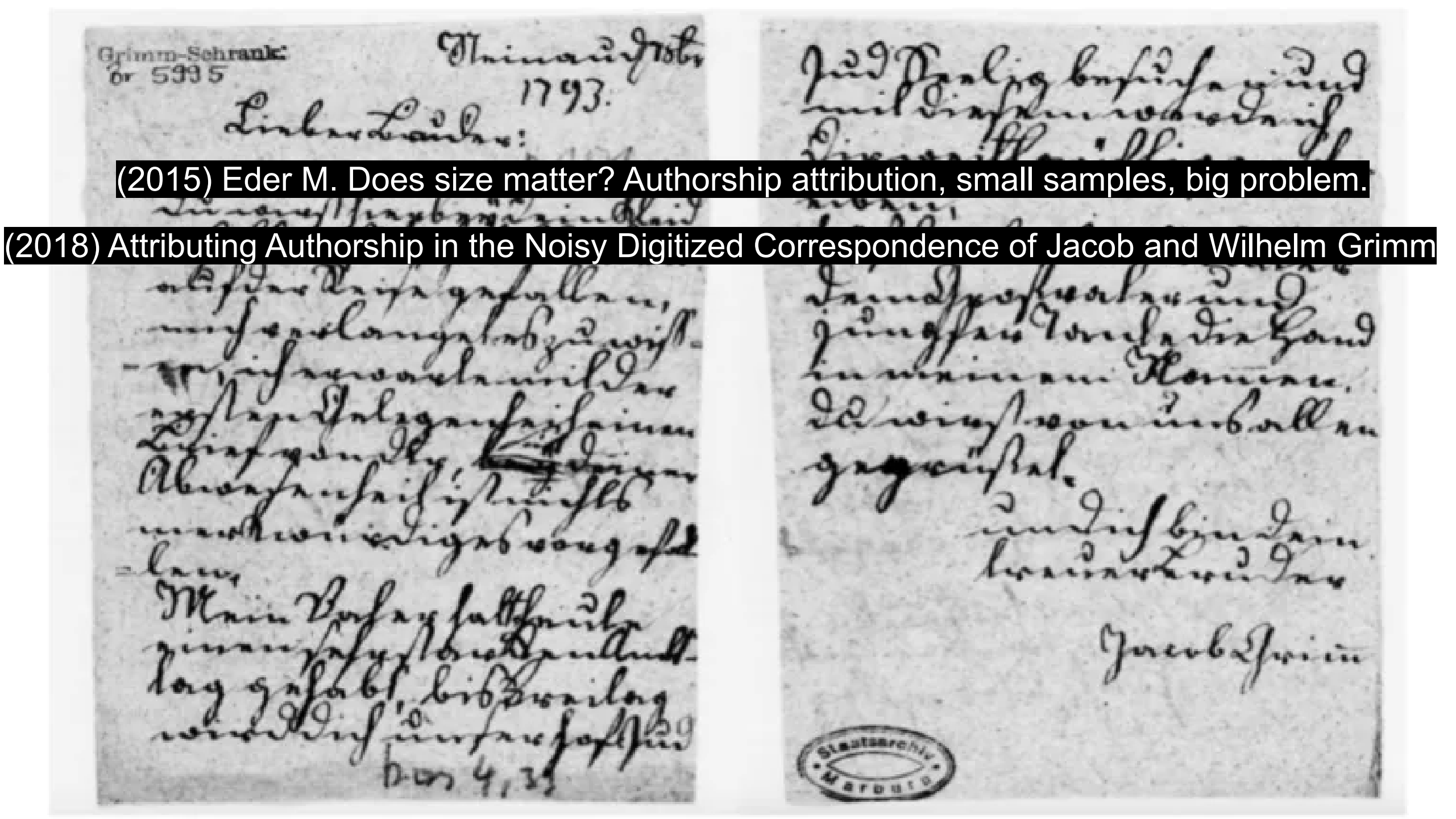
Коваленко А.П.
Мельников С.Ю.
Мещеряков Р.В.
Хвостенко В.М.

Академия криптографии РФ,
г. Москва

Как деградация обучающих корпусов влияет на точность моделей атрибуции авторства



(2015) Eder M. Does size matter? Authorship attribution, small samples, big problem.

