# Автономное распознавание демографических атрибутов пользователей социальных сервисов

**Аннотация**. При внесении информации в пустые графы пользовательского аккаунта различных веб-ресурсов многие случайно или специально не заполняют демографические атрибуты (пол, возраст, семенное положение, уровень образования, религиозные и политические взгляды). Важность этих данных заключается в том, что они оказывают положительное влияние на системы рекомендации, интернет-маркетинга и иные ресурсы, предусматривающие персонализацию результатов, увеличивая их эффективность.

Статья повествует об автоматическом определении демографических атрибутов пользователей социальною ресурса Twitter посредством текстовых сообщений и прочих открытых данных пользовательских аккаунтов. Его суть заключается в алгоритме машинного обучения, полностью автономном выстраивании базовой комбинации данных для обучения и тестирования, а также наличии обширного выбора языков и демографических атрибутов.

В ходе проведенного эксперимента подтвердились качественные характеристики и их высокий уровень при определении половой принадлежности, возраста и семейного статуса пользователя для английского, русского, немецкого, французского, итальянского и испанского языков. На английском языке в том числе определяется образование, религиозность и политические вкусы.

**Ключевые слова**: демографические параметры; демографические атрибуты, социальные сети; микроблоги; обработка текстовых сообщений; анализ текстов на естественном языке; оценка содержимого; компьютерная лингвистика; машинное обучение.

## Введение

Ежедневно увеличивается число пользователей сети Интернет, что способствует появлению все новых и новых ресурсов, позволяющих обмениваться информацией, делая доступными индивидуальную сторону личных сведений, в т.ч. и текстовую. Желание скрыть эти данные актуализирует методику частичного распознавания авторов сообщений посредствам демографических атрибутов.

Для интернет-маркетинга и рекомендаций подобный способ позволяет эффективно вести таргетированное продвижение продукции и различных предложений в социальных сетях, где собираются группы со схожими атрибутами. Подобные аспекты также полезны в практическом применении многих социальных дисциплин.

Демографические атрибуты делятся на два вида:

- категориальные – различаются по полу, национальности, расе, семейному положению, уровню образования, профессии, трудоустроенности, религиозным и политическим взглядам;

- численные – делятся по возрасту и уровню дохода. При отражении его в комплексе категорий, возможно рассмотрение атрибута в качестве категориального. Например, людей можно поделить по возрасту на группы. Именно рассматриваемые демографические характеристики составляют его социо-демографический портрет.

Статья повествует о методе определения демографических атрибутов пользователей сети Twitter посредством текстовых сообщений, который примечателен следующими параметрами:

- обширным выбором характеристик: пол, возраст, семейный статус, уровень образования, религиозные и политические взгляды;

- автоматизированный сбор и разметка блоков сообщений согласно параметрам;

- поддержка русского, английского, испанского, немецкого, французского и итальянского языков;

- высококачественный итог.

## 1. Обзор литературы

Суть распознавания демографических атрибутов анонимных пользователей социальных сетей посредством текстовых сообщений имеет схожесть с классической социолингвистикой, выраженное в выявлении языковой специфики различных социальных слоев общества и выделении их в определенные группы (автороведческая экспертиза). Однако имеются и некоторая специфика веб-распознавания, способствующая минимальной информативности о пользователях, создающая сложности в распознавании их характеристик:

- небольшое количество символов, которое оставляют пользователи для экономии времени;

- неформальный стиль сообщений, содержащий в себе нестандартные аббревиатуры, слэнг, неологизмы, а также включающий в себя орфографические и пунктуационные ошибки;

- лингвистические конструкции (теги), ссылки на профили других пользователей, описания собственных эмоций (эмотиконы) и т.д.;

- минимальное качество информации в содержании профиля (спам, фейковые аккаунты).

Представленные выше особенности и ограниченность классического метода атрибуции текста в электронных сообщениях пользователей сети Интернет побудили возникновение большого количества методик,

специализирующихся на сервисах мгновенных сообщений, электронной почте, форумах, блогах, социальных сетях и прочих ресурсах текстовых сообщений. В их основе лежит метод машинного обучения, классифицирующий пользователей согласно лингвистическим и иным критериям по классам и необходимым атрибутам. Пользовательские сообщения представляют собой строки из набора символов, содержащих в себе признаки и базу данных для последующей выборки.

### 1.1 Задача определения гендерной принадлежности

Одним из наиболее простых вариантов распознавания половой принадлежности пользователя Интернет-ресурсов считается его имя, размещенное непосредственно в профиле [1]. Однако его минусов является то, что в качестве реального имени может быть использован псевдоним или попросту универсальное имя, не имеющее гендерной привязки.

