MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 28, NO. 3, 2021
journal homepage: www.mais-journal.ru

THEORY OF DATA

Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose

A. M. Manakhova[1], N. S. Lagutina[1]

DOI: 10.18255/1818-1015-2021-3-260-279

[1]P. G. Demidov Yaroslavl State University, 14 Sovetskaya str., Yaroslavl 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received June 25, 2021
After revision August 23, 2021
Accepted August 25, 2021

This article is dedicated to the analysis of various stylometric characteristics combinations of different levels for the quality of verification of authorship of Russian, English and French prose texts. The research was carried out for both low-level stylometric characteristics based on words and symbols and higher-level structural characteristics.
All stylometric characteristics were calculated automatically with the help of the ProseRhythmDetector program. This approach gave a possibility to analyze the works of a large volume and of many writers at the same time. During the work, vectors of stylometric characteristics of the level of symbols, words and structure were compared to each text. During the experiments, the sets of parameters of these three levels were combined with each other in all possible ways. The resulting vectors of stylometric characteristics were applied to the input of various classifiers to perform verification and identify the most appropriate classifier for solving the problem. The best results were obtained with the help of the AdaBoost classifier. The average F-score for all languages turned out to be more than 92 %. Detailed assessments of the quality of verification are given and analyzed for each author. Use of high-level stylometric characteristics, in particular, frequency of using N-grams of POS tags, offers the prospect of a more detailed analysis of the style of one or another author. The results of the experiments show that when the characteristics of the structure level are combined with the characteristics of the level of words and / or symbols, the most accurate results of verification of authorship for literary texts in Russian, English and French are obtained. Additionally, the authors were able to conclude about a different degree of impact of stylometric characteristics for the quality of verification of authorship for different languages.

Keywords: stylometry; stylometric characteristics; authorship verification; natural language processing

INFORMATION ABOUT THE AUTHORS

Alla Mikhajlovna Manakhova correspondence author	orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru MSc student.
Nadezhda Stanislavovna Lagutina	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru Associate Professor, PhD in Physics and Mathematics.

For citation: A. M. Manakhova and N. S. Lagutina, “Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 260-279, 2021.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 28, № 3, 2021
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Анализ влияния стилометрических характеристик разного уровня на верификацию авторов художественных произведений

А. М. Манахова[1], Н. С. Лагутина[1]

DOI: 10.18255/1818-1015-2021-3-260-279

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, д. 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 25 июня 2021 г.
После доработки 23 августа 2021 г.
Принята к публикации 25 августа 2021 г.

Данная статья посвящена анализу влияния различных комбинаций стилометрических характеристик разного уровня на качество верификации авторства русских, английских и французских прозаических текстов. Исследование проводилось как для низкоуровневых стилометрических характеристик, основанных на словах и символах, так и для более высокоуровневых – структурных.

Подсчёт всех стилометрических характеристик был выполнен автоматически с помощью программы ProseRhythmDetector. Такой подход позволил провести анализ произведений большого объёма и многих писателей одновременно. В ходе работы каждому тексту были сопоставлены векторы стилометрических характеристик уровня символов, слов и структуры. При проведении экспериментов наборы параметров этих трёх уровней были скомбинированы между собой всеми возможными способами. Полученные векторы стилометрических характеристик были поданы на вход различным классификаторам для выполнения верификации и выявления наиболее подходящего классификатора для решения поставленной задачи. Лучшие результаты были получены с помощью классификатора AdaBoost. Средняя F-мера для всех языков оказалась более 92%. Детальные оценки качества верификации приведены для каждого автора и проанализированы. Использование высокоуровневых стилометрических характеристик, в частности, частоты использования N-грамм POS-тегов открывает перспективу более детального анализа стиля того или иного автора. Результаты экспериментов показывают, что при соединении характеристик уровня структуры с характеристиками уровня слов и/или символов получаются наиболее точные результаты верификации авторства для художественных текстов на русском, английском и французском языках. Дополнительно авторам удалось сделать вывод о разной степени влияния стилометрических характеристик на качество верификации авторства для различных языков.

Ключевые слова: стилометрия; стилометрические характеристики; верификация авторства; обработка естественного языка

ИНФОРМАЦИЯ ОБ АВТОРАХ

Алла Михайловна Манахова автор для корреспонденции	orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru магистрант.
Надежда Станиславовна Лагутина	orcid.org/0000-0002-6137-8643. E-mail: lagutinans@rambler.ru доцент, кандидат физико-математических наук.

Для цитирования: A. M. Manakhova and N. S. Lagutina, “Analysis of the Impact of the Stylometric Characteristics of Different Levels for the Verification of Authors of the Prose”, *Modeling and analysis of information systems*, vol. 28, no. 3, pp. 260-279, 2021.

Введение

Одной из проблем извлечения информации из неструктурированных данных является верификация автора текста на естественном языке [1]. Решение этой задачи заключается в определении принадлежности текста заданному автору. Основная часть исследований в этой области посвящена верификации авторов электронных писем и сообщений в социальных сетях [2]. Аналогичной задачей является разрешение споров об авторских правах [3]. Кроме анализа современных текстов, актуальной проблемой остаётся верификация авторов художественных произведений [4—6].

Основным методом автоматической верификации авторства является классификация текстов с помощью векторов стилометрических характеристик [7]. В первую очередь такими характеристиками являются простые текстовые функции, например, частоты слов и символов, n-граммы символов и слов, длины слов и предложений. В последнее время всё больше авторов обращают внимание на определение особенностей синтаксиса и грамматики. Количество таких параметров очень велико, может достигать нескольких тысяч. Однако не все характеристики вносят одинаковый вклад в решение задачи. Наиболее сложным для анализа стилем обладают художественные произведения. Поэтому авторы поставили перед собой задачу исследовать влияние различных групп стилометрических параметров на верификацию авторов повестей на английском, французском и русском языках.

1. Современное состояние исследований в области верификации авторства текстов

Задача верификации авторства может рассматриваться как математическая задача бинарной классификации, принадлежит ли рассматриваемый документ определённому классу или нет. Для текстов на естественном языке формируются характеристические числовые векторы, затем применяется один из методов классификации, которым очень часто является метод машинного обучения. Многие исследователи для вычисления вектора признаков документа используют, ставшие классическими в компьютерной лингвистике, параметры: частоты символов и слов, n-граммы слов, эмбеддинги слов [7]. Популярные низкоуровневые текстовые функции: униграммы слов и n-граммы символов использовали авторы статьи [8]. Они разработали метод проверки на основе внутреннего профиля автора, который создаёт текстовую модель, представляющую все тексты одного автора как общий вектор. Исследователи экспериментировали с корпусами прозы, газетными статьями, обзорами и другими жанрами из соревнований PAN-2014 и PAN-2015 на четырех языках: голландском, английском, греческом и испанском. Разработанный подход показал в экспериментах точность (accuracy) более 80 % и позиционировался авторами как независимый от языка. Ещё более высокий результат 90 % показал метод, описанный в работе [9]. Авторы выделили параметры стиля текста, не зависящие от языка и использовали метод выбора признаков SVM-RFE (Support Vector Machine Based on Recursive Feature Elimination) для удаления избыточных и нерелевантных характеристик из процесса обучения. Метод был применён к верификации авторства статей на четырёх языках: английском, греческом, испанском и немецком.

Однако использование параметрических векторов из простых характеристик имеет свои ограничения и недостатки. Авторы работы [10] обращают внимание, что надежность использования таких параметров в алгоритмах машинного обучения значительно снижается для коротких и тематически разнообразных текстов в социальных сетях. Авторы решают эту проблему попыткой визуализировать процесс принятия решений нейронной сетью. Для верификации авторства коротких обзоров Amazon исследователи используют сиамские нейронные сети. При обсуждении результатов авторы проводят лингвистический анализ внутренних весов сети с целью привязать результат к некоторым традиционным лингвистическим категориям. В данной работе используется корпус текстов большого объёма 9052606 отзывов, написанных 784649 авторами, что, конечно, способствует повыше-

нию качества решения задачи. Другие исследователи использовали простое моделирование текста с помощью эмбеддинга на основе векторного представления слов Word2vec [11] для верификации авторства коротких статей на английском языке. Они добились увеличения качества классификации в модели машинного обучения, основанной на слиянии трех разных архитектур: сверточных нейронных сетей, рекуррентно-сверточных нейронных сетей и машинных классификаторов опорных векторов. Окончательное решение получается путем объединения результатов трех моделей с использованием метода голосования. В результате экспериментов точность (accuracy) оказалась от 91 % до 97 %.

Однако большое количество современных исследований проблемы верификации авторов идёт по пути совершенствования характеристических векторов текстов. В качестве параметров добавляются более сложные текстовые функции.