(2018) Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm

ЗАДАЧА

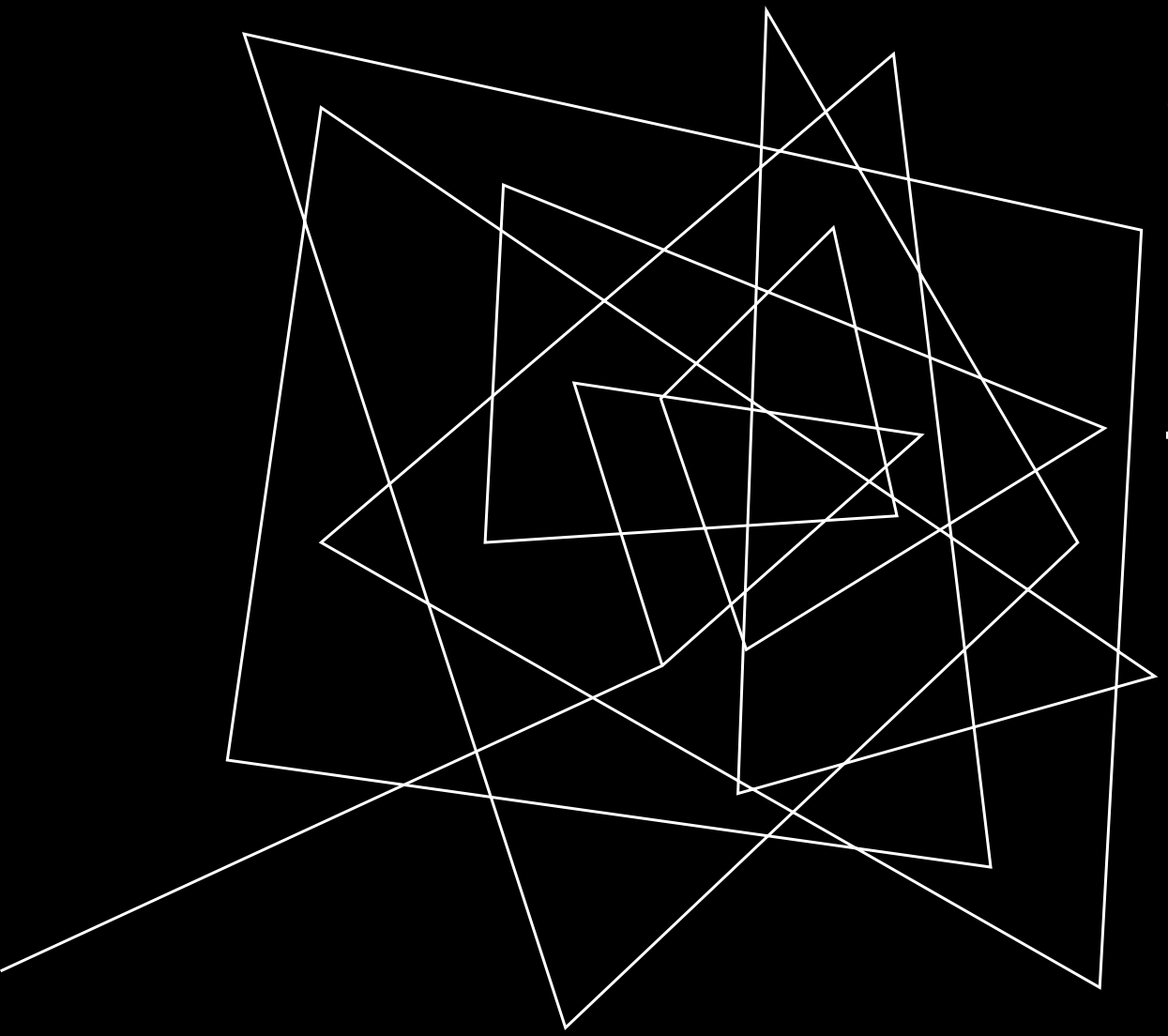
- Рост объёма пользовательского контента
- Основные трудности: короткие тексты, искажения, малые обучающие выборки.
- Цель работы: изучить, как два типа деградации обучающих данных одновременно влияют на точность классификаторов:
  - Случайные символьные замены (0-20 % шума).
  - Сокращение обучающего корпуса (с 800 до 50 текстов на автора).
- Тестовые тексты остаются неизменными.