Машинное обучение, часто применяемое на сегодняшний день в рассматриваемом вопросе, базируется на бинарной классификации. Т.е. используется информация, которая уже была ранее внесена пользователем в соответствующие графы. Например, в социальной сети Facebook кроме словарей имен берутся данные из полей «interested_in» и «relationship_status». Не исключается также соответствующее распределение друзей [2].

Многие исследователи рассматриваемой темы учитывают N-граммы символов и слов, содержащихся в сообщениях пользователей [3; 4]. Важно лишь выбирать максимально соответствующие заданным параметрам, чтобы исключить слишком большую выборку из «пустой» информации.

Для сети Twitter лучше всего N-граммам символов и слов извлекают не только из текстовых сообщений, но и метаданных, таких как «Screen name», «Full name», «Description» [5]. А для сети Facebook классификация осуществляется на основе статусов к уже вышеназванному посредством тематического моделирования [6].

Выделим перечень структурных признаков:

1) по символам (общему количеству, в т.ч. верхнего регистра);

2) по словам (общему числу, средней длине или фиксированному количеству знаков, похожести между собой);

3) по предложениям (суммарное число разных знаков пунктуации);

4) по всему тексту (общему числу предложений, абзацев, среднему количеству предложений и слов в абзаце и предложении).

Например, первые три вышеуказанных признака часто используются при классификации пользователей Youtube [7], а первые четыре – в новостных порталах и сервисах электронных писем [8]. Важно отметить, что они также применяются вместе с социолингвистическими признаками [9; 10]. Примером может выступать классификация авторов блогов [11].

В Twitter также используется классификация по первым k-словам, стеммам, хештегам, диграммам и триграммам, которые могут присутствовать у пользователей обоих полов [12; 13]. Некоторые авторы научных работ исследовали влияние имени на работу алгоритма [12] и непосредственное окружение пользователя [13].

### 1.2 Задача определения возраста

Выделяют следующие варианты определения возрастной группы пользователей:

- непрерывная переменная;
- возрастная категория (моложе 20 лет, 20-40 лет, 40 лет и старше).

Выделяют также жизненный этап человека (школьник, студент, пенсионер и пр.). Авторы различных работ по выделяли множество выборок:

- на два класса (например, 40-, 40+ [14]);
- на три класса (например, 13-17, 23-27, 33-42 [15; 16];
- регрессионно [17; 18].

Проводилось сравнение вариаций вышеназванных выборок с целью определения возрастной группы пользователей Twitter [19]. Выбору можно делать для блогов [15; 16], записей телефонных разговоров [14], сведений из социальных сетей [9; 19; 20]. Признаки, определяющие возраст, можно разделить следующим образом:

- по N-граммам слов и символов текстовых сообщений;
- по стилистике (частей речи, сленга, средней длине предложения, знакам препинания, акронимам, эмотиконам и пр.).

Не стоит забывать и про специфические особенности источника информации. В качестве примера стоит привести LiveJournal, для которого сыграет роль число друзей, постов, комментариев [21].

### 1.3 Определение других атрибутов

Чтобы определить политические предпочтения и этническую принадлежность пользователя в Twitter, используют не только непосредственные признаки профиля, но и специфику поведения (количество сообщений, их среднее число в день, интервалы между ними и т.д.), содержание сообщений (характерные слова для определенных слоев и групп людей), окружающую среду (друзей, подписчиков и пр.) [22]. Например, с этой целью посредством алгоритма машинного обучения «Метод опорных векторов» проводилось сравнение признаков по текстовым сообщениям и сообществам [23] пользователя.

Географическое положение в том же Twitter классифицируется с помощью метода тематического моделирования [24], например, по

распределению слов исходя из географических локаций [25].

## Выводы

В данной статье рассматривались вариации решения задач, связанных с распознаванием социо-демографических атрибутов пользователей сети Интернет, а именно гендерной принадлежности, возраста, политических предпочтений, этнической принадлежности и географического положения посредством методов машинного обучения. Наиболее часто применяют признаки, находящиеся в их текстовых сообщениях, являясь универсальным для любого ресурса. универсальным практически для любого веб-ресурса. При этом учет специфики лишь улучшает выборку, делая ее точнее.

Текстовые признаки можно разделить на независимые (структурные и N-граммы символов и слов) и зависимые от языка сообщений (социолингвистические). Однако в ходе исследования были выявлены минусы рассматриваемых методик:

- ограниченность атрибутов полом и возрастом;
- невозможность или минимальный функционал автоматических средств, собирающих сообщения размечающих их соответствующими атрибутами;
- невозможность или минимальный функционал фильтра от недостоверности сведений (спама и фейковых страниц);
- минимальное применение социолингвистических вариантов определения признаков, особенных для некоторых атрибутов и их показателей;
- минимальный выбор вариаций машинного обучения (вычленение признаков, выборка максимально информативных данных, обучение, классификация), что препятствует качественному анализу;
- невозможность моделирования одного атрибута в сочетании с другими, что препятствует качественному анализу;
- минимальное количество программных решений с ограниченным функционалом (определяют лишь пол и/или возраст), не способные интегрироваться с промышленными приложениями;
- отсутствие бесплатных решений на русском языке.