Ли и соавторы [12] предприняли попытку применить функции специфичные для предметной области. Они использовали 233 функции, включая 227 стилометрических функций и шесть новых специфических для социальных сетей функций. Стилометрические функции включали уровень символов: частоты отдельных букв, прописных букв, специальных символов; уровень слов: общее количество слов, средняя длина слова, количество слов с одним символом и т. д .; синтаксические: количество знаков препинания и функциональных слов, общее количество предложений. Набор специфичных для социальных сетей функций включал смайлики, сокращения, начало предложения без заглавной буквы, окончание предложения без знака препинания и отсутствие упоминания «Я» или «Мы» в сообщении. Они разработали алгоритмы и исследовали различные классификаторы для определения подлинности коротких сообщений социальной сети Facebook. Результаты экспериментов по проверке, является ли указанный пользователь автором данного сообщения, показали среднюю точность 79,6 % для 30 пользователей и 9259 сообщений. Это качество было достигнуто за счет стилометрических характеристик. Функции на основе предложений показали худшую производительность с точностью 53,6 %. Особенности социальных сетей не улучшили классификацию. Тот факт, что особенности, основанные на предложениях, не повлияли на качество классификации, можно объяснить особенностью коротких сообщений в социальных сетях, поскольку они редко состоят из большого количества предложений. Более интересно, что специальные символы социальных сетей влияют на решение данной конкретной задачи гораздо меньше, чем лексические и синтаксические особенности.

Определение авторства художественных произведений связывается в первую очередь со сложными стилистическими характеристиками текста. При решении вопроса о подлинности текстов Плиния [13], оценки принадлежности спорного произведения Данте [5], решения проблемы проверки авторства Гёте по отношению к анонимным статьям [4] рассматриваются стилометрические параметры основанные на специфических фразах, характерных для автора или времени написания. Однако авторы подчёркивают неоднозначность полученных результатов и необходимость продолжения исследований. Важность выбора релевантных характеристик подчёркивают авторы работы, посвящённой верификации античных авторов [14].

Исследователи используют параметры, вычисляемые на основе структуры текста. В статье [15] представлен подход к решению проблемы проверки авторства в области судебной медицины. Разработанный метод использует графы для представления лексических и синтаксических аспектов текстов. На основе этих структур данных вычисляются лингвистические функции, которые позволяют выявить стиль письма автора. Предлагаемый метод применяется к англоязычным документам.

Следует отметить, что исследователи национальных языков также обращают внимание на эффективность использования структурных характеристик текста [16]. В этой работе задача верификации авторов решается для книг на классическом арабском языке. Метод, предлагаемый учёными основан на сходстве лингвистических особенностей текста и применяет ряд лексических, мор-

фологических и синтаксических признаков и ансамблей признаков. В результате экспериментов точность достигла 87,1 %, однако корпус текстов состоял всего из 31 книги.

Следует обратить внимание, что появление сложных текстовых функций позволяет не только получить ответ классификатора, как чёрного ящика, но и проводить анализ результатов в рамках предметной области и лингвистики. Такие исследования могут оказаться очень полезными, так как дают возможность собирать информацию о деталях стиля написания текста, проводить качественный анализ ошибок, выявлять ограничения методов, прогнозировать возможность их применения в предметных областях. Например, авторам работы [17] удалось показать, что у каждого писателя есть свой стиль, который проявляется в использовании нескольких типов стилометрических характеристик, уникальных для отдельных авторов. Таким образом, выделение нескольких групп параметров текста на разных уровнях: символов, слов, структуры предложений, и исследование их влияния на качество верификации авторства является актуальной задачей в области автоматической обработки естественного языка.

2. Обзор характеристик

Стилометрический анализ текста включает в себя поиск и подсчёт различных стилометрических характеристик. Выбор этих характеристик текста является важнейшим этапом его исследования. Среди таких характеристик можно выделить несколько категорий:

1. Уровень символов:
   (a) количество отдельных букв;
   (b) общее количество букв;
   (c) количество отдельных символов;
   (d) общее количество символов
   (e) средняя длина предложения в символах;
2. Уровень слов:
   (a) количество слов;
   (b) количество предложений;
   (c) средняя длина предложений по количеству слов;
   (d) средняя длина слова;
   (e) частота встречаемости N-грамм из одного, двух и трёх слов;
3. Структурный уровень:
   (a) частота встречаемости POST N-грамм из одного, двух, трёх и четырёх слов.

Для проведения исследований в области верификации авторства авторами были выбраны именно эти стилометрические характеристики, потому что они позволяют наиболее точно определить авторский стиль произведения [18].

3. Классификация текстов

Классификация является одной из важных задач в рамках обработки естественного языка. Она решается с помощью специальных аналитических моделей, называемых классификаторами. В настоящее время существует большое количество различных видов классификаторов, для построения которых используются как статистические методы (логистическая регрессия), так и методы машинного обучения (нейронные сети, деревья решений, метод k-ближайших соседей, машины опорных векторов). Одним из наиболее важных этапов в задаче классификации текста является выбор классификатора. Без этого не возможно определить наиболее эффективную модель для алгоритма классификации текста. Это обусловлено тем, что решаемые задачи могут иметь особенности, связанные с числом классов или с объёмом и качеством исходных данных.

Для решения поставленной задачи авторами были использованы следующие классификаторы:

1. Классификатор DecisionTree: Одним из популярных алгоритмов классификации для анализа текста и данных является дерево решений [19]. Структура этого метода представляет собой иерархическую декомпозицию пространства признаков. Каждый лист дерева представляет собой значение целевой переменной, каждый внутренний узел соответствует одному из признаков. Дерево может быть построено разделением исходных наборов характеристик на подмножества, основанные на проверке значений этих признаков. Каждый узел дерева содержит условие ветвления по одному из признаков.
2. Классификатор Random Forest: Random Forest является ансамблем множества деревьев решений [20]. Это позволяет повысить точность классификации по сравнению с одним деревом. Результат классификации получается в итоге агрегирования ответов множества деревьев.
3. Классификатор SVM: Задача бинарной классификации с помощью метода опорных векторов (SVM) состоит в построении оптимальной разделяющей гиперплоскости в пространстве признаков текстов высокой размерности. Задача классификации на несколько классов с помощью метода опорных векторов заключается в переходе от задачи классификации на множество классов к множественной задаче разбиения на два класса. Первый вариант перехода соответствует стратегии «один против всех». Обучается несколько классификаторов, в соответствии с количеством классов. Классификатор с самым лучшим значением функции выхода присваивает текст к соответствующему классу. Второй вариант перехода соответствует стратегии «один против одного». Также обучается несколько классификаторов по количеству классов. Текст классифицируется в соответствии с тем к какому классу его отнесло большинство классификаторов.
4. Классификатор Gaussian Naive Bayes: Метод наивного байесовского классификатора, применяемого для классификации текстов [21], основан на теореме Байеса:

$$P(c|d) = \frac{P(d|c) * P(c)}{P(d)},$$

где, $P(c|d)$ — вероятность, что текст d принадлежит классу c; $P(d|c)$ — вероятность встретить текст d среди всех текстов класса c; $P(c)$ — безусловная вероятность встретить текст класса c в корпусе текстов; $P(d)$ — безусловная вероятность текста d в корпусе текстов.
5. Классификатор AdaBoost Классификатор AdaBoost используется с целью повышения точности алгоритмов классификации. Он лучше всего работает с обучающими алгоритмами, которые показывают наименее точные результаты (слабые обучающие алгоритмы). Наиболее распространенными алгоритмами, используемыми с AdaBoost, являются одноуровневые деревья решений. Кроме того, слабые классификаторы легко вычисляются, и поэтому появляется возможность объединять много сущностей алгоритма, для создания более сильного классификатора с помощью бустинга.

Векторы стилометрических характеристик были поданы на вход каждому из классификаторов для последующего анализа.

С целью проверки устойчивости классификаторов была применена техника пятикратной кроссвалидации. Тексты были разделены на пять частей, 80 % из которых составляли тренировочную выборку, а оставшиеся 20 % — тестовую. Оценка качества выполнялась с помощью таких параметров, как точность, полнота и F-мера.

4. Эксперименты

4.1. Корпус текстов

Для проведения экспериментов использовались корпуса художественной литературы на русском языке (724 фрагмента за период с 1832 до 2020 года), на английском языке (664 фрагмента за

период с 1816 до 2019 года) и на французском языке (500 фрагментов за период с 1823 года до 2019 года). В каждом из корпусов были представлены произведения 20 авторов. Размеры фрагментов варьируются от 15,000 до 20,000 слов.

4.2. Постановка экспериментов

Расчёт стилометрических характеристик для последующей классификации текстов выполнялся автоматически. Порядок проведения экспериментов был следующим. С помощью алгоритмов, разработанных ранее в рамках проекта ProseRhythmDetector [22], для текстов были подсчитаны стилометрические характеристики разного уровня, а затем записаны в csv-таблицы.