ТИПЫ И МОДЕЛИ ИСКАЖЕНИЙ ТЕКСТА

- Типы искажений: опечатки при наборе, OCR-ошибки, ошибки распознавания речи, машинный перевод, частичное дешифрование.
- В работе использованы случайные символьные замены (замена символа).
- Процесс генерации шума: для каждого алфавитного символа независимо с вероятностью p (0, 0.01, 0.03, 0.05, 0.10, 0.20) выбирается замена на случайную букву латиницы.

ОПИСАНИЕ КОРПУСА И ФОРМИРОВАНИЕ ВЫБОРОК

- Исходный корпус: англоязычные твиты 2009-2010 гг., 7000 самых популярных аккаунтов.
- Выбрано 50 авторов случайно, по 1000 твитов от каждого.
- Деление: 80 % обучающие (800 твитов/автор), 20 % тестовые (200 твитов/автор).
- Варианты объёма обучающего корпуса: 800, 750, 700,…, 50 (шаг = 50 дает 16 вариантов).
- Для каждого объёма создаём шесть зашумлённых версий (уровни шума 0-20 %).
- Тестовые наборы остаются чистыми и неизменными во всех экспериментах.



ПРИЗНАКИ АВТОРСКОГО СТИЛЯ

BOW (Bag-of-Words) – частотный вектор:

- токены
- биграммы токенов

TF-IDF – взвешенная частота:

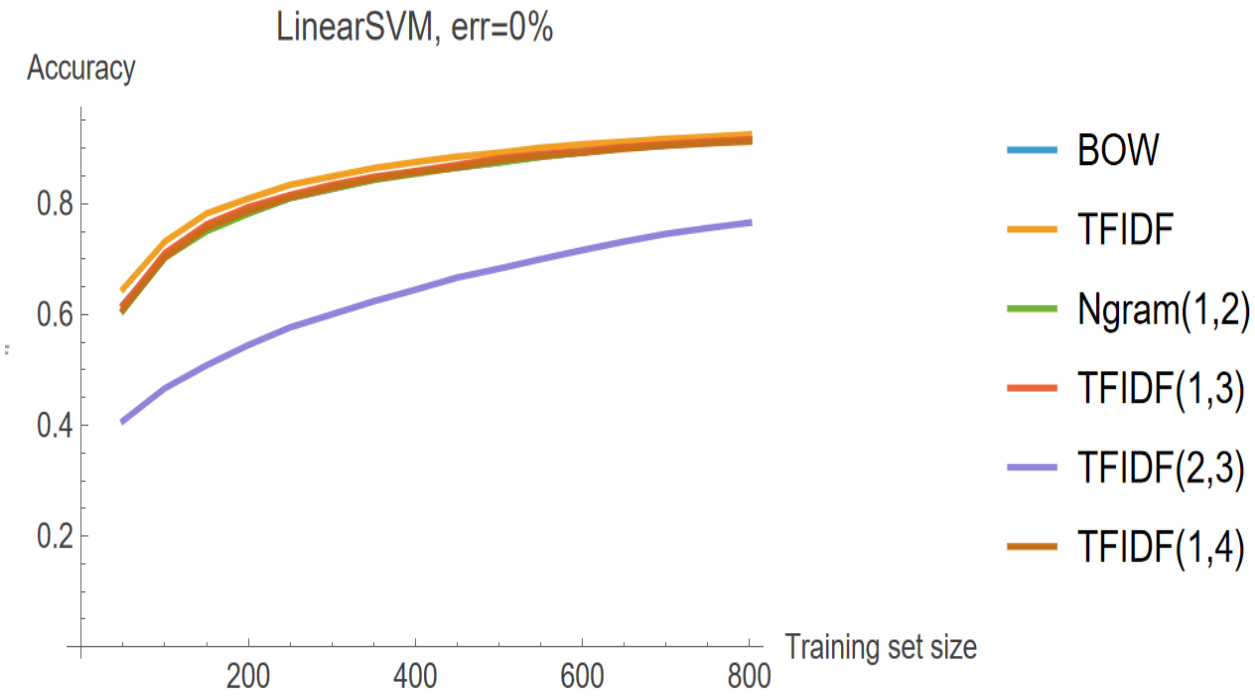
- токены (1-граммы)
- токены + биграммы + триграммы
- биграммы + триграммы
- токены + биграммы + триграммы + четырехграммы