## 2. Метод

Все ныне существующие методы, позволяющие определить демографические атрибуты пользователей базируются на машинном обучении с учителем для последующего разделения согласно лингвистическим и иным признакам с соответствующими различиями. Текстовое сообщение – это комбинация символов в строках, содержащая в себе важные сведения для

классификации. Сложность в том, что разметка нуждается в дополнительных источниках сведений, собираемых вручную. Представленная методика содержит ряд преимуществ, среди которых:

- автоматизированное выстраивания базы сведений;
- возможность извлекать различные типы признаков из текстовых сообщений пользователей Twitter;
- расширенный набор атрибутов: все графы Facebook профиля и иные данные, касающиеся предпочтений и интересов пользователя можно применить как атрибут;
- большой перечень языков.

В методичке предусмотрено несколько шагов:

- выстраивание базового комплекса сведений;
- подготовительное сканирование текстовых данных;
- выстраивание описательных характеристик;
- тренинг;
- группировка.

Каждый этап, исключая первый, работает с каждым атрибутом отдельно. При построении базового комплекса сведений собирается информация из сети Twitter, первоначально запрашивая лишь профиль пользователя, а при наличии его ссылки на Facebook, где количество атрибутов больше, идет запрос с последующим сохранением всех его доступных текстовых сообщения и извлекаются всех необходимые значения атрибутов. В итоге, в качестве элемента комплекса сведений по атрибутам и языку выступают символьные строки.

Подготовительное сканирование текстовых данных сопровождается распознаванием языка с помощью библиотеки language-detection. Заранее проводит работу фильтр, исключающий ненужное авторство (ретвиты), что значительно повышает качество конечной информации, определяя точность работы.

Выстраивание описательных характеристик позволяет определить лингвистические признаки, токены которых формируют комплекс характеристик в виде N-грамм от 1 до 3 с учётом их порядка. Регистры символов могут учитываться и нет. Конечный вектор признаков имеет сведения о наличии или отсутствии признака в его данных. Число копий во внимание не берутся.

В процессе тренинга строится классификация по алгоритму машин опорных векторов. При группировке за основу берутся текстовые сообщения и поля профиля произвольного пользователя. Осуществляется распределение по классам согласно языку и атрибутам и выдается итог.

## 3. Результаты экспериментов

Исследование по демографическим атрибутам имеет вид:

1) Получение размеченных сведений о пользователях с верными параметрами атрибута.

2) Тренинг классификатора согласно этой информации.

3) Группировка по этому же принципу остальных сведений.

4) Оценка качества группировки с помощью сравнения параметров атрибута в разных данных и конечном итоге.

Извлечение размеченных данных с корректными значениями происходит при сборе и разметке текстовых сообщений, объем которых составляет 500 пользователей для каждого атрибута и языка. Таким образом, тренинг осуществляется на 450 случайных пользователях, а тестирование – на остальных 50.

Для оценки качества применяют метрику точности (Accuracy), согласно которой в таблице 1 представлены итоги опыта с использованием N-грамм третьего порядка.

Таблица 1 – Результаты экспериментальных исследований

| **Тестовая задача** | **Язык** | **Точность, %** |
|---|---|---|
| Определение пола | Английский | 84 |
| | Русский | 86 |
| | Испанский | 94 |
| | Немецкий | 88 |
| | Французский | 94 |
| | Итальянский | 82 |
| Определение возраста | Английский | 94 |
| | Русский | 92 |
| | Испанский | 92 |
| | Немецкий | 80 |
| | Французский | 84 |
| | Итальянский | 94 |
| Определение семейного статуса | Английский | 98 |
| | Русский | 96 |
| | Испанский | 98 |
| | Немецкий | 94 |
| | Французский | 98 |
| | Итальянский | 94 |
| Определение уровня образованности | Английский | 92 |
| Определение принадлежности к религии | Английский | 94 |
| Определение политических предпочтений | Английский | 82 |

## Заключение

Рассмотренная методика подойдет для интернет-маркетинга с целью увеличения точности таргетированного продвижения товаров и услуг, результативности рекламных кампаний и прибыльности бизнеса. В политической среде с ее помощью можно осуществлять сбор сведений по избирателям и эффективнее расходовать рекламный бюджет. Правоохранительные органы с помощью представленной методики смогли бы глубже анализировать сообщения преступников и исключать их анонимность. Кроме того, метод позволяет, учитывая демографический профиль пользователей, проводить следующие рекомендации:

- товаров и услуг интернет-магазинов;
- пользователей для дружбы/следования в социальных сетях;
- телепередач.