Csv-таблицы были скомбинированны между собой различными способами:

- Характеристики уровня символов;
- Характеристики уровня слов;
- Характеристики структурного уровня;
- Характеристики уровня символов и слов;
- Характеристики уровня символов и структурного уровня;
- Характеристики уровня слов и структурного уровня;
- Характеристики уровня символов, слов и структурного уровня.

С помощью алгоритма, разработанного ранее в рамках проекта ProseRhythmDetector, была проведена классификация текстов по авторам. Результаты экспериментов были представлены в виде таблиц.

4.3. Результаты экспериментов

Для обозначения стилометрических характеристик различного уровня в таблицах с результатами были приняты следующие условные обозначения: характеристики уровня символов – Ch, характеристики уровня слов – W, характеристики структурного уровня – St.

Для выявления наилучшего классификатора было проведено сравнение результатов, показанных всеми вышеописанными классификаторами на основе совокупности характеристик уровня символов, слов и структуры. Сравнение осуществлялось по показателям точности, полноты и F-меры.

Точность — это число правильно положительно классифицированных текстов, поделенное на число всех положительно классифицированных текстов:

$$P = \frac{TP}{TP + FP}.$$

Полнотой называют число правильно классифицированных текстов, поделенное на число всех подходящих текстов:

$$R = \frac{TP}{TP + FN}.$$

F-мера — это среднее гармоническое точности и полноты:

$$F = \frac{2PR}{P + R}.$$

Сравнение проводилось по усреднённым показателям. Ниже представлены получившиеся таблицы.

На основе таблиц 1, 2, 3 можно сделать вывод, что лучшие результаты показал классификатор AdaBoost с наивысшими значениями по всем показателям для всех исследуемых языков. Исходя из этого он был выбран для получения более развёрнутых результатов.

Рассмотрим подробнее результаты, полученные после работы классификатора AdaBoost. В таблицах 4, 5, 6 представлены результаты работы классификатора на основе различных комбинаций

Table 1. Russian (Ch + W + St, average metrics)

Таблица 1. Русский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**97.2**	**90.4**	**93.7**
DecisionTree	78.9	77.8	78.4
GaussianNB	64.2	77.7	70.3
RandomForest	72.4	60.0	65.6
SVM	94.4	80.4	86.9

Table 2. English (Ch + W + St, average metrics)

Таблица 2. Английский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**97.2**	**87.6**	**92.2**
DecisionTree	79.8	79.8	79.8
GaussianNB	64.1	79.5	71.0
RandomForest	68.5	58.9	63.3
SVM	94.1	81.8	87.5

Table 3. French (Ch + W + St, average metrics)

Таблица 3. Французский язык (Ch + W + St, средние метрики)

Классификатор	Ср. точность	Ср. полнота	Ср. F-мера
AdaBoost	**98.0**	**88.6**	**93.1**
DecisionTree	78.7	77.8	78.2
GaussianNB	67.1	80.6	73.2
RandomForest	70.4	61.4	65.6
SVM	96.7	85.5	90.8

характеристик уровня символов, слов и структуры, описанных в предыдущем разделе. В таблицах отражены средние показатели по каждому из трёх параметров.

Первое, на что стоит обратить внимание, это высокие (больше 90) значения точности для каждого набора характеристик во всех трёх таблицах. Единственным исключением из этого правила является значение уровня символов, рассчитанное на основе корпуса английских текстов. Здесь значение уровня символов равняется 85,3. Средние значения полноты для всех трёх языков находятся в пределах от 80,0 до 90,0 за редким исключением.

Наиболее важным показателем является F-мера, демонстрирующая баланс точности и полноты. Этот показатель во всех трёх таблицах имеет значение не менее 79,5, что уверенно можно назвать хорошим результатом.

Анализируя таблицы можно сделать вывод, что наиболее высокое значение F-меры достигается при соединении характеристик на всех трёх уровнях. В случае с русским и французским языками уровни символов, слов и структуры по отдельности вносят примерно одинаковый вклад в общий результат, однако английский язык не подчиняется данной тенденции: основной вклад в итоговое значение вносит уровень слов, а наименьший (с разницей практически в 10 пунктов) уровень символов.

Результаты, полученные при классификации на основе попарного соединения характеристик уровня символов, слов и структуры, также не позволяют прийти к однозначному заключению сразу для трёх языков, касаемо вопроса о наиболее эффективной комбинации характеристик различного уровня. Для русского и французского языков наилучшее значение было получено при соединении символьного и структурного уровня (92,0 и 92,2 соответственно), однако для английского языка лучший показатель F-меры был достигнут при соединении структурного уровня с уровнем слов (91,7).

Однако стоит отметить, что для всех трёх языков различные попарные комбинации характеристик дали примерно одинаковые результаты.

Table 4. Russian (AdaBoost, all metrics)

Таблица 4. Русский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	90.4	83.4	86.8
W	94.6	83.0	88.4
St	93.2	82.4	87.5
Ch + St	96.0	88.3	92.0
Ch + W	95.4	86.8	90.9
W + St	96.2	86.5	91.1
Ch + W + St	97.2	90.4	93.7

Table 5. English (AdaBoost, all metrics)

Таблица 5. Английский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	85.3	74.4	79.5
W	92.7	84.0	88.1
St	90.6	81.6	85.9
Ch + St	93.1	86.7	89.8
Ch + W	94.5	85.3	89.7
W + St	95.2	88.4	91.7
Ch + W + St	97.2	87.6	92.2

Table 6. French (AdaBoost, all metrics)

Таблица 6. Французский язык (AdaBoost, все метрики)

Уровни	Ср. точность	Ср. полнота	Ср. F-мера
Ch	93.7	84.6	88.9
W	90.5	82.2	86.2
St	91.2	79.6	85.0
Ch + St	96.4	88.4	92.2
Ch + W	95.9	87.3	91.4
W + St	95.3	86.1	90.5
Ch + W + St	98.0	88.6	93.1

Следующим шагом рассмотрим расширенные таблицы, демонстрирующие все вышеописанные комбинации характеристик разного уровня на примере конкретных авторов. Помимо значений трёх основных параметров, в этих таблицах представленно и значение стандартного отклонения для каждого из них.

Таблицы 7, 8, 9, содержащие результаты исследования на материале русскоязычного корпуса, позволяют сделать вывод о том, что комбинация характеристик структурного уровня с уровнем символов или уровнем слов и позволяет получить если не самое высокое значение F-меры, то близкое к этому. Сочетание всех трёх уровней дало лучший результат для фрагментов авторства Тургенева, Алексея Толстого, Лескова, Пелевина, Горького, Пикуля, Достоевского, Рубанова и Стругацких. А при попарном сочетании характеристик структурного уровня с уровнем слов или уровнем символов, лучший результат был достигнут для фрагментов авторства Набокова, Маканина, Аксенова, Славниковой, Льва Толстого, Водолазкина, Солженицина и Гоголя.

К тому же следует обратить внимание на минимальность стандартного отклонения F-меры при комбинировании характеристик сразу трёх уровней: в большинстве случаев оно не превышает 5,0.

Table 7. Russian (AdaBoost, all metrics, all authors)