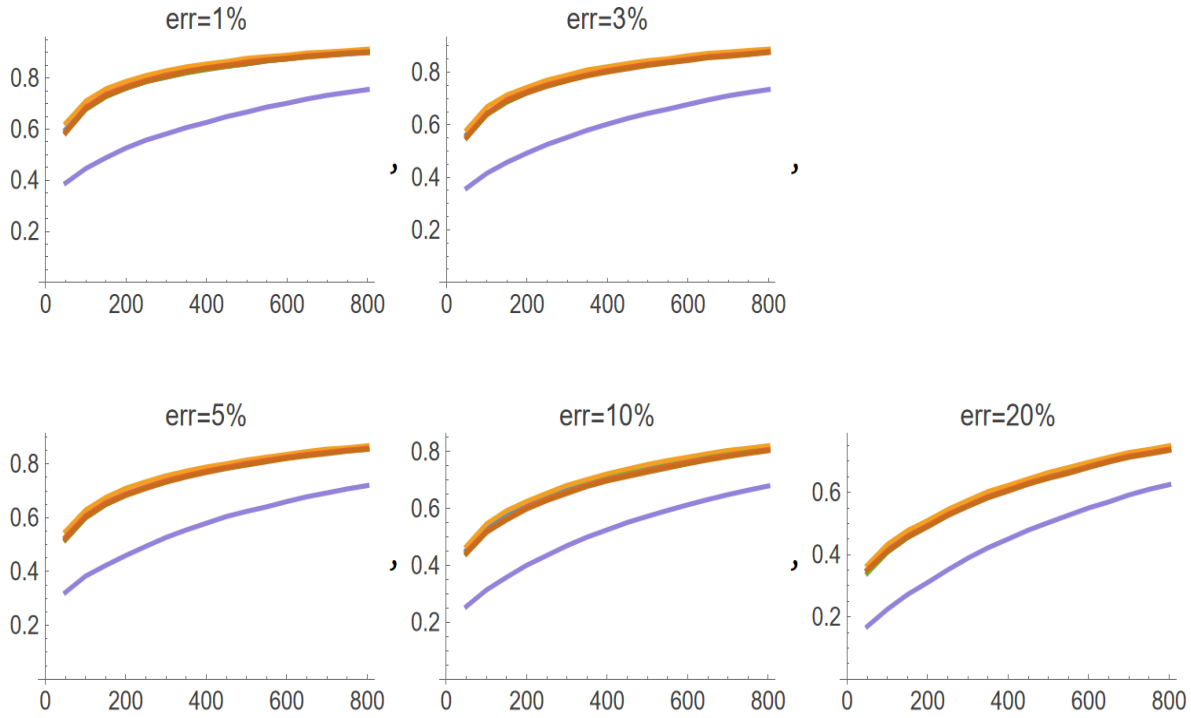
Векторы разреженные, высокоразмерные (десятки тысяч признаков).

КЛАССИФИКАТОРЫ

Метод	Пакет Scikit-learn
SVM (линейное ядро)	LinearSVC()
Logistic Regression	LogisticRegression()
Naive Bayes (Multinomial)	MultinomialNB()
Random Forest	RandomForestClassifier()
Decision Tree	DecisionTreeClassifier()
k-Nearest Neighbors	KNeighborsClassifier()