С помощью рассматриваемого метода можно осуществлять анализ разных направлений в сфере обработки персональных данных пользователей сети Интернет. В качестве примера можно привести поиск групп и кластеризацию граф в социальных сетях, персонализацию информационного поиска, т.е. высокий уровень точности поискового запроса, простые решения задач компьютерной лингвистики путем исключения многозначности и омонимов.

## Список литературы


1. Sloan L. Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. / L. Sloan [et al.] – Sociological Research Online. – 2013. – Т. 18. – №. 3. – p. 7.
2. Tang C. What's in a name: A study of names, gender inference, and gender behavior in facebook. / C. Tang [et al.] – Database Systems for Adanced Applications. – Springer Berlin Heidelberg, 2011. – pp. 344-356.
3. Miller Z. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. / Z. Miller, B. Dickinson, W. Hu – International Journal. – 2012. – Т. 2.
4. Deitrick W. Gender identification on twitter using the modified balanced winnow. / W. Deitrick [et al.] – Communications and Network. – 2012. – Т. 4. – №. 3. – pp. 189-195.
5. Burger J. D. Discriminating gender on Twitter. / J. D. Burger [et al.] – Proceedings of the Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2011. – pp. 1301-1309.
6. Schwartz H. A. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. / H. A. Schwartz [et al.] – PloS one. – 2013. – Т. 8. – №. 9. – p. 73791.
7. Filippova K. User demographics and language in an implicit social

network. – Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. – Association for Computational Linguistics, 2012. – pp. 1478-1488.

8. Cheng N. Author gender identification from text. / N. Cheng, R. Chandramouli, K. P. Subbalakshmi – Digital Investigation. – 2011. – T. 8. – №. 1. – pp. 78-88.

9. Rao D. Classifying latent user attributes in twitter. / D. Rao [et al.] – Proceedings of the 2nd international workshop on Search and mining user-generated contents. – ACM, 2010. – pp. 37-44.

10. Rao D. Hierarchical Bayesian Models for Latent Attribute Detection in Social Media. / D. Rao [et al.] – ICWSM. – 2011.

11. Mukherjee A. Improving gender classification of blog authors. / A. Mukherjee, B. Liu – Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 207-217.

12. Liu W. What's in a Name? Using First Names as Features for Gender Inference in Twitter. / W. Liu, D. Ruths – 2013 AAAI Spring Symposium Series. – 2013.

13. Al Zamal F. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. / F. Al Zamal, W. Liu, D. Ruths – ICWSM. – 2012.

14. Garera N. Modeling latent biographic attributes in conversational genres. / N. Garera, D. Yarowsky – Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – Association for Computational Linguistics, 2009. – Vol. 2, pp. 710-718.

15. Schler J. Effects of Age and Gender on Blogging. / J. Schler [et al.] – AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. – 2006. – pp. 199-205.

16. Goswami S. Stylometric analysis of bloggers' age and gender. / S. Goswami, S. Sarkar, M. Rustagi – Third International AAAI Conference on Weblogs and Social Media. – 2009.

17. Nguyen D. Author age prediction from text using linear regression. / D. Nguyen, N. A. Smith, C. P. Rosé – Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. – Association for Computational Linguistics, 2011. – pp. 115-123.

18. C. van Heerden C. Combining regression and classification methods for improving automatic speaker age recognition. / C. van Heerden [et al.] – Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. – IEEE, 2010. – pp. 5174-5177.

19. Nguyen D. "How Old Do You Think I Am?": A Study of Language and Age in Twitter. / D. Nguyen [et al.] – Seventh International AAAI Conference

on Weblogs and Social Media. – 2013.

20. Peersman C. Predicting age and gender in online social networks. / C. Peersman, W. Daelemans, L. Van Vaerenbergh – Proceedings of the 3rd international workshop on Search and mining user-generated contents. – ACM, 2011. – pp. 37-44.

21. Rosenthal S. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. / S. Rosenthal, K. McKeown. – Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. – Association for Computational Linguistics, 2011. – Vol. 1, pp. 763-772.

22. Pennacchiotti M. A Machine Learning Approach to Twitter User Classification. / M. Pennacchiotti, A. M. Popescu – ICWSM. – 2011.

23. Conover M. D. Predicting the political alignment of twitter users. / M. D. Conover [et al.] – Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom). – IEEE, 2011. – pp. 192-199.

24. Eisenstein J. A latent variable model for geographic lexical variation / J. Eisenstein [et al.] – Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. – Association for Computational Linguistics, 2010. – pp. 1277-1287.

25. Cheng Z. You are where you tweet: a content-based approach to geo-locating twitter users. / Z. Cheng, J. Caverlee, K. Lee – Proceedings of the 19th ACM international conference on Information and knowledge management. – ACM, 2010. – pp. 759-768.