Таблица 7. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	И С Тургенев	95.1	4.7	79.4	10.5	86.5	6.4
W	И С Тургенев	87.8	10.5	81.8	10.5	84.3	5.2
St	И С Тургенев	92.0	5.1	86.6	4.8	88.2	2.0
Ch + St	И С Тургенев	94.8	5.2	90.8	8.1	85.9	9.7
Ch + W	И С Тургенев	95.0	6.1	86.8	8.8	92.5	5.0
W + St	И С Тургенев	92.6	6.3	86.9	8.5	91.4	5.9
Ch + W + St	И С Тургенев	95.9	6.3	91.9	7.5	93.5	1.3
Ch	А К Толстой	91.8	8.9	83.1	13.8	84.1	3.2
W	А К Толстой	94.2	4.5	81.4	7.7	88.4	5.0
St	А К Толстой	92.2	8.7	86.1	10.9	83.6	8.0
Ch + St	А К Толстой	91.5	8.5	85.7	2.6	90.2	6.5
Ch + W	А К Толстой	95.9	4.2	84.3	11.5	91.3	5.0
W + St	А К Толстой	91.9	4.0	87.1	11.4	92.9	4.7
Ch + W + St	А К Толстой	92.5	6.0	90.7	8.3	93.5	2.3
Ch	Н С Лесков	77.4	16.8	71.6	12.9	74.8	13.2
W	Н С Лесков	89.2	12.6	78.2	8.9	79.7	3.2
St	Н С Лесков	84.6	18.3	74.5	5.7	75.7	15.2
Ch + St	Н С Лесков	91.9	9.9	76.3	7.5	75.4	5.4
Ch + W	Н С Лесков	99.3	0.3	81.0	6.7	83.9	5.2
W + St	Н С Лесков	87.6	19.2	71.9	12.7	89.7	12.6
Ch + W + St	Н С Лесков	94.4	10.2	77.5	9.0	93.0	4.5
Ch	В О Пелевин	91.5	12.8	91.0	5.2	87.8	8.5
W	В О Пелевин	96.6	4.9	80.1	8.9	86.7	6.2
St	В О Пелевин	88.8	3.3	83.9	2.2	84.7	7.4
Ch + St	В О Пелевин	94.9	4.0	89.9	7.1	92.0	5.6
Ch + W	В О Пелевин	97.0	3.3	94.6	6.6	97.3	2.8
W + St	В О Пелевин	93.0	3.5	89.5	4.2	89.8	7.3
Ch + W + St	В О Пелевин	96.5	3.9	95.1	4.1	96.7	1.9
Ch	В В Набоков	91.5	6.3	81.5	7.6	85.7	4.1
W	В В Набоков	96.4	3.8	73.2	11.5	71.5	4.6
St	В В Набоков	86.1	9.6	74.8	10.0	85.9	5.5
Ch + St	В В Набоков	95.4	4.8	76.3	14.5	90.3	6.5
Ch + W	В В Набоков	91.9	7.0	82.2	6.5	89.1	5.8
W + St	В В Набоков	93.5	7.3	72.4	5.5	82.1	13.1
Ch + W + St	В В Набоков	93.3	7.4	89.5	7.0	88.0	7.5
Ch	М Горький	95.9	6.7	87.8	5.5	90.9	6.1
W	М Горький	95.2	5.9	80.9	8.7	89.8	4.3
St	М Горький	99.3	0.4	84.5	5.9	94.3	4.3
Ch + St	М Горький	98.2	2.8	93.7	5.3	92.0	2.9
Ch + W	М Горький	99.8	0.2	91.8	8.5	94.3	4.6
W + St	М Горький	99.1	0.9	88.3	6.5	93.5	1.7
Ch + W + St	М Горький	99.6	0.4	91.9	5.4	96.3	2.5
Ch	М А Булгаков	86.0	11.5	75.9	14.8	69.0	13.8
W	М А Булгаков	96.9	4.9	78.0	17.4	81.2	10.7
St	М А Булгаков	82.1	20.9	72.6	17.3	71.8	7.0
Ch + St	М А Булгаков	86.7	8.6	83.0	11.0	74.3	15.2
Ch + W	М А Булгаков	96.0	6.5	81.4	17.6	88.1	10.5
W + St	М А Булгаков	99.2	0.3	78.0	12.9	77.9	15.1
Ch + W + St	М А Булгаков	99.4	0.3	82.6	9.5	87.7	11.7

Рассмотрим таблицы 10, 11, 12, в которых отражены результаты исследования, полученные на материале англоязычного корпуса. Анализ полученных значений F-меры для каждой из возмож-

Table 8. Russian (AdaBoost, all metrics, all authors)

Таблица 8. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	В С Пикуль	94.3	4.2	88.9	7.0	88.8	7.4
W	В С Пикуль	96.9	5.1	87.1	11.4	92.7	4.7
St	В С Пикуль	99.5	0.4	93.7	5.8	96.2	4.7
Ch + St	В С Пикуль	98.2	2.9	94.3	5.1	94.9	2.9
Ch + W	В С Пикуль	97.0	3.3	90.5	5.8	95.2	4.1
W + St	В С Пикуль	98.6	2.0	89.5	5.8	96.5	2.4
Ch + W + St	В С Пикуль	98.6	1.7	92.7	9.8	97.5	3.0
Ch	Д И Рубина	94.9	3.2	81.8	10.1	85.0	7.4
W	Д И Рубина	93.4	7.0	86.7	6.9	89.1	1.4
St	Д И Рубина	93.9	6.7	72.0	7.3	82.0	5.8
Ch + St	Д И Рубина	96.1	3.8	81.2	4.3	85.0	9.3
Ch + W	Д И Рубина	95.7	5.3	91.5	5.2	94.2	3.0
W + St	Д И Рубина	99.0	0.4	89.0	4.7	86.2	7.1
Ch + W + St	Д И Рубина	97.8	2.7	87.6	4.9	92.9	4.2
Ch	В С Маканин	96.9	4.7	90.8	9.4	91.9	4.7
W	В С Маканин	97.9	3.2	88.4	4.3	95.9	5.1
St	В С Маканин	99.4	0.3	88.1	7.3	88.2	6.1
Ch + St	В С Маканин	98.6	2.1	95.0	4.5	97.6	4.8
Ch + W	В С Маканин	99.6	0.3	91.8	4.1	95.7	4.7
W + St	В С Маканин	98.2	3.3	94.3	5.3	95.9	3.7
Ch + W + St	В С Маканин	99.6	0.4	94.3	6.8	96.4	3.3
Ch	В П Аксенов	82.3	14.0	73.9	9.7	71.9	8.0
W	В П Аксенов	88.2	20.3	73.7	11.1	74.4	6.4
St	В П Аксенов	85.6	13.6	68.0	7.8	78.3	8.2
Ch + St	В П Аксенов	95.5	6.3	79.3	7.7	82.4	7.4
Ch + W	В П Аксенов	84.0	17.9	76.0	15.4	83.3	3.1
W + St	В П Аксенов	98.8	0.7	79.1	8.5	87.4	6.6
Ch + W + St	В П Аксенов	99.0	0.6	88.6	8.2	85.4	4.8
Ch	Ф М Достоевский	86.7	7.0	90.0	7.4	86.4	6.8
W	Ф М Достоевский	88.8	7.9	77.7	3.7	89.0	4.3
St	Ф М Достоевский	94.6	3.2	94.5	3.1	93.2	2.4
Ch + St	Ф М Достоевский	94.9	4.6	95.7	4.6	94.1	6.0
Ch + W	Ф М Достоевский	88.8	7.9	77.7	3.7	89.0	4.3
W + St	Ф М Достоевский	95.3	3.9	94.5	1.7	93.9	5.0
Ch + W + St	Ф М Достоевский	98.5	2.2	97.7	3.0	95.6	1.6
Ch	А В Рубанов	97.4	2.7	92.3	4.4	95.5	4.6
W	А В Рубанов	98.1	3.0	93.5	5.6	91.2	5.5
St	А В Рубанов	97.1	3.8	88.8	3.0	93.9	3.7
Ch + St	А В Рубанов	99.8	0.4	91.3	1.6	95.5	5.0
Ch + W	А В Рубанов	98.1	2.6	93.2	9.3	96.4	3.5
W + St	А В Рубанов	98.1	3.0	93.5	5.6	91.2	5.5
Ch + W + St	А В Рубанов	99.8	0.3	95.8	3.6	97.2	3.5

ных комбинаций стилометрических характеристик показывает, что использование характеристик структурного уровня часто позволяет улучшить результат. Это видно на примере сочинений Генти (G A Henty), Моэма (W S Maugham), Честертона (G K Chesterton), Мойес (J Moyes), Элиот (G Eliot), Коллинза (W Collins), Троллопа (A Trollope), Лэнга (A Lang), Пратчетта (T Pratchett), Смит (Z Smith), Геймана (N Gaiman), Джеймса (H James), Харди (T Hardy), Роулинг (J K Rowling) и Макьюэна (I McEwan).

Сочетание характеристик всех трёх уровней тоже позволяет добиться улучшения результата верификации, хотя это прослеживается не так явно, как для русскоязычных авторов. Для текстов англоязычного корпуса наилучшие результаты при комбинации характеристик всех трёх уровней удалось получить для текстов 6 авторов: Генти, Моэма, Элиота, Смит, Харди и Макьюэна. К тому же,

Table 9. Russian (AdaBoost, all metrics, all authors)