РЕЗУЛЬТАТЫ: SVM

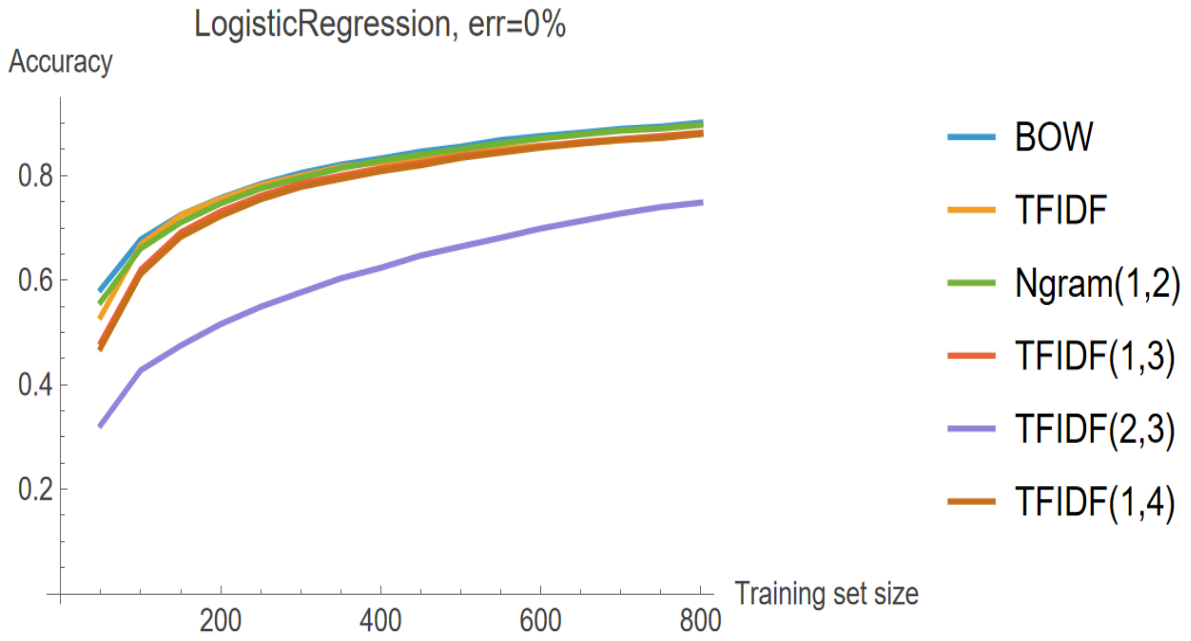


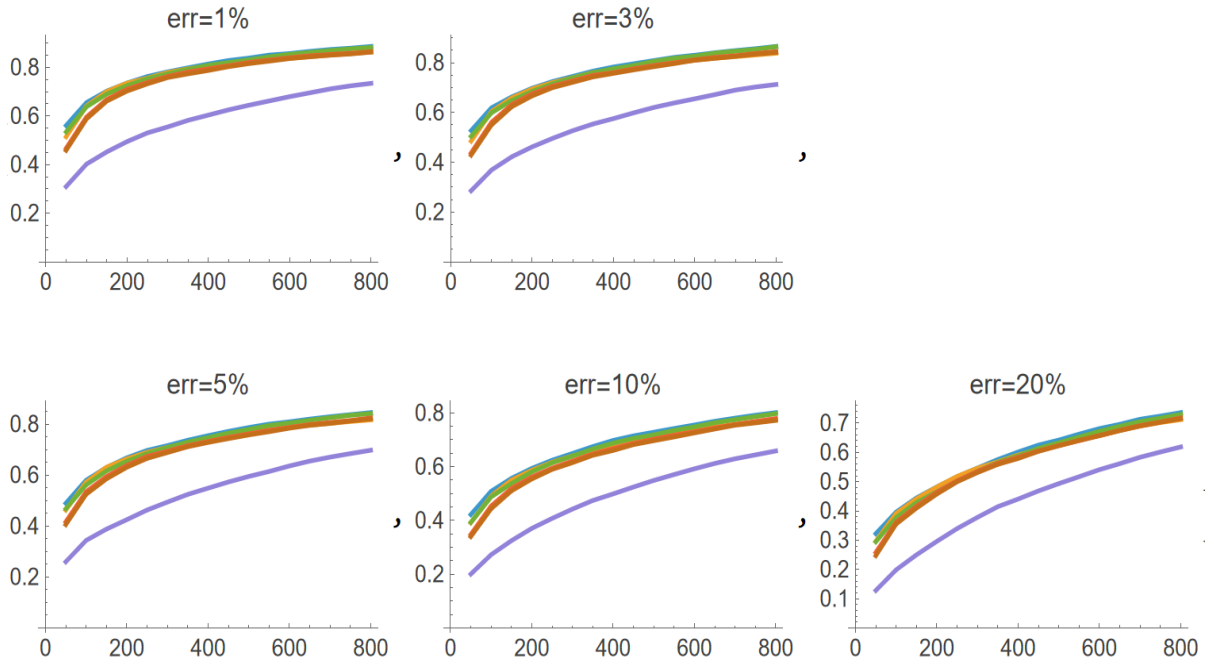


Точность классификатора SVM с линейным ядром в зависимости от полноты эталонных наборов данных (обучающие корпуса не искажены).

Точность классификатора SVM с линейным ядром в зависимости от полноты эталонных наборов данных (обучающие корпуса искажены на 1, 3, 5, 10, 20%).