Таблица 9. Русский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	О А Славникова	98.4	2.9	93.1	5.5	96.3	4.4
W	О А Славникова	99.6	0.2	88.7	11.3	97.3	2.5
St	О А Славникова	99.7	0.4	96.7	4.4	98.8	1.5
Ch + St	О А Славникова	99.9	0.1	96.6	4.3	98.4	2.0
Ch + W	О А Славникова	99.4	0.3	94.7	6.9	95.9	3.8
W + St	О А Славникова	99.6	0.2	88.7	11.3	97.3	2.5
Ch + W + St	О А Славникова	99.9	0.2	97.3	3.3	98.0	2.5
Ch	Стругацкие	83.6	3.9	81.4	13.3	86.5	7.9
W	Стругацкие	91.7	3.9	82.8	3.4	87.3	6.1
St	Стругацкие	91.9	8.4	68.5	7.5	73.4	12.7
Ch + St	Стругацкие	95.0	5.4	87.6	8.4	87.0	5.7
Ch + W	Стругацкие	95.6	5.2	88.5	9.5	88.3	7.9
W + St	Стругацкие	91.7	3.9	82.8	3.4	87.3	6.1
Ch + W + St	Стругацкие	95.2	3.8	92.4	4.1	89.5	5.8
Ch	Л Н Толстой	71.7	15.7	63.8	13.0	66.6	10.0
W	Л Н Толстой	87.3	11.0	73.2	8.6	83.1	13.6
St	Л Н Толстой	94.0	6.5	71.1	4.8	83.8	10.5
Ch + St	Л Н Толстой	94.4	9.9	77.6	7.4	87.9	5.0
Ch + W	Л Н Толстой	83.6	18.7	74.4	8.9	68.5	12.7
W + St	Л Н Толстой	87.3	11.0	73.2	8.6	83.1	13.6
Ch + W + St	Л Н Толстой	88.9	20.2	75.0	13.9	87.5	4.4
Ch	Е Г Водолазкин	96.9	4.9	87.2	5.1	92.5	1.8
W	Е Г Водолазкин	99.7	0.4	94.1	3.0	90.0	7.1
St	Е Г Водолазкин	94.0	5.3	90.6	7.2	91.8	5.7
Ch + St	Е Г Водолазкин	99.7	0.1	100.0	0.0	97.5	2.2
Ch + W	Е Г Водолазкин	99.7	0.3	87.0	10.5	96.5	2.0
W + St	Е Г Водолазкин	99.7	0.4	94.1	3.0	90.0	7.1
Ch + W + St	Е Г Водолазкин	99.9	0.2	97.1	5.7	96.8	3.9
Ch	А А Проханов	95.4	4.8	96.8	3.8	95.2	4.9
W	А А Проханов	96.4	3.2	98.3	3.3	97.2	1.6
St	А А Проханов	96.4	4.8	93.0	4.7	95.2	5.9
Ch + St	А А Проханов	97.2	3.1	98.7	2.5	96.1	4.7
Ch + W	А А Проханов	99.9	0.2	97.5	3.2	100.0	0.0
W + St	А А Проханов	96.4	3.2	98.3	3.3	97.2	1.6
Ch + W + St	А А Проханов	99.7	0.3	98.1	3.2	97.9	2.7
Ch	А И Солженицын	91.4	6.6	86.0	9.1	83.5	4.0
W	А И Солженицын	98.3	2.1	90.9	5.3	94.6	3.0
St	А И Солженицын	93.2	4.9	85.2	9.4	90.7	4.4
Ch + St	А И Солженицын	98.0	2.9	89.8	5.8	92.1	4.0
Ch + W	А И Солженицын	94.2	4.6	91.7	5.6	93.4	4.2
W + St	А И Солженицын	98.3	2.1	90.9	5.3	94.6	3.0
Ch + W + St	А И Солженицын	96.6	3.4	87.7	4.6	91.8	3.8
Ch	Н В Гоголь	89.1	9.7	72.4	5.0	78.4	2.7
W	Н В Гоголь	99.4	0.5	71.7	12.5	76.6	18.0
St	Н В Гоголь	99.4	0.4	75.5	12.8	87.9	6.4
Ch + St	Н В Гоголь	99.5	0.7	83.7	10.7	88.3	6.9
Ch + W	Н В Гоголь	97.1	5.2	76.4	12.6	85.8	11.4
W + St	Н В Гоголь	99.4	0.5	71.7	12.5	76.6	18.0
Ch + W + St	Н В Гоголь	99.4	0.3	84.7	13.8	81.6	17.2

стандартное отклонение F-меры для комбинации характеристик всех трёх уровней в большинстве случаев не превышает 5,0.

Таблицы 13, 14, 15 демонстрируют результаты экспериментов, проведённых с корпусами французских текстов. В результате анализа этих таблиц удалось определить, что стилометрические

Table 10. English (AdaBoost, all metrics, all authors)

Таблица 10. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	G A Henty	93.9	4.4	78.2	5.5	82.1	5.0
W	G A Henty	94.9	5.1	81.4	10.8	84.3	10.0
St	G A Henty	81.3	17.3	88.0	11.9	86.4	5.5
Ch + St	G A Henty	89.8	8.7	87.8	6.7	88.4	7.4
Ch + W	G A Henty	92.7	6.3	83.2	13.3	86.4	3.3
W + St	G A Henty	93.3	7.0	82.0	8.3	88.4	5.9
Ch + W + St	G A Henty	95.9	3.8	94.0	5.5	88.7	4.4
Ch	W S Maugham	87.7	8.2	78.2	9.0	77.4	6.4
W	W S Maugham	92.1	5.8	90.8	10.8	88.2	8.7
St	W S Maugham	92.8	3.0	86.3	6.1	82.1	9.3
Ch + St	W S Maugham	93.1	5.2	86.2	7.5	82.0	8.8
Ch + W	W S Maugham	96.0	4.1	86.5	9.2	84.9	5.6
W + St	W S Maugham	97.0	3.3	87.7	7.9	92.4	5.4
Ch + W + St	W S Maugham	96.5	3.4	86.6	8.4	93.1	4.3
Ch	G K Chesterton	83.7	11.8	62.2	4.2	68.6	6.5
W	G K Chesterton	98.4	2.3	85.1	5.3	93.8	3.8
St	G K Chesterton	97.9	2.9	83.7	12.2	92.8	2.4
Ch + St	G K Chesterton	92.6	9.8	91.1	8.6	84.3	8.8
Ch + W	G K Chesterton	97.0	3.2	90.7	5.8	90.6	6.8
W + St	G K Chesterton	99.5	0.3	89.4	6.5	96.2	3.2
Ch + W + St	G K Chesterton	97.9	3.0	89.4	9.4	93.1	4.6
Ch	J Moyes	74.4	16.2	69.3	9.8	65.9	12.3
W	J Moyes	83.4	18.0	83.9	5.2	84.2	10.2
St	J Moyes	97.2	4.2	81.3	7.0	82.3	8.8
Ch + St	J Moyes	97.1	3.8	84.0	9.1	89.5	6.6
Ch + W	J Moyes	96.7	5.2	81.9	5.4	87.7	5.0
W + St	J Moyes	97.1	4.4	85.5	6.5	92.2	6.6
Ch + W + St	J Moyes	97.6	3.4	87.0	7.9	91.7	4.8
Ch	G Eliot	66.6	16.5	60.6	8.1	64.6	8.6
W	G Eliot	88.7	12.2	73.8	11.4	85.3	6.7
St	G Eliot	92.0	9.7	68.4	9.7	81.5	6.4
Ch + St	G Eliot	91.4	9.9	80.2	11.1	89.6	4.2
Ch + W	G Eliot	99.1	0.6	76.5	9.8	81.8	7.2
W + St	G Eliot	96.4	6.4	85.2	11.4	78.7	6.2
Ch + W + St	G Eliot	94.0	7.4	85.9	6.1	92.8	7.3
Ch	W Collins	91.6	6.2	75.2	5.9	84.3	1.6
W	W Collins	98.2	2.1	86.6	3.8	92.4	6.3
St	W Collins	93.3	4.9	82.3	8.0	93.0	6.4
Ch + St	W Collins	96.3	5.9	87.6	6.8	88.1	9.4
Ch + W	W Collins	97.7	3.2	84.2	18.4	91.5	3.6
W + St	W Collins	98.2	3.1	92.0	7.0	92.0	5.0
Ch + W + St	W Collins	97.3	4.3	89.1	4.2	92.2	6.8
Ch	A Trollope	96.7	3.1	83.2	6.1	95.1	3.0
W	A Trollope	96.9	4.7	91.8	5.2	96.4	1.8
St	A Trollope	97.7	3.8	93.3	6.2	95.3	2.9
Ch + St	A Trollope	96.7	3.6	94.2	6.0	95.7	2.3
Ch + W	A Trollope	98.3	2.4	94.6	5.0	94.7	5.4
W + St	A Trollope	99.8	0.2	94.5	2.9	97.3	1.6
Ch + W + St	A Trollope	98.4	2.2	90.9	3.2	96.6	1.9

характеристики структурного уровня вносят значительный вклад в улучшение качества верификации авторства. Это подтверждают результаты, полученные на основе произведений Франса (A France), Гара (R Gard), Колетт (Colette), Панколь (K Pancol), Мопассана (G Maupassant), Золя (É Zola), Гюго (V Hugo), Роллана (R Rolland), Жида (A Gide), Леви (M Levy), Бальзака (H Balzac)

Table 11. English (AdaBoost, all metrics, all authors)

Таблица 11. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	H Bindloss	90.4	6.6	78.9	11.6	84.6	8.8
W	H Bindloss	98.5	1.7	88.1	3.5	89.1	3.6
St	H Bindloss	97.1	3.6	91.4	7.9	88.5	3.7
Ch + St	H Bindloss	99.7	0.3	91.6	2.7	93.9	4.0
Ch + W	H Bindloss	98.2	2.3	92.5	6.3	96.6	3.2
W + St	H Bindloss	97.8	2.9	91.0	6.2	92.9	4.7
Ch + W + St	H Bindloss	99.5	0.5	94.9	4.5	95.2	4.2
Ch	K Atkinson	94.3	6.8	86.1	10.1	89.5	9.2
W	K Atkinson	99.3	0.7	93.8	5.2	98.5	3.0
St	K Atkinson	96.9	5.7	80.0	17.5	86.4	12.0
Ch + St	K Atkinson	99.5	0.4	89.1	8.1	95.9	3.4
Ch + W	K Atkinson	99.9	0.2	98.8	2.5	97.0	2.5
W + St	K Atkinson	99.8	0.2	94.0	8.0	95.6	4.0
Ch + W + St	K Atkinson	99.9	0.2	95.1	6.1	93.4	7.0
Ch	Sir W Scott	92.7	6.7	81.6	8.6	89.8	6.2
W	Sir W Scott	93.2	9.8	88.3	8.3	91.3	7.3
St	Sir W Scott	95.4	7.9	87.5	7.3	84.6	7.3
Ch + St	Sir W Scott	93.1	6.0	89.3	9.7	94.5	4.7
Ch + W	Sir W Scott	97.7	4.1	92.6	4.7	95.7	3.9
W + St	Sir W Scott	97.8	3.3	97.0	3.6	95.2	4.2
Ch + W + St	Sir W Scott	97.1	4.8	91.7	5.8	94.1	3.4
Ch	C Kingsley	87.5	8.4	77.2	7.9	79.7	7.9
W	C Kingsley	93.0	8.2	91.0	7.4	94.9	3.1
St	C Kingsley	91.1	4.9	83.4	6.0	84.4	8.0
Ch + St	C Kingsley	96.6	3.6	90.3	5.9	92.2	3.9
Ch + W	C Kingsley	98.2	3.1	87.3	9.1	93.7	8.8
W + St	C Kingsley	97.9	3.3	97.1	5.9	91.1	8.9
Ch + W + St	C Kingsley	99.5	0.5	94.3	5.0	93.7	4.9
Ch	A Lang	71.6	20.3	73.3	12.4	78.5	13.3
W	A Lang	87.8	9.7	71.8	3.5	78.4	6.5
St	A Lang	90.2	7.5	72.4	6.8	84.5	8.1
Ch + St	A Lang	96.8	4.0	76.6	8.4	89.9	5.2
Ch + W	A Lang	85.9	8.3	74.8	6.2	79.6	9.4
W + St	A Lang	89.0	21.0	83.3	10.9	86.7	4.3
Ch + W + St	A Lang	99.1	0.4	79.3	8.6	84.1	10.1
Ch	T Parsons	90.0	12.2	85.9	6.5	87.6	7.3
W	T Parsons	99.1	0.3	84.2	12.1	92.3	8.1
St	T Parsons	95.1	5.2	72.7	18.6	85.1	6.2
Ch + St	T Parsons	84.5	18.6	97.2	3.5	94.4	4.9
Ch + W	T Parsons	99.6	0.2	96.0	5.0	97.3	3.3
W + St	T Parsons	99.5	0.2	88.6	9.1	92.7	2.0
Ch + W + St	T Parsons	99.6	0.5	89.1	9.6	92.2	11.6
Ch	T Pratchett	92.0	6.6	88.2	7.4	86.4	9.3
W	T Pratchett	98.9	1.9	97.3	3.1	95.1	6.0
St	T Pratchett	100.0	0.0	93.0	5.6	98.7	1.7
Ch + St	T Pratchett	99.9	0.2	99.9	0.2	96.1	5.3
Ch + W	T Pratchett	98.8	1.9	94.8	6.7	95.3	2.0
W + St	T Pratchett	97.7	3.1	99.1	1.6	97.3	1.6
Ch + W + St	T Pratchett	97.4	3.3	99.1	1.6	97.7	2.9

и Пруста (M Proust). Комбинация сразу трёх уровней стилометрических характеристик помогла получить лучшие результаты для 3 авторов: Франса, Золя, Роллана и Леви, что меньше, чем для русскоязычных и англоязычных писателей. Среднее отклонение для значений F-меры по корпусу

Table 12. English (AdaBoost, all metrics, all authors)

Таблица 12. Английский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	Z Smith	83.3	18.1	64.0	9.9	75.5	8.2
W	Z Smith	95.7	6.8	79.9	13.6	89.8	7.4
St	Z Smith	75.7	22.4	76.6	12.5	75.0	8.4
Ch + St	Z Smith	99.3	0.7	81.5	12.1	87.4	8.5
Ch + W	Z Smith	93.6	7.0	84.4	14.1	90.8	5.3
W + St	Z Smith	95.5	7.8	85.0	9.4	85.9	7.9
Ch + W + St	Z Smith	99.5	0.3	78.3	18.0	94.3	7.0
Ch	N Gaiman	86.2	8.6	75.1	7.7	80.3	7.0
W	N Gaiman	91.1	6.7	74.9	8.2	74.8	11.1
St	N Gaiman	87.3	6.0	79.4	8.7	81.3	6.6
Ch + St	N Gaiman	84.9	5.9	78.9	8.1	88.1	5.5
Ch + W	N Gaiman	89.6	9.3	76.2	7.9	81.1	10.2
W + St	N Gaiman	92.9	7.9	84.9	13.3	82.8	7.1
Ch + W + St	N Gaiman	92.3	5.2	79.6	8.8	75.8	14.0
Ch	H James	64.2	20.0	52.8	6.7	49.3	0.2
W	H James	64.2	20.2	73.7	14.0	74.7	13.1
St	H James	59.2	20.0	71.7	12.5	69.6	19.0
Ch + St	H James	59.2	20.2	76.6	20.7	76.7	17.4
Ch + W	H James	79.2	24.9	55.0	10.0	76.7	17.3
W + St	H James	69.2	24.7	74.4	21.3	67.0	14.4
Ch + W + St	H James	89.3	19.9	66.6	18.3	70.3	11.2
Ch	T Hardy	88.4	10.1	74.5	3.7	81.1	5.9
W	T Hardy	91.5	9.7	82.2	5.1	81.2	11.1
St	T Hardy	86.7	5.0	84.4	7.4	83.9	8.7
Ch + St	T Hardy	97.7	2.6	82.8	7.0	87.8	5.8
Ch + W	T Hardy	83.1	17.6	82.1	4.2	87.6	6.0
W + St	T Hardy	95.7	4.6	83.8	12.7	80.1	5.6
Ch + W + St	T Hardy	95.4	4.5	81.6	9.7	90.0	3.5
Ch	J K Rowling	90.2	5.2	65.7	9.8	70.8	12.9
W	J K Rowling	93.7	9.8	78.3	9.9	78.3	9.8
St	J K Rowling	90.7	7.5	80.1	11.6	91.4	3.1
Ch + St	J K Rowling	99.3	1.0	89.5	3.3	91.8	5.2
Ch + W	J K Rowling	92.3	8.0	82.9	7.1	86.3	7.4
W + St	J K Rowling	96.6	5.4	83.7	5.9	89.6	8.6
Ch + W + St	J K Rowling	99.4	0.5	87.8	4.2	91.2	3.4
Ch	I McEwan	81.4	5.7	78.5	8.4	79.0	7.7
W	I McEwan	95.4	6.5	82.9	10.7	85.6	5.1
St	I McEwan	94.7	5.2	76.9	10.8	83.6	6.4
Ch + St	I McEwan	94.2	5.6	79.5	4.8	86.7	6.3
Ch + W	I McEwan	96.4	3.1	90.8	3.2	88.8	8.9
W + St	I McEwan	94.1	10.5	90.7	6.1	89.0	6.8
Ch + W + St	I McEwan	99.3	1.0	90.8	6.2	91.2	3.9

франкоязычных текстов оказалось несколько выше, чем для текстов, рассмотренных ранее: здесь в большинстве случаев оно попадало в промежуток от 5,0 до 10,0.