РЕЗУЛЬТАТЫ:
LOGISTIC REGRESSION





Точность классификатора Logistic Regression в зависимости от полноты эталонных наборов данных (обучающие корпуса не искажены).

Точность классификатора Logistic Regression в зависимости от полноты эталонных наборов данных (обучающие корпуса искажены на 1, 3, 5, 10, 20%).

	800	600	400	200	50
Err=0%					
SVM	0.92	0.91	0.88	0.83	0.65
LR	0.9	0.88	0.85	0.78	0.58
NB	0.89	0.87	0.83	0.76	0.53
RF	0.88	0.86	0.82	0.75	0.57
DT	0.84	0.82	0.78	0.71	0.53
kNN	0.72	0.7	0.67	0.63	0.51
Err=5%					
SVM	0.89	0.87	0.83	0.77	0.58
LR	0.87	0.84	0.8	0.72	0.53
NB	0.86	0.84	0.79	0.71	0.49
RF	0.84	0.81	0.76	0.7	0.51
DT	0.81	0.78	0.73	0.65	0.47
kNN	0.65	0.63	0.6	0.56	0.46
Err=20%					
SVM	0.75	0.71	0.64	0.54	0.37
LR	0.74	0.7	0.62	0.52	0.32
NB	0.74	0.7	0.63	0.52	0.33
RF	0.71	0.67	0.59	0.48	0.3
DT	0.7	0.65	0.57	0.46	0.28
kNN	0.42	0.39	0.37	0.32	0.25

BOW	TFIDF	Ngram(1,2)	TFIDF(1,3)	TFIDF(1,4)

Лучшие результаты для каждого классификатора для трех значений уровня искажений (0%, 5% и 20%) и пяти вариантов размеров обучающего множества (800, 600, 400, 200 и 50 текстов).

Цвет ячеек таблицы соответствует той характеристике авторского стиля (BOW, TFIDF, Ngram(1,2), TFIDF(1,3), TFIDF(1,4)), которая для данного классификатора обеспечила лучший результат.

ЗАКЛЮЧЕНИЕ

Проведена оценка точности атрибуции авторства коротких текстов в случае, когда тексты обучающих корпусов подвергались случайным символьным искажениям, и производилась редукция обучающих корпусов. Использовался корпус текстов Twitter с 50 авторами по 1000 текстов от каждого автора. Атрибутируемые авторские тексты при этом не менялись. Число обучающих текстов в авторских коллекциях варьировалось от 800 до 50, а уровень искажений изменялся в пределах от 0% до 20%.

Для каждого варианта объема обучающего корпуса и уровня искажений проводилось обучение классификаторов и их тестирование. Оценивалась точность атрибуции авторства с помощью классификаторов SVM, Logistic Regression, Naive Bayes, Random Forest, Decision Tree, kNN. В качестве признаков авторского стиля использовались BOW (Bag of Words) и TF-IDF для токенов и N-грамм. При деградации обучающих наборов авторских текстов большинство рассмотренных методов атрибуции (кроме kNN) показали схожие снижения уровня точности.

Метод опорных векторов с использованием TF-IDF оказался лучшим при уменьшении количества обучающих текстов и при сильных искажениях. Без искажений он показал точность 0.92, при искажениях 5% и 20% достигнута точность 0.89 и 0.75 соответственно. Метод **Logistic Regression с использованием BOW занимает второе место**, обеспечивая в тех же условиях точности 0.9, 0.87 и 0.74. Методы Decision Tree и Random Forest сильнее всех страдают при уменьшении количества обучающих данных и росте искажений. **Метод kNN показал нестабильные и наихудшие по точности результаты**.

По материалам исследования в ближайшем номере журнала **Discrete and Continuous Models and Applied Computational Science** выходит наша статья:
V.M. Khvostenko, A.P. Kovalenko, S.Yu. Melnikov, R.V. Meshcheryakov «Modeling Authorship Attribution for Short Texts Under Training Corpus Degradation».

Работа частично поддержана грантом РНФ 24-11-00340 Исследование и разработка методов обработки слабоструктурированной информации на естественных языках в условиях сильных шумов для решения задач безопасности.

Российский научный фонд

СПАСИБО ЗА ВНИМАНИЕ!

ВОПРОСЫ К ДОКЛАДЧИКУ:
VICTOR.KHVOSTENKO@GMAIL.COM