5. Заключение

На основе проведённого исследования можно сделать вывод о высокой значимости стилометрических характеристик структурного уровня в решении задачи верификации авторства. Соединение характеристик структурного уровня с характеристиками уровня слов и/или символов позволило получить наиболее точные результаты во время экспериментов на корпусе русскоязычных (85,0 %), англоязычных (75,0 %) и франкоязычных (60,0 %) художественных текстов (значения в процентах рассчитано как отношение количества авторов, для которых лучший результат удалось получить

Table 13. French (AdaBoost, all metrics, all authors)

Таблица 13. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	G Flaubert	99.1	0.7	89.7	6.7	92.5	5.5
W	G Flaubert	97.9	4.0	97.9	4.0	95.5	4.1
St	G Flaubert	98.0	3.4	95.5	5.9	92.7	4.7
Ch + St	G Flaubert	99.7	0.3	90.7	8.7	92.0	8.7
Ch + W	G Flaubert	96.6	4.2	95.1	6.5	98.3	2.3
W + St	G Flaubert	99.9	0.2	98.0	4.0	95.2	4.4
Ch + W + St	G Flaubert	99.9	0.2	96.6	6.6	97.9	2.6
Ch	A France	97.1	4.4	77.5	14.0	83.3	10.4
W	A France	95.7	7.0	89.9	8.7	83.5	9.6
St	A France	95.4	5.3	81.8	9.7	87.2	7.6
Ch + St	A France	99.8	0.3	91.4	8.0	94.7	3.4
Ch + W	A France	97.4	3.7	89.5	6.6	89.7	8.3
W + St	A France	99.3	0.5	95.2	6.0	91.5	7.3
Ch + W + St	A France	99.4	0.6	88.7	11.5	96.0	3.7
Ch	F Cusset	83.9	19.7	76.8	13.0	73.4	16.2
W	F Cusset	89.0	19.7	67.7	11.0	81.4	8.1
St	F Cusset	83.7	20.5	68.8	5.4	72.6	3.5
Ch + St	F Cusset	96.5	4.5	73.3	16.5	84.7	7.6
Ch + W	F Cusset	76.9	23.7	79.1	12.9	88.1	8.2
W + St	F Cusset	84.8	19.3	75.8	15.6	77.9	10.1
Ch + W + St	F Cusset	94.0	7.1	83.7	10.7	85.3	9.3
Ch	F Beigbeder	92.3	9.9	92.3	6.8	94.9	2.8
W	F Beigbeder	97.0	4.3	87.1	6.6	75.7	14.5
St	F Beigbeder	97.3	3.6	82.5	9.6	85.1	10.0
Ch + St	F Beigbeder	99.6	0.5	90.4	7.7	92.1	4.1
Ch + W	F Beigbeder	99.7	0.4	94.2	7.3	97.7	4.6
W + St	F Beigbeder	97.4	4.0	84.6	12.7	92.2	5.7
Ch + W + St	F Beigbeder	97.6	4.1	90.6	8.1	95.3	4.4
Ch	R Gard	94.2	6.6	85.4	8.4	86.7	8.8
W	R Gard	75.3	9.8	77.4	9.2	78.0	4.9
St	R Gard	93.3	6.9	76.4	3.2	85.1	5.4
Ch + St	R Gard	99.0	0.9	88.8	9.5	90.4	2.8
Ch + W	R Gard	97.0	3.6	83.6	7.4	86.5	6.8
W + St	R Gard	96.6	4.4	82.1	12.7	85.6	7.3
Ch + W + St	R Gard	93.3	5.0	80.3	9.0	85.5	5.6
Ch	Colette	99.6	0.4	92.1	6.6	97.2	4.0
W	Colette	87.5	13.0	76.0	13.0	79.7	11.5
St	Colette	91.0	12.7	89.9	5.1	76.5	11.2
Ch + St	Colette	99.7	0.3	94.6	4.9	99.2	1.6
Ch + W	Colette	97.1	5.1	92.1	7.0	91.0	5.6
W + St	Colette	99.3	0.5	80.4	8.3	94.3	5.1
Ch + W + St	Colette	99.4	0.6	91.6	7.1	91.0	7.5
Ch	K Pancol	93.5	7.2	93.6	4.3	88.7	6.3
W	K Pancol	98.0	2.4	81.5	7.7	87.3	6.0
St	K Pancol	92.2	4.4	85.2	3.1	86.8	10.7
Ch + St	K Pancol	95.7	7.8	94.7	4.8	97.1	2.4
Ch + W	K Pancol	99.3	0.4	91.1	2.5	95.0	6.4
W + St	K Pancol	94.3	5.0	88.4	9.7	91.9	6.6
Ch + W + St	K Pancol	99.8	0.3	96.1	5.1	93.8	2.7

путём соединения структурных характеристик с характеристиками уровня символов и/или слов к общему числу авторов в корпусе). Кроме того, значения среднего отклонения для параметра F-меры в большинстве случаев не превышает отметку 5.0 для русских и английских текстов и 10.0 для французских. Таким образом, полученные результаты позволяют выдвинуть гипотезу о разной

Table 14. French (AdaBoost, all metrics, all authors)

Таблица 14. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	A Nothomb	97.4	4.0	88.4	10.3	82.1	17.4
W	A Nothomb	91.3	10.2	90.2	14.3	95.0	6.9
St	A Nothomb	92.0	10.0	85.5	14.9	80.3	10.7
Ch + St	A Nothomb	96.2	6.7	80.0	12.3	90.0	11.2
Ch + W	A Nothomb	99.7	0.4	84.1	9.1	96.8	4.2
W + St	A Nothomb	96.1	6.9	89.1	9.6	92.0	7.8
Ch + W + St	A Nothomb	99.8	0.3	90.0	8.2	86.4	10.9
Ch	G Maupassant	89.7	12.7	77.4	10.2	83.5	7.8
W	G Maupassant	92.8	7.8	79.0	13.8	76.4	9.6
St	G Maupassant	83.4	21.1	74.7	7.4	79.3	8.2
Ch + St	G Maupassant	95.4	6.4	80.3	10.2	81.6	17.2
Ch + W	G Maupassant	78.9	25.0	78.3	12.5	83.7	17.8
W + St	G Maupassant	86.8	11.7	79.8	4.3	90.3	6.0
Ch + W + St	G Maupassant	99.1	0.5	84.6	15.6	84.1	9.5
Ch	É Zola	99.5	0.3	83.0	12.0	87.2	19.2
W	É Zola	84.1	20.1	77.8	13.6	77.1	7.4
St	É Zola	94.0	9.8	81.4	11.0	87.1	13.7
Ch + St	É Zola	99.7	0.3	91.7	10.5	91.8	8.5
Ch + W	É Zola	99.5	0.6	81.2	17.6	84.5	18.4
W + St	É Zola	99.2	0.4	75.7	10.1	87.9	4.3
Ch + W + St	É Zola	98.0	3.2	89.8	19.9	92.2	5.6
Ch	J G Verne	99.5	0.5	90.3	9.3	89.0	19.7
W	J G Verne	97.2	3.9	93.2	6.0	89.5	7.6
St	J G Verne	92.1	8.0	79.1	5.0	81.3	15.0
Ch + St	J G Verne	99.7	0.3	96.0	8.0	96.5	7.0
Ch + W	J G Verne	99.4	0.8	93.5	5.4	97.0	4.1
W + St	J G Verne	99.0	0.3	81.3	12.0	95.8	5.2
Ch + W + St	J G Verne	99.5	0.8	91.6	5.5	96.9	4.0
Ch	J P Modiano	85.6	20.5	74.9	13.4	91.2	4.5
W	J P Modiano	99.7	0.4	87.3	12.6	94.9	4.6
St	J P Modiano	89.2	19.9	83.3	11.8	79.0	17.4
Ch + St	J P Modiano	99.5	0.6	87.4	12.4	83.7	19.4
Ch + W	J P Modiano	99.6	0.6	90.8	7.6	93.6	6.3
W + St	J P Modiano	99.6	0.4	85.2	9.9	91.0	10.7
Ch + W + St	J P Modiano	99.6	0.2	94.7	6.9	91.9	9.5
Ch	V Hugo	86.7	9.6	66.6	12.2	83.2	10.9
W	V Hugo	88.6	9.6	69.9	10.7	72.8	15.1
St	V Hugo	91.8	8.2	66.3	16.4	74.5	14.3
Ch + St	V Hugo	75.1	22.9	86.5	17.7	74.6	14.3
Ch + W	V Hugo	96.6	4.8	64.3	15.2	74.1	10.7
W + St	V Hugo	98.7	0.6	83.7	5.5	84.8	7.1
Ch + W + St	V Hugo	95.6	6.7	74.9	3.2	77.1	10.8
Ch	G Musso	99.6	0.4	96.1	5.1	93.6	6.6
W	G Musso	78.9	24.4	78.2	14.5	81.5	9.6
St	G Musso	75.7	23.0	66.5	13.9	87.2	7.5
Ch + St	G Musso	99.7	0.3	86.4	18.8	91.6	7.6
Ch + W	G Musso	99.9	0.2	90.1	9.4	90.7	10.6
W + St	G Musso	86.4	20.0	94.0	8.0	88.0	7.2
Ch + W + St	G Musso	96.5	6.6	85.5	10.9	91.6	5.6

степени влияния стилометрических характеристик на качество верификации авторства для различных языков. Это означает, что для каждого языка необходимы самостоятельные исследования для получения наиболее эффективных алгоритмов решения задач определения авторского стиля.

Table 15. French (AdaBoost, all metrics, all authors)

Таблица 15. Французский язык (AdaBoost, все метрики, все авторы)

Характеристики	Автор	Точность	Ст. Откл.	Полнота	Ст. Откл.	F-мера	Ст. Откл.
Ch	A St Exupery	89.3	20.4	76.6	19.9	77.3	10.3
W	A St Exupery	96.1	7.1	80.9	18.7	83.9	19.2
St	A St Exupery	87.3	14.3	64.0	18.7	75.0	16.2
Ch + St	A St Exupery	89.4	20.2	86.7	8.1	87.1	10.9
Ch + W	A St Exupery	89.7	20.1	85.0	13.3	95.4	6.7
W + St	A St Exupery	99.1	0.4	87.6	10.5	86.6	13.3
Ch + W + St	A St Exupery	99.4	0.7	84.1	12.2	86.5	19.5
Ch	R Rolland	99.4	0.6	81.0	16.8	82.5	9.4
W	R Rolland	80.4	18.3	68.1	8.1	80.8	13.5
St	R Rolland	93.5	11.0	71.2	13.8	83.0	12.4
Ch + St	R Rolland	99.2	0.4	80.0	8.3	78.1	17.8
Ch + W	R Rolland	99.0	0.6	83.8	18.3	83.3	0.7
W + St	R Rolland	86.7	19.9	82.8	12.0	84.1	12.1
Ch + W + St	R Rolland	97.3	3.7	81.3	6.3	87.0	8.2
Ch	A Gide	99.6	0.6	93.6	5.9	94.3	3.1
W	A Gide	87.1	9.4	81.2	2.7	87.0	8.9
St	A Gide	95.9	4.0	86.3	3.4	85.9	7.7
Ch + St	A Gide	95.2	5.5	96.2	4.7	97.1	3.8
Ch + W	A Gide	96.4	6.8	97.1	3.7	93.2	6.6
W + St	A Gide	89.8	9.5	87.1	6.5	82.1	7.7
Ch + W + St	A Gide	99.7	0.4	92.6	4.9	91.9	7.2
Ch	M Levy	84.2	18.5	82.5	9.3	86.5	4.6
W	M Levy	82.9	9.7	76.6	16.1	72.5	13.6
St	M Levy	84.6	18.6	73.8	12.5	83.2	6.9
Ch + St	M Levy	89.0	20.2	81.8	6.7	90.9	8.4
Ch + W	M Levy	99.4	0.6	92.6	6.2	84.7	17.9
W + St	M Levy	95.8	6.8	78.7	13.0	88.4	6.4
Ch + W + St	M Levy	96.0	7.1	80.6	9.1	91.4	5.8
Ch	H Balzac	87.1	8.7	75.6	11.8	68.6	12.3
W	H Balzac	92.4	10.0	87.8	9.2	90.2	7.0
St	H Balzac	93.6	7.1	81.5	5.7	88.9	4.9
Ch + St	H Balzac	99.5	0.6	91.2	5.3	93.2	2.1
Ch + W	H Balzac	95.6	6.9	84.8	11.0	89.7	5.9
W + St	H Balzac	99.3	0.3	94.2	5.2	95.3	4.7
Ch + W + St	H Balzac	97.5	3.5	94.3	4.8	93.4	6.0
Ch	M Proust	97.8	3.4	98.6	2.9	98.8	1.5
W	M Proust	98.9	1.5	96.1	5.2	94.9	4.9
St	M Proust	99.7	0.4	98.3	3.3	99.5	1.0
Ch + St	M Proust	100.0	0.0	100.0	0.0	95.2	5.7
Ch + W	M Proust	99.9	0.2	95.2	4.1	98.5	1.9
W + St	M Proust	98.3	3.3	99.0	2.0	100.0	0.0
Ch + W + St	M Proust	99.7	0.6	100.0	0.0	97.3	5.4

Использование высокоуровневых стилометрических характеристик открывает перед учёными широкую перспективу для исследований в области автоматической обработки текстов на естественном языке. Анализ частоты использования N-грамм POS-тегов является шагом в сторону построения структурных шаблонов, которые могут быть использованы для более детального анализа стиля того или иного автора.

References

[1] N. P. Tuchkova and O. M. Ataeva, "Podhody k izvlecheniyu znanij v nauchnyh predmetnyh oblastyah", *Informacionnye i matematicheskie tekhnologii v nauke i upravlenii*, no. 2 (18), pp. 5–18, 2020.

[2] A. Altamimi, N. Clarke, S. Furnell, and F. Li, "Multi-platform authorship verification", in *Proceedings of the Third Central European Cybersecurity Conference*, 2019, pp. 1–7.

[3] O. Halvani, L. Graner, and R. Regev, "Taveer: An interpretable topic-agnostic authorship verification method", in *Proceedings of the 15th International Conference on Availability, Reliability and Security*, 2020, pp. 1–10.

[4] M. Kestemont, G. Martens, and T. Ries, "A computational approach to authorship verification of johann wolfgang goethe's contributions to the frankfurter gelehrte anzeigen (1772–73)", *Journal of European Periodical Studies*, vol. 4, no. 1, pp. 115–143, 2019.

[5] S. Corbara, A. Moreo, F. Sebastiani, and M. Tavoni, "The epistle to cangrande through the lens of computational authorship verification", in *International Conference on Image Analysis and Processing*, Springer, 2019, pp. 148–158.

[6] V. A. Drozdov, "Ob avtorstve poemy «'Ushshak-name» s tochki zreniya akademicheskogo vostokovedeniya i novejshih komp'yuternyh tekhnologij", *Orientalistika*, vol. 3, no. 5, pp. 1360–1378, 2020.

[7] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, and B. Stein, "Overview of the cross-domain authorship verification task at pan 2020", in *CLEF*, 2020.

[8] N. Potha and E. Stamatatos, "Intrinsic author verification using topic modeling", in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, ACM, 2018, pp. 1–7.

[9] S. Adamovic, V. Miskovic, M. Milosavljevic, M. Sarac, and M. Veinovic, "Automated language-independent authorship verification (for indo-european languages)", *Journal of the Association for Information Science and Technology*, vol. 70, no. 8, pp. 858–871, 2019.

[10] B. Boenninghoff, S. Hessler, D. Kolossa, and R. M. Nickel, "Explainable authorship verification in social media via attention-based similarity learning", in *2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 36–45.

[11] N. E. Benzebouchi, N. Azizi, M. Aldwairi, and N. Farah, "Multi-classifier system for authorship verification task using word embeddings", in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, IEEE, 2018, pp. 1–6.

[12] J. S. Li, L.-C. Chen, J. V. Monaco, P. Singh, and C. C. Tappert, "A comparison of classifiers and features for authorship authentication of social networking messages", *Concurrency and Computation: Practice and Experience*, vol. 29, no. 14, e3918, 2017.

[13] E. Tuccinardi, "An application of a profile-based method for authorship verification: Investigating the authenticity of pliny the younger's letter to trajan concerning the christians", *Digital Scholarship in the Humanities*, vol. 32, no. 2, pp. 435–447, 2017.

[14] P. B. Reddy, T. M. Mohan, P. V. K. Raja, and T. R. Reddy, "A novel approach for authorship verification", in *Data Engineering and Communication Technology*, Springer, 2020, pp. 441–448.

[15] E. Castillo, O. Cervantes, and D. Vilarino, "Authorship verification using a graph knowledge discovery approach", *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 6, pp. 6075–6087, 2019.

[16] H. Ahmed, "The role of linguistic feature categories in authorship verification", *Procedia computer science*, vol. 142, pp. 214–221, 2018.

[17] M. A. Al-Khatib and J. K. Al-qaoud, "Authorship verification of opinion articles in online newspapers using the idiolect of author: A comparative study", *Information, Communication & Society*, pp. 1–19, 2020.

[18] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, "A survey on stylometric text features", in *Proceedings of the 25th Conference of Open Innovations Association (FRUCT)*, IEEE, 2019, pp. 184–195.

[19] Y. Polin, T. Zudilova, I. Ananchenko, and T. Vojtyuk, "Derevya reshenij v zadachah klassifikacii: osobennosti primeneniya i metody povysheniya kachestva klassifikacii", *Sovremennye naukoemkie tekhnologii*, no. 9, pp. 59–63, 2020.

[20] B. Xu, X. Guo, Y. Ye, and J. Cheng, "An improved random forest classifier for text categorization.", *JCP*, vol. 7, no. 12, pp. 2913–2920, 2012.

[21] S.-B. Kim, K.-S. Han, H.-C. Rim, and S. H. Myaeng, "Some effective techniques for naive bayes text classification", *IEEE transactions on knowledge and data engineering*, vol. 18, no. 11, pp. 1457–1466, 2006.

[22] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries", in *Proceedings of the 26th Conference of Open Innovations Association (FRUCT)*, IEEE, 2020, pp. 247–255.